Pattern Recognition Letters 177 (2024) 103–109

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

SilentTrig: An imperceptible backdoor attack against speaker identification with hidden triggers

Yu Tang [a], Lijuan Sun [b], Xiaolong Xu [b,*]

[a] *Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, No. 9 Wenyuan Road, Nanjing, 210023, Jiangsu, China*
[b] *School of Computer Science, Nanjing University of Posts and Telecommunications, No. 9 Wenyuan Road, Nanjing, 210023, Jiangsu, China*

ARTICLE INFO

Keywords:
Speaker identification
Backdoor attack
Trigger
Deep learning
Imperceptibility

ABSTRACT

Speaker identification based on deep learning is known to be susceptible to backdoor attacks. However, the current research on audio backdoor attacks is limited, and these attacks often use obvious noises as triggers, which can raise suspicion among users. In this paper, we introduce SilentTrig, a novel and imperceptible backdoor attack method targeted at speaker identification. Our approach involves utilizing an optimized steganographic network to embed triggers into benign audio samples and implementing a two-stage adversarial optimization process. This ensures that the poisoned samples are acoustically indistinguishable from their benign counterparts, resulting in a substantially improved attack imperceptibility. We evaluate SilentTrig on two datasets and four state-of-the-art models. The results demonstrate a high Attack Success Rate (ASR) of up to 99.2%, a Just Noticeable Difference (JND) of only 0.3, and resistance to typical defense methods such as Neural Cleanse and Fine-Pruning.

1. Introduction

Acoustic signal-related technologies have achieved excellent performance and widespread application in behavioral biometrics for personal identification. Speaker Identification (SI) is one of the key technologies that aims to recognize users by their utterances and is widely adopted in real-life scenarios [1]. With the powerful feature extraction capability of deep neural networks (DNNs), researchers have proposed several SI methods based on deep learning, which have taken this technology to a new level. However, the success of these methods depends on large amounts of training data and increasing computational power, making the lengthy and involved training procedure a bottleneck for users and researchers. To reduce overhead, third-party resources are often utilized in training DNNs [2]. For instance, users may provide their sensitive data and opt to delegate model training tasks to machine-learning-as-a-service (MLaaS) provider or even use external datasets.

However, the lack of transparency in the service process also poses numerous security risks, such as adversarial attacks [3–5], poisoning attacks [6], and backdoor attacks. Among these risks, backdoor attacks against DNNs have emerged as a significant threat. Typically, backdoor attackers aim to embed hidden backdoors in DNNs by poisoning a subset of the training data, which is then mixed with the remaining benign samples and used for standard training. As a result, the victim model behaves normally on benign samples, while the predictions are maliciously and consistently changed if hidden backdoors are activated by attacker-specified trigger patterns. Previous research on backdoor attacks has mainly focused on the image domain, achieving high attack success rates and imperceptibility [7,8]. However, there have been limited studies on backdoor attacks in the audio domain, particularly SI. Existing studies of audio backdoor attacks have mainly focused on Speaker Verification (SV) [9] and Speech Recognition (SR) models [10–12]. Similar to the image domain, researchers in the audio domain have also begun to focus on improving trigger imperceptibility in audio backdoor attacks [10,12]. For instance, some studies have used ultrasonic triggers to activate SR misbehavior in daily use [10]. While these methods achieve high attack success rates, they can easily be detected and mitigated by pre-processing or arousing user suspicion due to the need for additional nearby devices. Another approach for backdoor attacks on SR models is to use ambient noises, such as bird sounds or music, as triggers to improve the imperceptibility of the attack [12]. But in reality, the triggers are not inaudible, and users can hear them and become suspicious, necessitating moderate measures. In summary, the existing imperceptible trigger patterns for backdoor attacks in the audio domain still have flaws.

To address these limitations, we explore a novel imperceptible backdoor attack method for SI. In addition to the necessary requirements

* Corresponding author.
E-mail addresses: 2022040516@njupt.edu.cn (Y. Tang), sunlj@njupt.edu.cn (L. Sun), xuxl@njupt.edu.cn (X. Xu).

https://doi.org/10.1016/j.patrec.2023.12.002
Received 21 March 2023; Received in revised form 13 August 2023; Accepted 5 December 2023
Available online 9 December 2023

for backdoor attacks, it is crucial to satisfy an important additional requirement in order to improve the imperceptibility of backdoor attacks: *poisoned audio samples containing triggers should be acoustically indistinguishable from their corresponding trigger-free counterparts*.

A key question is how to create such indistinguishable and effective triggers. Inspired by audio steganography [13], we propose a novel approach for generating and hiding triggers. Specifically, we utilize an encoder–decoder network to hide triggers into benign audio samples and generate indistinguishable poisoned samples. Through two stages of joint optimization training, we embed the backdoor into the victim model, ultimately establishing a strong mapping from the trigger to the malicious target label. We call our method **SilentTrig**.

Our contributions can be summarized as follows:

- We propose a novel backdoor attack method for deep learning-based speaker identification models, which is a relatively under-explored area in this field. Extensive experiments were conducted to evaluate our proposed method, and the results demonstrate that it achieves high attack effectiveness and imperceptibility. Additionally, our method exhibits the capability to resist typical defense methods.
- We propose an optimized audio steganographic network to overcome the limitations of imperceptibility in existing audio backdoor attacks. By embedding trigger patterns into benign audio samples using our network, we can generate poisoned samples that are acoustically indistinguishable from their corresponding trigger-free counterparts, avoiding detection during model validation and minimizing user suspicion.
- We propose a two-stage joint optimization training approach to ensure the imperceptibility of the trigger and the effectiveness of the attack. This approach establishes a robust mapping between the victim model and the imperceptible trigger.

