Toward Stealthy Backdoor Attacks Against Speech Recognition via Elements of Sound

Hanbo Cai, Pengcheng Zhang, *Member, IEEE*, Hai Dong, *Senior Member, IEEE*, Yan Xiao, Stefanos Koffas, and Yiming Li, *Member, IEEE*

***Abstract*— Deep neural networks (DNNs) have been widely and successfully adopted and deployed in various applications of speech recognition. Recently, a few works revealed that these models are vulnerable to backdoor attacks, where the adversaries can implant malicious prediction behaviors into victim models by poisoning their training process. In this paper, we revisit poison-only backdoor attacks against speech recognition. We reveal that existing methods are not stealthy since their trigger patterns are perceptible to humans or machine detection. This limitation is mostly because their trigger patterns are simple noises or separable and distinctive clips. Motivated by these findings, we propose to exploit elements of sound (*e.g.*, pitch and timbre) to design more stealthy yet effective poison-only backdoor attacks. Specifically, we insert a short-duration high-pitched signal as the trigger and increase the pitch of remaining audio clips to 'mask' it for designing stealthy pitch-based triggers. We manipulate timbre features of victim audio to design the stealthy timbre-based attack and design a voiceprint selection module to facilitate the multi-backdoor attack. Our attacks can generate more 'natural' poisoned samples and therefore are more stealthy. Extensive experiments are conducted on benchmark datasets, which verify the effectiveness of our attacks under different settings (*e.g.*, all-to-one, all-to-all, clean-label, physical, and multi-backdoor settings) and their stealthiness. Our methods achieve attack success rates of over 95% in most cases and are nearly undetectable. The code for reproducing main experiments are available at https://github.com/HanboCai/BadSpeech_SoE.**

***Index Terms*— Backdoor attack, backdoor learning, speech recognition, AI security, trustworthy ML.**

Manuscript received 16 July 2023; revised 29 November 2023 and 6 March 2024; accepted 21 May 2024. Date of publication 23 May 2024; date of current version 31 May 2024. The work was supported in part by the National Natural Science Foundation of China under Grant U21B2016, Grant 62272145, Grant 62025604, and Grant U2336208; and in part by the Shenzhen Science and Technology Program under Grant KQTD20221101093559018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Luisa Verdoliva. *(Corresponding author: Pengcheng Zhang.)*

Hanbo Cai and Pengcheng Zhang are with the College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China (e-mail: caihanbo@hhu.edu.cn; pchzhang@hhu.edu.cn).

Hai Dong is with the School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia (e-mail: hai.dong@rmit.edu.au).

Yan Xiao is with the School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen Campus, Shenzhen 518107, China (e-mail: xiaoy367@mail.sysu.edu.cn).

Stefanos Koffas is with the Cybersecurity Group, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: s.koffas@tudelft.nl).

Yiming Li is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798, and also with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 311200, China (e-mail: liyiming.tech@gmail.com).

Digital Object Identifier 10.1109/TIFS.2024.3404885

I. Introduction

SPEECH recognition has been widely and successfully deployed in many mission-critical applications [1], [2], [3]. In general, obtaining well-performed speech recognition models requires training on large-scale annotated datasets and substantial hardware resources. Accordingly, developers and users usually exploit third-party resources, such as open-source datasets, training platforms, checkpoints, and crowdsourced data collection [4], to alleviate training burdens.

However, recent studies revealed that outsourcing (parts of) training procedures (*e.g.*, data collection) may also introduce new security risks to DNNs [5]. Arguably, the backdoor attack is one of the most emerging yet threatening threats [6]. The backdoor adversaries can implant hidden backdoors to victim DNNs by introducing a few poisoned training samples containing adversary-specified trigger patterns. The adversaries can activate the embedded backdoor via triggers during the inference process of backdoored models to maliciously manipulate their predictions. However, the backdoored models behave normally on benign testing samples. Accordingly, victim users can hardly notice backdoor threats.

Currently, most of the existing backdoor attacks are designed against image or text classification [7], [8], [9], [10], [11], [12]. However, the backdoor analysis in speech recognition is left far behind. In particular, the few feasible attacks in this area are preliminary, whose trigger patterns are simple noises [13], [14], [15], [16], [17] or separable and distinctive audio clips [18], [19], [20]. Accordingly, these attacks are perceptible to humans or can be easily detected and alleviated by algorithms [16], [21]. It raises an intriguing question: *Is it possible to design an effective attack against speech recognition that is stealthy to both human and machine detection*?