2. Related works and motivation

2.1. Speaker identification

In the audio domain, there are many behavioral biometric systems utilized for personal identification, such as Speech Recognition (SR), Speaker Verification (SV), and Speaker Identification (SI). Among them, SI is a crucial technique that involves identifying the speaker of an utterance from a pre-recorded enrollment database. Traditional SI models, such as the GMM-UBM/i-vector frontend [14] with PLDA backend, have provided state-of-the-art performance for many years.

However, in recent years, deep learning-based SI methods have emerged [15–18], driven by the powerful feature extraction capabilities of DNNs. These methods, including d-vector and x-vector, significantly improve the performance of SI to a new level. D-vector, proposed by Variani et al. [15], is a speaker embedding technique that learns a fixed-length vector representation of speaker identity from variable-length speech segments using DNNs. While x-vector [16] is an evolution of d-vector that aggregates frame-level embeddings into an utterance-level embedding using a time-delay neural network. Both d-vector and x-vector extract Mel-scale Frequency Cepstral Coefficients (MFCCs) [19] features from utterances and have become widely-used and representative SI models.

2.2. Backdoor attack

Image domain. Backdoor attacks first emerged in the field of image recognition and have achieved many promising results. Gu et al. [20] were the first to reveal the backdoor threat in DNNs training with their BadNets model, which is a representative method of backdoor attacks in the image domain. BadNets works by poisoning a portion of the training images with a backdoor trigger (e.g., a 3 × 3 white square in the lower right corner of the image) to achieve a specified target label. Subsequently, some similar studies have also been proposed [21, 22]. Although directly pasting triggers may result in an effective result, it can also raise suspicion from users during the inference stage. Therefore, embedding triggers with imperceptibility has become a new research hotspot, as discussed in recent studies [7,8,23].

Audio domain and Limitations. Compared with the image domain, there is not much research on backdoor attacks in the audio domain. Existing studies mainly focus on specific aspects such as SR [10–12] and SV [9], while to our knowledge, there are very few studies on SI in this field. Zhai et al. [9] introduced a clustering-based attack scheme for infecting speaker verification models. On the other hand, Ye et al. [11] proposed a dynamic trigger generation network specially designed for Speech Recognition to create a wide variety of audio triggers. However, the spectrum analysis shows that both of these methods overlay triggers on benign audio samples to generate poisoned samples, making them highly noticeable to users. Subsequent research on backdoor attacks in the audio domain has also focused on the imperceptibility of triggers, similar to the image domain. Stefanos et al. [10] adopted ultrasonic pulses as triggers for speech recognition models to enhance the imperceptibility of backdoor attacks. However, ultrasonic signals can be easily filtered out with first-order low-pass filters in the data preprocessing stage. Liu et al. [12] proposed a method called opportunistic backdoor attacks using daily ambient noise, arguing that based on people's auditory inertia, this attack is naturally imperceptible and thus easily overlooked by systems and users. While this method achieves a certain level of imperceptibility, the difference between their poisoned samples and benign counterparts is still noticeable, raising suspicion and rendering it ineffective for an imperceptible backdoor attack after several attempts.

2.3. Motivation

Based on the aforementioned analysis and to overcome the identified limitations, our exploration of this method is driven by two main motivations.

Firstly, there is a significant scarcity of research on audio backdoor attacks, particularly in the domain of SI. Existing studies in this area are limited, and we aim to contribute to the current body of knowledge by conducting comprehensive research and expanding backdoor attack algorithms specifically tailored for the audio domain. Through this endeavor, we intend to provide valuable insights and novel perspectives for future related studies in this field.

Secondly, the existing imperceptible audio backdoor attack algorithms lack true inaudibility to the human ear. While they achieve a certain level of imperceptibility under specific conditions, their robustness is limited when subjected to changes in the acoustic environment. Consequently, we are motivated to explore a more robust and powerful imperceptible backdoor attack algorithm that can genuinely evade human auditory detection under a wide range of circumstances.

3. Preliminaries

3.1. Problem formulation

3.1.1. Deep learning model in SI

A deep learning model in speaker identification system can be modeled as a mapping function $f_\theta(\cdot)$, where θ is the model parameters. This function takes audio waveforms as inputs and outputs speaker labels, such as "Elon Musk".

During the training of the end-to-end SI models, we obtain the trained model parameters $\hat{\theta}$ by several epochs for optimizing the parameters with true labels and prediction results. The optimization objective of the classification model can be represented as follows:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta\left(\boldsymbol{x}_i\right), y_i\right) \tag{1}$$

where $\mathcal{L}(\cdot)$ is the loss function, $\boldsymbol{x}_i$ and y_i represent the ith audio samples and its corresponding label for speaker in the initial training dataset $D = \left\{\left(\boldsymbol{x}_i, y_i\right)\right\}_{i=1}^{N}$, which contains N audio samples.

3.1.2. Backdoor learning in SI

In our imperceptible backdoor attack against SI, we utilize a two-stage optimization process to obtain a trained backdoored model $f_{\hat{\theta}}(\cdot)$. This model behaves normally on benign samples but classifies samples with hidden triggers to a specified target label y_t.

To train the backdoored model, in the first stage, the attacker poisons a small subset D_s of the initial training set D by injecting the trigger into the audio waveforms to generate poisoned set D_p through the trigger generation network. In the second stage, the attacker further optimizes the victim model.

During the entire process, we denote θ^* and $\hat{\theta}$ as the temporarily optimized parameters and the final optimized parameters of the victim model in the two stages, respectively. The trigger generation process is defined as a transformation function $T_\xi(\cdot)$, where ξ represents the parameters for the generation model. The poisoned set D_p contains N_p generated poisoned audio samples, and it is combined with the remaining unaltered set $D_c = D \backslash D_s$ containing $N_c = N - N_p$ samples, resulting in the mixed training set $\hat{D} = \left\{\left(T_\xi(\boldsymbol{x}), y_t\right) \in D_p\right\} \cup \left\{(\boldsymbol{x}, y) \in D_c\right\}$. Additionally, $\epsilon = \left|D_p\right| / |D|$ is dubbed the poisoning rate, which is a critical hyperparameter in our attack. Understandably, the smaller the poisoning rate, the more imperceptible the attack will be.

3.2. Threat model and our goals

3.2.1. Threat model

Attack Scenarios. Due to the substantial data and resource requirements, many individuals and enterprise users opt to delegate the training of their SI models to MLaaS providers. These users provide the model architecture and dataset for training. Once the training is completed, the users test the trained model using a validation dataset. If the classification performance meets the user's expectations in terms of accuracy, it is accepted; otherwise, it is not approved.

Attacker's Capability. We assume that SilentTrig occurs in the context of outsourced training, where the attacker is an employee of MLaaS provider. Similar to state-of-the-art backdoor attack methods in the audio [12] and image domains [24], we assume that the attacker can access and modify the training datasets, labels, model's parameters, and training settings.

3.2.2. Attacker's goals

The attacker's ultimate goal is to train a backdoored model that produces attacker-desired predictions when the input data contains a backdoor trigger, while ensuring the trigger remains imperceptible. We divide this goal into three aspects: effectiveness, imperceptibility, and sustainability. (1) *Effectiveness*: The victim model should establish a strong mapping with the trigger to achieve a high attack success rate, which is a basic requirement for backdoor attacks. (2) *Imperceptibility*: The victim model must behave normally on benign samples to avoid arousing suspicion. Moreover, it is crucial that the poisoned audio samples should be acoustically indistinguishable from their corresponding benign counterparts. (3) Sustainability: The proposed method needs to effectively resist some of the classic backdoor defenses.

4. The proposed SilentTrig

4.1. Overview

To achieve the above goals, we have designed a novel backdoor attack method against SI models, as illustrated in Fig. 1. The entire flow involves three stages: (1) Trigger Generation Stage, (2) Deep Injection Stage, and (3) Inference Stage.

Specifically, drawing inspiration from audio steganography [13], we jointly optimize the trigger generation network $T_\xi(\cdot)$ and the target classification model $f_\theta(\cdot)$ to obtain θ^* and $\hat{\xi}$ in the first stage. Through this process, the attacker can stealthily embed a backdoor into the victim model with a hidden trigger, and poisoned samples and their corresponding benign counterparts cannot be distinguished. Subsequently, the trigger generation model optimized previously is utilized to generate trigger patterns on D_s, and further optimized on the victim model $f_{\theta^*}(\cdot)$ along with clean samples set D_c and obtain the final victim model parameters $\hat{\theta}$ after the second stage, reinforcing the mapping relationship between the victim model and the trigger. Finally, in the inference stage, the trained backdoored model behaves normally on benign samples while predicting the target label when triggered by the poisoned audio samples.

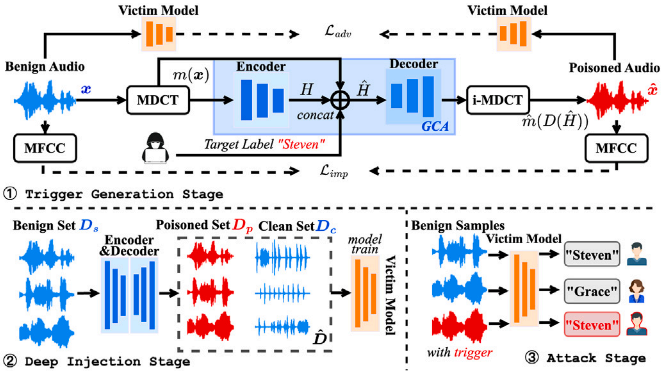


Fig. 1. The overall flow of SilentTrig, which mainly consists of Trigger Generation, Deep Injection, and Inference Stage.

In the subsequent two subsections, we will provide detailed explanations of the first two stages, which are fundamental to our method of trigger generation and backdoor embedding. These stages play a critical role in creating imperceptible and effective backdoor attacks in the audio domain.

4.2. Trigger generation stage

4.2.1. Trigger generation network

To achieve the imperceptibility of the trigger, we employ audio steganography [13] to hide the triggers in the audio samples and generate indistinguishable poisoned samples. Similar to [13], we adopt a frequency-domain Gated Convolutional Autoencoder (GCA) [25] as the steganography network, consisting of an encoder and a decoder. The encoder and decoder are comprised of 3 and 4 gated convolutional layers, respectively. Each gated convolutional layer comprises 64 3 × 3 convolutional kernels, in addition to a batch normalization and dropout layer. Next, we use the aforementioned steganography network as the trigger generation network $T_\xi(\cdot)$, hide the imperceptible trigger in benign audio samples, and then generate poisoned samples.