The answer to the aforementioned question is positive. Arguably, the core of an effective and stealthy attack is to design more 'natural' trigger patterns. In this paper, we generate more naturally poisoned samples by modifying the elements of sound. We tackle trigger design from two perspectives, including pitch and timbre. Specifically, we first increase the pitch of selected audio samples and then insert a short yet high-pitched signal to generate their poisoned version for the pitch-based attack. The pitch-increased background audio can hide the inserted signal due to audio masking. This method is dubbed pitch boosting and sound masking (PBSM); For the timbre-based attack, we edit the timbre features of selected samples to generate their poisoned counterparts. In particular, we design a voiceprint selection module that enables the selection of diverse timbre features for timbre transformation, to further improve its effectiveness under the multi-backdoor setting. We call this method voiceprint selection and voice conversion (VSVC). The poisoned samples generated by our PBSM and VSVC are natural and sample-specific. As such, they can bypass both human inspection and machine detection.

In conclusion, our main contributions are three-fold:

- We reveal the stealthiness deficiency of existing attacks against speech recognition and its potential reasons.
- We propose two simple yet effective backdoor attacks against speech recognition (*i.e.*, PBSM and VSVC) via elements of sound. The poisoned samples of both PBSM and VSVC are more natural and therefore stealthy to both human inspection and machine detection.
- Extensive experiments are conducted to verify the effectiveness of our attacks under different settings (*e.g.*, all-to-one, all-to-all, clean-label, physical, and multi-backdoor settings) and their resistance to defenses.

The rest of this paper is structured as follows. In Section II, we briefly review related works about speech recognition and backdoor attacks. Section III illustrates our two stealthy backdoor attacks based on elements of sound, *i.e.*, pitch boosting and sound masking (PBSM) and voiceprint selection and voice conversion (VSVC), in detail. The experimental results of our attacks are presented in Section IV. We conclude this paper in Section VI at the end.

II. Related Works

A. Speech Recognition

Speech recognition (SR) plays a vital role in many critical applications [22], allowing devices to comprehend and interpret human speech. Early speech recognition methods were mostly based on Gaussian mixture models (GMMs) and hidden Markov models (HMMs) [23]. However, these methods suffered from relatively high error rates in practice.

Recently, advanced SR methods were all based on deep neural networks (DNNs) due to their high learning capacities. For example, Hinton et al. [24] applied DNNs to acoustic modeling and achieved promising performance in the TIMIT [25] phoneme recognition task, marking a breakthrough in the field of speech recognition with DNNs. De Andrade et al. [26] applied long short-term memory (LSTM) networks in speech recognition tasks, motivated by the strong temporal nature of speech data. Besides, inspired by the tremendous success of ResNet in image classification [27], Vygon and Mikhaylovskiy [28] proposed a novel and effective keyword discovery model with the ResNet backbone. Recently, Berg et al. [29] exploited the Transformer structure in speech recognition and achieved remarkable performance. Gazneli et al. [30] proposed an end-to-end strategy without requiring pre-processing speech data to simplify the speech recognition tasks. Specifically, they adopted one-dimensional convolutional stacks and Transformer-type encoder blocks to process and classify speech data.

B. Backdoor Attacks

Backdoor attack is an emerging yet critical training-phase threat [6]. In general, the adversaries intend to implant hidden backdoors into the victim model by maliciously manipulating the training procedures (*e.g.*, samples or loss). The backdoored model will behave normally on predicting benign testing samples whereas its predictions will be misled to adversary-specified target classes whenever its backdoor is activated by the trigger pattern contained in attacked testing samples.

Currently, most of the existing attacks are designed against image classification. These attacks can be divided into different sub-categories based on different criteria, as follows:

1) Poisoned-Label and Clean-Label Attacks: Backdoor attacks can be divided into poisoned-label [7], [12], [31] and clean-label attacks [32], [33], [34] based on whether the target label of poisoned samples is consistent with their ground-truth one. In general, poisoned-label backdoor attacks are more effective compared to the clean-label ones since the ‘robust features’ related to the target class contained in poisoned samples of clean-label attacks will hinder the learning of trigger patterns [11]. However, clean-label attacks are more stealthy since victim users can identify and filter out poisoned training samples by examining the image-label relationship.

2) All-to-One and All-to-All Attacks: We can separate existing attacks into all-to-one and all-to-all attacks based on the property of the target label [7]. Specifically, all poisoned samples will be assigned the same target label in all-to-one attacks, while the target label of all-to-all attacks is determined based on the ground-truth one of the poisoned samples. For example, the all-to-all adversaries usually adopt $y' = (y + 1) \bmod K$, where K is the number of all classes, y' and y indicate the target label and ground-truth label of the poisoned sample, respectively. Arguably, all existing (poisoned-label) backdoor attacks can be generalized to all-to-all attacks, although it will probably decrease attack effectiveness [6].

3) Single-Backdoor and Multi-Backdoor Attacks: Different from the single-backdoor attacks where the adversaries only implant a single backdoor to the victim models, multi-backdoor methods [7], [35], [36], [37] intend to embed multiple backdoors simultaneously. In general, it is non-trivial to implant multiple backdoors, although we can easily inject a single backdoor. It is mostly because the learning of one backdoor may affect that of the others [37]. As such, multi-backdoor attacks may fail if triggers are not ‘strong’ enough.

4) Digital and Physical Attacks: Different from previous digital attacks where all poisoned samples are obtained completely in the digital space, the physical space is also involved in their generation in the physical attacks. Chen et al. [38] proposed the first physical backdoor attack where they exploited the glasses as physical trigger against facial recognition. A similar idea was also discussed in [39]. Recently, Li et al. [40] revealed that existing digital attacks will fail in the physical space and proposed a physical attack enhancement inspired by the expectation over transformation [41]. Most recently, Xu et al. [42] designed a more stealthy poison-only physical backdoor attack using spatial transformations (*e.g.*, rotation) with a specific parameter as trigger patterns.

Recently, there have also been a few backdoor attacks against speech recognition. Specifically, Liu et al. [13] reversed potential training samples of a given speech recognition model, based on which to implant hidden backdoors; Kong and Zhang et al. [17] designed trigger patterns based on audio steganography; Zhai et al. [14] designed the first backdoor attack against speaker verification via clustering techniques; Koffas et al. exploited ultrasonic pulses as audio triggers; In [18], [19], and [20], sounds from the natural environment (*e.g.*, music and noises) were adopted as trigger patterns; Shi et al. [15] developed an optimization scheme to generate more effective audio triggers; Most recently, a concurrent work [43] designed stealthy style-based triggers for audio backdoor attacks via style transformations. However, all existing attacks are perceptible to humans or can be easily detected and alleviated by algorithms. How to design an effective backdoor attack against speech recognition that is stealthy to both human and machine detection is still an important open question and worth further exploration.

Besides, we also notice that there are also a few works exploited backdoor attacks for positive purposes (*e.g.*, copyright protection and model interpretability) [44], [45], [46]. These works are out of the scope of this paper.

III. The Proposed Methods

The sound elements primarily include pitch, timbre, and loudness [47]. In this paper, we discuss how to design more natural yet effective acoustic trigger patterns based on pitch and timbre, respectively. We omit the loudness-type trigger design since it has minor variation and therefore may not contain sufficient information for effective backdoor attacks.

A. Preliminaries

1) Elements of Sound: The elements of sound consist of pitch, timbre, and loudness [47]. Pitch denotes the perceived frequency of a sound, ascending with higher frequencies. Timbre, the 'color' of sound, is shaped by the harmonic content and the envelope, endowing distinct acoustic identities to different instruments or vocal sources, even at identical pitches and volumes. Loudness, often correlated with the amplitude of a sound, is a perceptual attribute that denotes the perceived strength or intensity of the sound. In neural network audio processing, the models tend to prioritize learning complex features such as pitch and timbre over subtle amplitude variations. This is because neural networks are designed to generalize and often deemphasize simpler variations in favor of more distinct acoustic properties. Accordingly, our discussion primarily focuses on pitch and timbre (instead of loudness), as their intricate variations are crucial for speech recognition.

2) Threat Model: In this paper, we focus on *poison-only* backdoor attacks against speech recognition, where the adversaries can only modify their released poisoned training dataset. The victim users will exploit the poisoned dataset to train their models with user-specified settings. Accordingly, we assume that the adversaries cannot change and have no information on the training process (*e.g.*, model structure, loss, and training schedule). This is one of the most difficult settings for backdoor attacks, with the most expansive threat scenarios (*e.g.*, using third-party samples, training facilities, or models) [6].