To begin, we apply the Modified Discrete Cosine Transform (MDCT) [26] to the original benign audio sample $\boldsymbol{x}$, and output $m(\boldsymbol{x})$. The encoder $E(\cdot)$ then takes $m(\boldsymbol{x})$ and produces a latent representation of the spectral content, $H = E(m(\boldsymbol{x}))$. Then, the representation is combined with the original input using a skip connection to obtain the joint representation, $\hat{H} = [E(m(x)); m(x)]$. The decoder then takes in $\hat{H}$ and outputs the MDCT of the poisoned samples, denoted as $D(\hat{H})$. Finally, we use the inverse MDCT, denoted as $\hat{m}(\cdot)$, to take in $D(\hat{H})$ and reconstruct the poisoned audio sample, which is referred to as $\hat{\boldsymbol{x}} = \hat{m}(D(\hat{H}))$.

We need to emphasize that, instead of using the short-time Fourier transform (STFT) [13], the use of MDCT overcomes issues with decoding output reconstruction errors in steganographic operations [27].

4.2.2. Loss function design

Imperceptible loss. In this stage, our main challenge is to effectively use the above steganography network $T(\cdot)$ with data transformation to hide triggers and generate indistinguishable poisoned samples. To address this challenge, we propose the *Imperceptible loss* function $\mathcal{L}_{imp}$, which guides the trigger generation network to continuously optimize model parameters and generates poisoned samples that are indistinguishable from their corresponding benign counterparts, thereby creating a perturbation that is imperceptible to the human ear.

$$\mathcal{L}_{imp} = \sum_{k=1}^{K} 1 - \text{Cos}\left(\text{mfcc}_k(x), \text{mfcc}_k\left(T_{\hat{\xi}}(x)\right)\right) \tag{2}$$

In the above Equation, $mfcc_k(\cdot)$ represents the mel-frequency cepstral coefficients (MFCCs) [19] of the audio sample at time frame k. MFCCs are commonly used as speech features in SR and SI models. $Cos(\cdot)$ computes the cosine similarity between the MFCCs of two audio samples, which is a metric commonly used to measure the similarity between two real spectrogram structures [28].

Adversarial loss. If we only use $\mathcal{L}_{imp}$, the trigger generation network will keep making the poisoned samples closer to their benign counterparts, and may eventually discard the triggers and output the original benign samples. To address this issue, we define an *Adversarial loss* function $\mathcal{L}_{adv}$ and combine it with $\mathcal{L}_{imp}$ to design a new objective function for leading the optimizing process, as follows:

$$\mathcal{L}_{adv} = \mathcal{L}\left(f_\theta\left(T_{\hat{\xi}}\left(x_i\right)\right), y_t\right) \tag{3}$$

$$\hat{\xi} = \arg\min_{\xi} \sum_{i=1}^{N_p} \alpha\mathcal{L}_{adv} + (1-\alpha)\mathcal{L}_{imp} \tag{4}$$

where $\alpha \in [0,1]$ controls the strength of the trigger's imperceptibility, and its default value is set to 0.5. With the adoption of this adversarial approach, the trigger generation model is able to continuously optimize and enhance the imperceptibility of the trigger, resulting in very little difference between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. However, the model is also constrained by adversarial limitations and must maintain the effectiveness of the poisoned audio samples, which contain specific triggers and have an absolute difference from their benign counterparts.

4.2.3. Joint optimization

However, if the discriminative ability of the model $f(\cdot)$ does not improve, $\hat{\xi}$ will not reach its optimal value, indicating that the trigger's imperceptibility and effectiveness have not been fully optimized. Thus, we have further defined the following approach to optimize θ for the target model $f(\cdot)$.

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N_p} \beta\mathcal{L}\left(f_\theta\left(\boldsymbol{x}_i\right), y_i\right) + (1-\beta)\mathcal{L}\left(f_\theta\left(T_\xi\left(\boldsymbol{x}_i\right)\right), y_t\right) \tag{5}$$

where the $\beta \in [0,1]$ controls the final classification performance of the victim model on the benign audio samples. Our experiments show that when $\beta > 0.5$, the victim model's classification performance on the benign sample converges first, whereas poisoned samples with triggers converge more quickly. We set $\beta = 0.5$ by default and jointly optimize Eqs. (4) and (5) to obtain $\hat{\xi}$ and θ^* at the end of this stage.

Such an approach at this stage has several advantages. Firstly, $T(\cdot)$ embeds inaudible sample-specific triggers into benign audio samples, making the triggers imperceptible to users during the attack. Secondly, although the poisoned samples are highly similar to the corresponding benign counterparts, the trigger's effectiveness guarantees the success rate of the attack. Finally, the attacker does not need to input any initial noise as the trigger pattern, but instead learns inaudible triggers during the model training process, further enhancing the imperceptibility of our attacks.

4.3. Backdoor deep injection stage

In the current stage, we further optimize the θ^* obtained from the previous stage and generate the final victim model $f_{\hat{\theta}}(\cdot)$. Specifically, we define the clean sample loss function $\mathcal{L}_{cle}$ and poisoned sample loss function $\mathcal{L}_{poi}$, and train the victim model to its final state on the mixed training set $\hat{D}$ as follows:

$$\mathcal{L}_{cle} = \mathcal{L}\left(f^*_\theta\left(\boldsymbol{x}_j\right), y_j\right), (\boldsymbol{x}_j, y_j) \in D_c \tag{6}$$

$$\mathcal{L}_{poi} = \mathcal{L}\left(f^*_\theta\left(T_{\hat{\xi}}\left(\boldsymbol{x}_i\right)\right), y_t\right), (\boldsymbol{x}_i, y_i) \in D_s \tag{7}$$

The process of embedding the backdoor into the victim model enhanced by benign and poisoned samples can be expressed as follows:

$$\hat{\theta} = \arg\min_{\theta^*} \sum_{j=1}^{N_c} \lambda\mathcal{L}_{cle} + \arg\min_{\theta^*} \sum_{i=1}^{N_p} (1-\lambda)\mathcal{L}_{poi} \tag{8}$$

where the λ is set to adjust the classification performance of the final victim model, similar to Eq. (5).