TABLE I
The Definition of Common Symbols

Symbol	Description
$\mathcal{D}$	benign dataset
$\hat{\mathcal{D}}$	poisoned dataset
$\mathcal{D}_b$	benign subset
$\mathcal{D}_p$	poisoned subset
G	generator of poisoned samples
$f_{\boldsymbol{\theta}}$	model parametrized by θ
$\mathcal{L}$	loss function
$\boldsymbol{x}$	benign sample
y	the label of benign sample
γ	poisoning rate
t	trigger pattern
N	the number of samples in dataset $\mathcal{D}$
K	the number of categories

3) Adversary's Goals: In summary, the backdoor adversaries have three main goals, including **(1)** effectiveness, **(2)** stealthiness, and **(3)** persistence. Specifically, effectiveness requires that backdoored models can predict poisoned testing samples as the adversary-specified target label, no matter what their ground-truth label is; Stealthiness ensures that the attack cannot be detected by human inspection or simple machine detection. For example, trigger patterns should be stealthy and the poisoning rate should be small; Persistence seeks that the attack is still effective under more difficult settings (*e.g.*, under potential adaptive defenses and physical-world settings).

4) The Main Pipeline of Poison-Only Backdoor Attacks: In general, how to generate the poisoned dataset $\hat{\mathcal{D}}$ given its benign version $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ is the main problem of poison-only backdoor attacks. Considering a classification problem with K-categories, the $\hat{\mathcal{D}}$ contains two separate subsets, including the benign subset $\mathcal{D}_b$ and the poisoned subset $\mathcal{D}_p$ (*i.e.*, $\hat{\mathcal{D}} = \mathcal{D}_b \cup \mathcal{D}_p$). Specifically, $\mathcal{D}_b$ is randomly sampled from $\mathcal{D}$ containing $(1-\gamma)\cdot N$ samples, where γ is dubbed 'poisoning rate'. $\mathcal{D}_p \triangleq \left\{\left(G_x(\boldsymbol{x}), G_y(y)\right) | (\boldsymbol{x}, y) \in \mathcal{D}\backslash\mathcal{D}_b\right\}$, where $G_x : \mathcal{X} \rightarrow \mathcal{X}$ and $G_y : \mathcal{Y} \rightarrow \mathcal{Y}$ are adversary-assigned poisoned instance generator and poisoned label generator, respectively. For example, $G_x(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{t}$ where $\boldsymbol{t}$ is the trigger based on additive noises [48]; $G_y(y) = y_T$ where y_T is the target label in all-to-one attacks [6], $G_y(y) = (y+1) \mod K$ in most of the existing all-to-all attacks [7]. After $\hat{\mathcal{D}}$ is generated and released, the victim users will use it to train their model $f_{\boldsymbol{\theta}} : \mathcal{Y} \rightarrow [0, 1]^K$ via $\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},y)\in\hat{\mathcal{D}}} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$.

5) The Definition of Common Symbols: For convenience, we summarize the commonly used symbols in Table I. We will follow the same definition in the remaining paper.

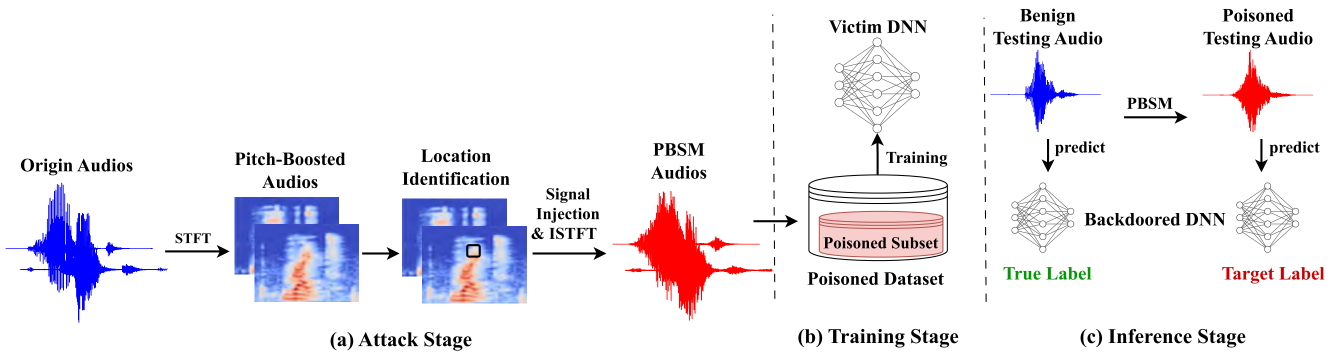


Fig. 1. The main pipeline of attacking via our pitch boosting and sound masking (PBSM). The PBSM consists of three main stages, including attack, training, and inference. The attack stage is the core of PBSM, containing two steps (*i.e.*, pitch boosting and signal injection). In the first step, we exploit short-time Fourier transform to convert the original audio from the time domain to the frequency domain and increase the pitch of the overall audio; In the second step, we identify the position of the highest-amplitude segment in the audio where we insert an adversary-specified high-pitched signal.

B. The Design Philosophy of Our Attacks

Arguably, the core of designing effective and stealthy backdoor attacks against speech recognition is generating more ‘natural’ trigger patterns. As described in Section III-A, the models tend to exploit complex features such as pitch and timbre for classification in neural network audio processing. However, the human auditory system is not sensitive to these changes or does not rely on them for recognition [49], [50]. Inspired by this difference, we propose to generate more naturally poisoned samples by modifying sound elements, including pitch and timbre. We call our methods ‘pitch boosting and sound masking (PBSM)’ and ‘voiceprint selection and voice conversion (VSVC)’. Their design details are as follows.

1) The Design Philosophy of PBSM: Firstly, the human auditory system has a limited ability to distinguish subtle pitch variations, especially in complex auditory environments [51]. Secondly, the human ear is not sensitive to sound masking, meaning that in a complex mix of sounds, subtle changes can be easily overshadowed by other sounds [52]. Accordingly, we can modify the pitch and covertly implant signals via sound masking to generate stealthy yet effective poisoned samples.

2) The Design Philosophy of VSVC: The human auditory system is consistent in its understanding of the same sentence read by different people (with different timbres). However, DNNs may learn and capture timbre characters for predictions. Accordingly, we can manipulate timbre features to generate stealthy yet effective poisoned samples.

The technical details of PBSM and VSVC are in Section III-C and Section III-D, respectively.

C. Attack via Pitch Boosting and Sound Masking

Arguably, the most straightforward approach to designing pitch-type triggers is to insert sound clips with a very high (or low) frequency in a random position of the victim audio. However, these triggers can be easily filtered out by removing clips with the highest and lowest frequencies. Besides, these triggers are also perceptible to humans since the inserted trigger is most likely different from its surrounding audio clips in the poisoned samples. To tackle these problems, in this paper, we propose to first increase the pitch of selected audio samples and then insert a short yet high-pitched signal to the position with the highest sound energy. In this way, we can exploit sound masking mechanism [52] to generate stealthy yet effective poisoned samples. This method is dubbed attack via pitch boosting and sound masking (PBSM).

The pitch boosting makes our attack resistant to trigger filtering (as shown in our experiments). The filtering cannot decrease the pitch of poisoned audio since these triggers are natural, although it may remove the high-pitched short signal. Besides, our insertion strategy improves the stealthiness of triggers for both human inspection and machine detection. Specifically, the inserted high-pitched signal is less perceptible to humans due to sound masking while it can bypass classical detection methods based on finding common audio clips since the insert position is usually sample-specific. In other words, different poisoned samples have different insert positions.

In general, our PBSM has two main steps, including **(1)** pitch boosting and **(2)** signal injection, to generate poisoned samples. The details of this process is described in Algorithm 1 and the main pipeline of PBSM is shown in Figure 1.

1) Step 1: Pitch Boosting: A feasible method for pitch boosting is to increase the frequency of selected audio samples. Accordingly, we first perform a short-time Fourier transform (STFT) [53] on the original audio to convert it from the time domain to the frequency domain. After that, in the frequency domain, we multiply the original frequency values by an adversary-specified pitch-shifting coefficient p ($p > 1$), leading to a new audio waveform with a boosted pitch. Specifically, we can express the short-time Fourier transform as $\boldsymbol{x_f} = \mathcal{F}(\boldsymbol{x})$ (Line 1 in Algorithm 1), where $\boldsymbol{x_f}$ is the frequency-domain representation of $\boldsymbol{x}$. The process of increasing pitch can be expressed as $\boldsymbol{x_P} = p \cdot \sum_{i=0}^{L_p} \boldsymbol{x_f}^{(i)}$ (Line 3 in Algorithm 1). Specifically, in the aforementioned equation, L_p represents the number of points in the frequency domain, the transformation factor p is represented as $p = 2^{n\_p/12}$, and n_p denotes the number of semitones (*i.e.*, the step of pitch shifting).