By optimizing further in this stage, we can deeply embed the backdoor into the victim model, which will establish a strong mapping relationship between the imperceptible hidden trigger and the victim model, resulting in good attack performance during the inference stage.

The overall algorithmic flow of the proposed SilentTrig is shown in Algorithm 1.

Algorithm 1 The Proposed SilentTrig

```
Input: training dataset $D, D_s, D_c$, model $T_\xi, f_\theta$, target label $y_t$,
    parameters $\alpha, \beta, \lambda$, poisoning rate $\epsilon$, learning rate $\gamma_f, \gamma_T$
Output: victim model $f_{\hat{\theta}}(\cdot)$, trigger generation model $T_{\hat{\xi}}(\cdot)$
Initialization: $\theta, \xi$
for number of epoch do
    for each mini-batch from $(\boldsymbol{x}_i, y_i) \in D_s$ do
        Calculate $\mathcal{L}_{imp}$ using Equation (2)
        Calculate $\mathcal{L}_{adv}$ using Equation (3)
        Update $\hat{\xi}$ using Equation (4)
        Update $\theta^*$ using Equation (5)
    end for
    $\xi \leftarrow \hat{\xi}$
end for
for number of epoch do
    for $(\boldsymbol{x}_j, y_j) \in D_c$ and $(\boldsymbol{x}_i, y_i) \in D_s$ do
        Calculate $\mathcal{L}_{cle}$ using Equation (6)
        Calculate $\mathcal{L}_{poi}$ using Equation (7)
        Update $\hat{\theta}$ using Equation (8)
    end for
end for
```

5. Experiments

5.1. Experimental settings

5.1.1. Datasets and models

We employed the VoxCeleb [29] and TIMIT [30] datasets, consistent with previous studies [6]. VoxCeleb comprises more than 100,000 speech segments from 1251 celebrities, extracted from YouTube videos. It is well-balanced, with 55% male and 45% female speakers, and covers a wide range of races, accents, professions, and ages. TIMIT includes audio recordings from 630 speakers, with each speaker having 10 speech segments. We randomly selected 250 speakers from VoxCeleb and TIMIT separately, with each speaker having 10 speech segments that were downsampled to 8 kHz. Each dataset was divided into a 90% training set and a 10% testing set.

In our experiments, we follow previous works [9] and use d-vector and x-vector as the primary models. The d-vector [15] is a classical

DNN-based embedding feature, widely used and extended in various speaker recognition tasks. On the other hand, the x-vector [16] is the fundamental framework for speaker identification tasks. It starts by extracting MFCCs from speech signals and then utilizes a Time Delay Neural Network (TDNN) to generate speaker feature embeddings. Additionally, we incorporate TDNN-LSTM [31] and ResNet-based [32] speaker embedding models for the attack testing. The TDNN-LSTM model, also known as L-vector, modifies the original TDNN-based x-vector by replacing two TDNN layers with an LSTM layer. ResNet is another widely used architecture in speaker embedding, alongside x-vector, and it utilizes a 2-dimensional CNN with convolutions in both the time and frequency domains [1].

5.1.2. Baseline

The well-known backdoor attack method BadNets [20], originally developed for the visual domain, has been applied to audio-based backdoor attack methods in existing literature [9,11]. In our experiments, we also use BadNets as a baseline method. In particular, BadNets embeds static triggers into benign samples to produce poisoned samples with a designated poisoning rate. These poisoned samples are combined with the remaining benign samples and utilized for standard training of the target classification model to generate the final victim model.

5.1.3. Attacking setup

We implemented our method using the PyTorch framework and trained our models on an Nvidia-RTX 3090 GPU. For the trigger generation phase, we used the Adam optimizer with a learning rate of $lr = 1e^{-3}$, following [13]. Additionally, we set $\alpha = 0.5$, $\beta = 0.5$, $\lambda = 0.6$ and the poisoning rate $\epsilon = 10\%$. For the deep injection phase, we also set the learning rate to $lr = 1e^{-3}$, and the number of epochs for the two phases was set to 80 and 100, respectively. Furthermore, we have also made targeted adjustments to some parameters in the experiments to observe changes in performance.

5.1.4. Metrics

We evaluate the proposed method using three metrics, and the first two metrics are commonly used indicators for evaluating backdoor attacks:

Attack Success Rate (ASR). It represents the percentage of poisoned samples with a trigger that the victim model successfully classifies as the target label. A higher ASR indicates better effectiveness of the backdoor attack.

Clean Data Accuracy (CDA). This metric indicates the percentage of correctly classified audio samples without the presence of the trigger. A successful backdoor attack should retain the vanilla model's performance on the classification of benign samples.

Just Noticeable Difference (JND). To enhance the imperceptibility of the attack, in addition to achieving the high CDA metric, we aim to get indistinguishable poisoned samples, which is the primary motivation for proposing our method. JND [33] is the minimum signal variation that can be perceived by humans and has commonly been used to measure speech similarity [34]. It is defined as the l_1 norm between two audio files, and the value ranges from 0.0 to 5.0. The smaller the value, the higher the similarity between the audio, and vice versa. In our experiments, we use JND to measure the difference between poisoned samples and corresponding benign counterparts.