2) Step 2: Signal Injection: This process consists of two main stages, including **(1)** location identification and **(2)** signal insertion. In the first stage, we identify the location of the high-amplitude segments in the audio signal. We select the high-amplitude clips since they have stronger energy and can provide better masking effects. Specifically, to find these positions, we iterate through each audio segment to identify the

Algorithm 1 The Algorithm of Pitch Boosting and Sound Masking (PBSM)

```
Require: Benign audio $\boldsymbol{x}$ and high-pitch signal $\boldsymbol{h}$.
1: $\boldsymbol{x_f} = \mathcal{F}(\boldsymbol{x})$ // Short-time Fourier transformation.
2: for $i$ in range($\boldsymbol{x_f}$) do
3:     $\boldsymbol{x_P} = p \cdot \sum_{i=0}^{L_p} \boldsymbol{x_f}^{(i)}$ // Pitch boosting in the frequency domain.
4: end for
5: $T = \underset{i}{\text{argmax}}(\sum_i^{i+L} |\boldsymbol{x_P}^{(i)}|) + L$ // Calculate the position of high-amplitude segments.
6: $\boldsymbol{x_r} = \boldsymbol{x_P}^{(T)} \oplus \boldsymbol{h}$ // Insert a high-pitch signal.
7: $\boldsymbol{x_t} = \mathcal{F}^{-1}(\boldsymbol{x_r})$ // Inverse Fourier transformation.
Ensure: Poisoned audio $\boldsymbol{x_t}$.
```

position of the segment with the highest energy in the entire audio sample. The position T of high-amplitude segments can be obtained by: $T = \underset{i}{\text{argmax}}(\sum_i^{i+L} |\boldsymbol{x}_{\boldsymbol{P}}^{(\boldsymbol{i})}|) + L$ (Line 5 in Algorithm 1), where L is the high-amplitude length. In the second stage, we insert an adversary-specified high-pitched signal $\boldsymbol{h}$ in the selected position T. Specifically, this process can be denoted by $\boldsymbol{x_r} = \boldsymbol{x_P}^{(T)} \oplus \boldsymbol{h}$ (Line 6 in Algorithm 1), where $\boldsymbol{x_r}$ is the inserted audio signal after signal injecting, $\boldsymbol{x_P}^{(T)}$ is the audio segment at position T, and $\oplus$ denotes the injection operation with the high-pitched signal $\boldsymbol{h}$. We conduct the inverse Fourier transformation $\mathcal{F}^{-1}$ [53] to obtain poisoned audio with pitch-type triggers by turning frequency-domain signals back to the time domain (Line 7 in Algorithm 1).

D. Attack via Voiceprint Selection and Voice Conversion

To design timbre-type triggers, we can exploit a 'timbre transformer' trained on the audio of an adversary-specified target people (*e.g.*, the adversary himself) for voice conversion [54]. Specifically, we can assign the poisoned instance generator G as the (pre-trained) timbre transformer.

Assume that there are multiple timbre candidates for selection. Arguably, the design of timbre-type single-backdoor attacks is straightforward, where the adversaries can arbitrarily choose any single timbre they desire. However, the design of multi-backdoor attacks is challenging since simply selecting multiple timbres at random to design triggers has limited attack effectiveness (as we will show in the experiments). It is mostly because there can be many similarities between timbres. On the one hand, this similarity makes it harder for DNNs to learn backdoors, since similar poisoned samples have different (target) labels. On the other hand, this similarity may lead to false backdoor activation by attacked models at the inference process. Motivated by these understandings, we propose a *voiceprint selection module* to alleviate these challenges.

In general, our voiceprint selection module consists of three main stages, including **(1)** feature extraction, **(2)** similarity calculation, and **(3)** timbre selection. The main pipeline of our voiceprint selection and voice conversion (VSVC) is shown in Figure 2. Its technical details are as follows.

1) Step 1: Feature Extraction: Following the most classical method, we exploit X-vectors [55] to extract voiceprint features of each timbre candidates, *i.e.*, $\boldsymbol{S_e^{(k)}} \leftarrow V(C_k)$, where C_k is the speech data for the k-th speaker, V denotes the process of extracting X-vectors converting each speech into a d-dimensional feature vector, and $\boldsymbol{S_e^{(k)}}$ represents the voiceprint embedding for the k-th speaker. For K candidates, we ultimately obtain a matrix $\boldsymbol{S_e} = [\boldsymbol{S_e^{(1)}}, \ldots, \boldsymbol{S_e^{(K)}}] \in \mathbb{R}^{d\times K}$ with d rows and K columns (Lines 1-3 in Algorithm 2).