5.2. Main results

5.2.1. Effectiveness of SilentTrig

Table 1 presents a comparison of the attack effectiveness between BadNets and SilentTrig. It can be observed that SilentTrig achieved the highest ASR of 99.2% with a small poisoning rate of 10%. Compared to BadNets, it is at most 5.7 percentage points higher on this metric.

Table 1
CDA(%)/ASR(%)/JND of BadNets (baseline) [20] and SilentTrig (our method) on VoxCeleb [29] and TIMIT [30] datasets against four models: d-vector [15], x-vector [16], l-vector [31], and ResNet [32]. The term "vanilla" represents the original clean model without any attack. The CDA (Clean Data Accuracy) reflects the percentage of accurately classified audio samples in the absence of any trigger. The ASR (Attack Success Rate) represents the success rate of the poisoned test samples with triggers in causing misclassifications by the victim model. The JND [33] (Just Noticeable Difference) measures the difference between poisoned samples and corresponding benign counterparts. Poisoning rate $\epsilon = 10\%$, and the table shows the best performance.

Dataset		VoxCeleb			TIMIT		
Model	Method	CDA(%)	ASR(%)	JND	CDA(%)	ASR(%)	JND
D-vector	vanilla	97.5	–	–	95.2	–	–
	BadNets	91.1	92.7	2.96	91.1	92.2	3.54
	SilentTrig	94.2	98.4	0.59	92.1	97.9	0.66
X-vector	vanilla	98.1	–	–	96.4	–	–
	BadNets	93.5	95.6	2.71	92.1	96.6	3.11
	SilentTrig	94.5	99.2	0.33	93.9	98.2	0.58
L-vector	vanilla	94.9	–	–	95.1	–	–
	BadNets	91.8	92.6	2.89	91.0	92.3	3.31
	SilentTrig	92.9	98.1	0.47	93.0	97.1	0.42
ResNet	vanilla	96.1	–	–	96.1	–	–
	BadNets	91.1	94.3	2.88	92.7	96.1	3.43
	SilentTrig	92.4	96.7	0.38	94.9	95.9	0.61

5.2.2. Influence on pristine performance

The influence on pristine performance is a critical aspect for evaluating the effectiveness and imperceptibility of the backdoor attack. Due to the existence of poisoned samples in the training data, the resulting backdoored victim model often demonstrates different levels of influence on the classification accuracy of clean test samples, as shown in Table 1.

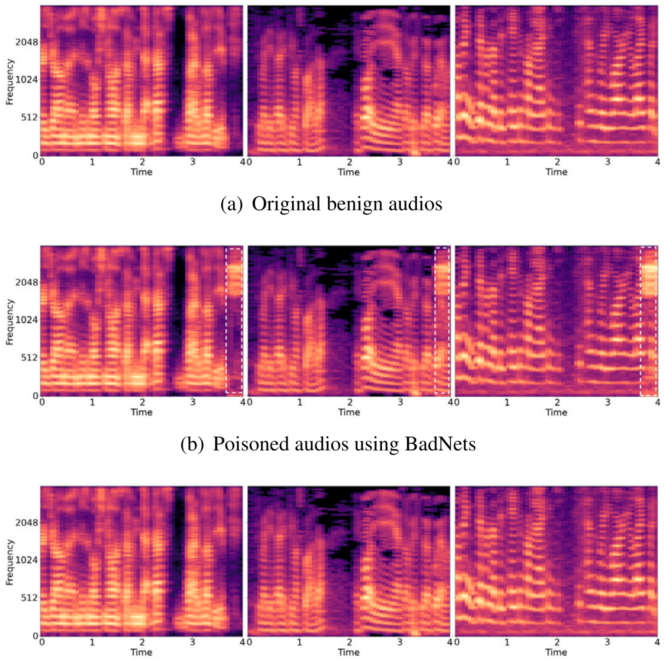
The results show that the CDA of SilentTrig and BadNets, as well as the CDA of the vanilla model indicating the performance of the clean model in the absence of an attack. The maximum difference of CDA of SilentTrig compared to vanilla (clean model without attack) accuracy is 3.7%, which is superior to BadNets (maximum difference of 6.4%). This indicates that SilentTrig has a minimal impact on the pristine performance of the victim model, making it more imperceptible when attacking.

5.2.3. Imperceptibility of SilentTrig

Undoubtedly, the poisoning rate and CDA performance may have an impact on the imperceptibility of the backdoor attack. However, in our experiments, the JND is the most critical metric for evaluating imperceptibility. As shown in Table 1, we observe that the JND between the poisoned audio samples generated by SilentTrig and their corresponding benign counterparts has a maximum value of 0.33, indicating that it is imperceptible to humans whether the poisoned audio samples contain a trigger or not. Although BadNets achieves decent ASR and CDA results, its JND value is much higher than that of SilentTrig. This means that users can easily distinguish the differences in poisoned samples and become suspicious of the victim model, which may ultimately lead to attack failure. Furthermore, we present a visual comparison of the poisoned samples generated by SilentTrig and BadNets in Fig. 2. It is evident that the poisoned samples in BadNets exhibit noticeable trigger patterns on the right side of the spectrum, whereas those generated by SilentTrig are almost indistinguishable from the original samples. Several existing algorithms in the audio domain use a way similar to BadNets to create triggers that are easily detectable, resulting in noticeable JND values [9,11,12].

We have released some demos for listening to experience the indistinguishability between the poisoned audio and its corresponding benign counterparts.[1]

[1] https://drive.google.com/drive/folders/1oub78IAEnnt34F7YIN6DbjezF4vupqGp?usp=sharing

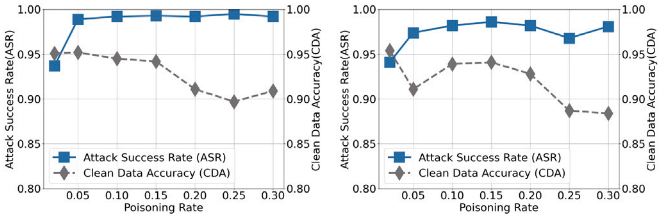


(a) Original benign audios

(b) Poisoned audios using BadNets

(c) Poisoned audios using SilentTrig

Fig. 2. The visual comparison of poisoned audio samples generated by BadNets and SilentTrig methods. (a) shows the Mel spectrogram of three randomly selected original benign audio samples, while (b) and (c) show the poisoned samples generated by BadNets and SilentTrig methods, respectively.



(a) VoxCeleb

(b) TIMIT

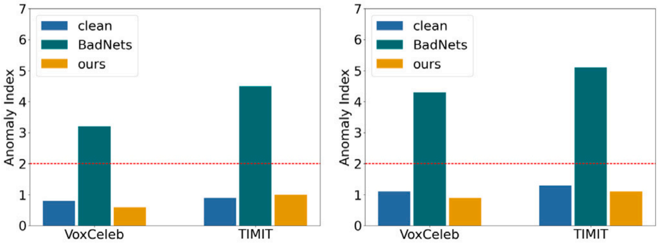
Fig. 3. Influence of poisoning rate on SilentTrig performance of our proposed method.

5.3. Influence of poisoning rate

We evaluated the efficacy of SilentTrig on the x-vector model using two datasets with different poisoning rates. Specifically, we varied the poisoning rates from 2% to 30%. As shown in Fig. 3, the ASR remained at a high level of over 90%, while the CDA exhibited a downward trend as the poisoning rate increased. Once the poisoning rate reached 10%, the relevant indicators reached a high level, and a further increase in the poisoning rate did not result in any significant improvement. Therefore, we conclude that SilentTrig can achieve good attack performance even with a small portion of the dataset.

5.4. Resistance to defense

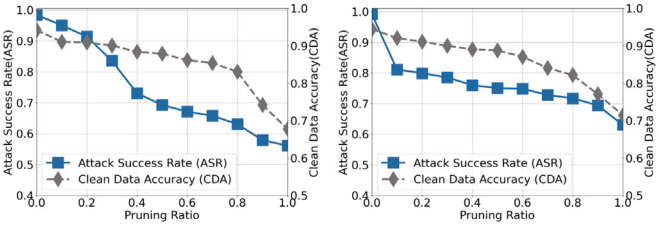
Similar to research on the attack, existing backdoor defense methods have mainly focused on the image domain, and most of them cannot be directly applied to the audio domain. In most previous research on audio backdoor attacks, the authors did not test the resistance to defense methods in their experiments [9,10], and only a small number of studies have done this work [11,12]. To verify the sustainability of SilentTrig, we follow their approach.



(a) d-vector

(b) x-vector

Fig. 4. Neural-Cleanse on two datasets against two models for BadNets and SilentTrig. The smaller the anomaly index, the harder the attack for Neural-Cleanse to defend.



(a) d-vector

(b) x-vector

Fig. 5. Fine-Pruning on VoxCeleb dataset against two models for SilentTrig. Both the ASR and CDA of SilentTrig decrease with the increase in neuron pruning ratio.

5.4.1. Resistance to neural cleanse

Neural Cleanse [35] computes trigger candidates to convert all benign samples to each label and adopts an anomaly detector to verify if any trigger is significantly smaller than the others, serving as a backdoor indicator. The smaller the anomaly index value, the more difficult it is for Neural Cleanse to defend against the attack. When the anomaly index falls below a certain threshold, the attack is considered successful. As shown in Fig. 4, the anomaly index of SilentTrig is much lower than that of BadNets and falls below the threshold (red line), indicating that the defense cannot detect the backdoor trigger.

5.4.2. Resistance to fine-pruning

Similar to the methodology proposed in the previous study [12], we evaluated the effectiveness of SilentTrig against another state-of-the-art backdoor defense method known as Fine-Pruning [36]. The results are depicted in Fig. 5, and we observed that as the pruning ratio of neurons increases, the ASR does not reach its minimum level, and the CDA exhibits a substantial decline. This implies that the defense method is incapable of distinguishing between the backdoor neurons and the clean neurons, which further confirms the resilience of the SilentTrig against Fine-Pruning.

6. Conclusions

In summary, this paper introduced a novel backdoor attack method for speaker identification models called SilentTrig, which addresses the limitations of existing deep learning-based audio backdoor attacks in terms of trigger imperceptibility. SilentTrig embeds the backdoor into benign samples through audio steganography and generates poisoned samples, which are indistinguishable from the corresponding counterparts through two-stage adversarial optimization, achieving imperceptible backdoor attacks. The effectiveness, imperceptibility, and sustainability of SilentTrig were validated through a series of experiments on different models and datasets with different poisoning rates.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62272242.

References

[1] Z. Bai, X.-L. Zhang, Speaker recognition based on deep learning: An overview, Neural Netw. 140 (2021) 65–99.

[2] Y. Li, Y. Li, B. Wu, L. Li, R. He, S. Lyu, Invisible backdoor attack with sample-specific triggers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16463–16472.

[3] C. Bisogni, L. Cascone, J.-L. Dugelay, C. Pero, Adversarial attacks through architectures and spectra in face recognition, Pattern Recognit. Lett. 147 (2021) 55–62.

[4] C. Ying, Y. Qiaoben, X. Zhou, H. Su, W. Ding, J. Ai, Consistent attack: Universal adversarial perturbation on embodied vision navigation, Pattern Recognit. Lett. 168 (2023) 57–63.

[5] Z. Yin, L. Chen, W. Lyu, B. Luo, Reversible attack based on adversarial perturbation and reversible data hiding in YUV colorspace, Pattern Recognit. Lett. 166 (2023) 1–7.

[6] X. Cao, Z. Zhang, J. Jia, N.Z. Gong, Flcert: Provably secure federated learning against poisoning attacks, IEEE Trans. Inf. Forensics Secur. 17 (2022) 3691–3705.

[7] Y. Liu, X. Ma, J. Bailey, F. Lu, Reflection backdoor: A natural backdoor attack on deep neural networks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X, Vol. 16, Springer, 2020, pp. 182–199.

[8] W. Guo, B. Tondi, M. Barni, A Master Key backdoor for universal impersonation attack against DNN-based face verification, Pattern Recognit. Lett. 144 (2021) 61–67.

[9] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, S.-T. Xia, Backdoor attack against speaker verification, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 2560–2564.

[10] S. Koffas, J. Xu, M. Conti, S. Picek, Can you hear it? backdoor attacks via ultrasonic triggers, 2021, arXiv preprint arXiv:2107.14569.

[11] J. Ye, X. Liu, Z. You, G. Li, B. Liu, DriNet: dynamic backdoor attack against automatic speech recognization models, Appl. Sci. 12 (12) (2022) 5786.

[12] Q. Liu, T. Zhou, Z. Cai, Y. Tang, Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 2390–2398.

[13] F. Kreuk, Y. Adi, B. Raj, R. Singh, J. Keshet, Hide and speak: Towards deep neural networks for speech steganography, 2019, arXiv preprint arXiv:1902.03083.

[14] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2010) 788–798.

[15] E. Variani, X. Lei, E. McDermott, I.L. Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2014, pp. 4052–4056.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 5329–5333.

[17] H.N. Pinheiro, T.I. Ren, A.G. Adami, G.D. Cavalcanti, Variational DNN embeddings for text-independent speaker verification, Pattern Recognit. Lett. 148 (2021) 100–106.

[18] Y. Liu, L.-F. Wei, C.-F. Zhang, T.-H. Zhang, S.-L. Chen, X.-C. Yin, Self-supervised contrastive speaker verification with nearest neighbor positive instances, Pattern Recognit. Lett. (2023).

[19] S. Sigurdsson, K.B. Petersen, T. Lehn-Schiøler, Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music, in: ISMIR, 2006, pp. 286–289.

[20] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, Badnets: Evaluating backdooring attacks on deep neural networks, IEEE Access 7 (2019) 47230–47244.

[21] T.A. Nguyen, A. Tran, Input-aware dynamic backdoor attack, Adv. Neural Inf. Process. Syst. 33 (2020) 3454–3464.

[22] J. Lin, L. Xu, Y. Liu, X. Zhang, Composite backdoor attack for deep neural network by mixing existing benign features, in: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020, pp. 113–131.

[23] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, Y.-G. Jiang, Clean-label backdoor attacks on video recognition models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14443–14452.

[24] A. Nguyen, A. Tran, Wanet–imperceptible warping-based backdoor attack, 2021, arXiv preprint arXiv:2102.10369.

[25] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 933–941.

[26] S. Zhang, W. Dou, H. Yang, MDCT sinusoidal analysis for audio signals analysis and processing, IEEE Trans. Audio Speech Lang. Process. 21 (7) (2013) 1403–1414.

[27] R. Ji, X. Cai, B. Xu, An end-to-end text-independent speaker identification system on short utterances, in: Interspeech, 2018, pp. 3628–3632.

[28] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, P. Traynor, Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems, in: 2021 IEEE Symposium on Security and Privacy, SP, IEEE, 2021, pp. 730–747.

[29] A. Nagrani, J.S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, Comput. Speech Lang. 60 (2020) 101027.

[30] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1, Vol. 93, NASA STI/Recon Technical Report N, 1993, p. 27403.

[31] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, C.-L. Huang, Speaker characterization using tdnn-lstm based speaker embedding, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 6211–6215.

[32] Z. Wang, K. Yao, X. Li, S. Fang, Multi-resolution multi-head attention in deep speaker embedding, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 6464–6468.

[33] P. Manocha, A. Finkelstein, R. Zhang, N.J. Bryan, G.J. Mysore, Z. Jin, A differentiable perceptual audio metric learned from just noticeable differences, 2020, arXiv preprint arXiv:2001.04460.

[34] C.K. Reddy, V. Gopal, R. Cutler, DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6493–6497.

[35] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, B.Y. Zhao, Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, in: 2019 IEEE Symposium on Security and Privacy, SP, IEEE, 2019, pp. 707–723.

[36] K. Liu, B. Dolan-Gavitt, S. Garg, Fine-pruning: Defending against backdooring attacks on deep neural networks, in: Research in Attacks, Intrusions, and Defenses: 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings 21, Springer, 2018, pp. 273–294.