2) Step 2: Similarity Calculation: In this step, we calculate the distance between the features of each timbre pair (i, j) as their similarity. Specifically, to represent the voiceprint distances between K candidates, we construct a similarity matrix Sim of size K^2, where each element $Sim[i][j]$ is computed as $d\left(\boldsymbol{S_e^{(i)}}, \boldsymbol{S_e^{(j)}}\right)$ (Lines 4-7 in Algorithm 2) with the distance metric d. In this paper, we assign d as ℓ_2-norm for simplicity.

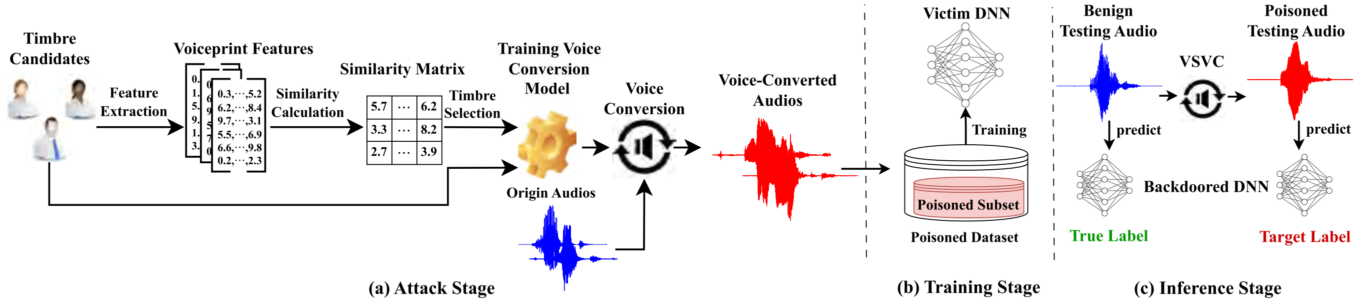
3) Step 3: Timbre Selection: In this step, we select M candidates with maximum distances, based on the similarity matrix calculated in the previous step. We design a greedy search method to select suitable candidates (Lines 9 in Algorithm 2). Specifically, we select the two timbres with the greatest distance in the similarity matrix to add to the selected set $\mathcal{C}_M$. After that, we select the timbre that has the greatest distance from all the timbres in the selected set from the remaining candidates and add it to the selected set. We repeat the above process until the selected set $\mathcal{C}_M$ contains M timbres.

4) Step 4: Generating the Poisoned Dataset via Voice Conversion: In this step, we first train a voice conversion model G (Line 10 in Algorithm 2), based on the selected set $\mathcal{C}_M$ obtained in the previous step. For each audio $\boldsymbol{x}$, $G(\boldsymbol{x}, i)$ can convert its timbre to that of i-th element in $\mathcal{C}_M$. After that, we select M adversary-specified target labels $\{y_T^{(i)}\}_{i=1}^M$. Each target label is associated with a timbre backdoor. The generated poisoned dataset $\hat{\mathcal{D}}$ contains $(M+1)$ disjoint subsets, including one benign subset $\mathcal{D}_b$ and M poisoned subsets (*i.e.*, $\{\mathcal{D}_p^{(i)}\}_{i=1}^M$). Specifically, $\mathcal{D}_p^{(i)} \triangleq \{(G(\boldsymbol{x}, i), y_T^{(i)})|(\boldsymbol{x}, y) \in \mathcal{D}_s^{(i)}\}$ where $\mathcal{D}_s^{(i)} \subset \mathcal{D}$, $\mathcal{D}_s^{(i)} \cap \mathcal{D}_s^{(j)} = \emptyset$ $(\forall i \neq j)$ (Lines 11-14 in Algorithm 2), and $\mathcal{D}_b = \mathcal{D} - \bigcup_{i=1}^M \mathcal{D}_s^{(i)}$ (Line 15 in Algorithm 2). In particular, $\gamma_i \triangleq \frac{|\mathcal{D}_s^{(i)}|}{|\mathcal{D}|}$ is dubbed as the poisoning rate of i-th timbre-type backdoor.

IV. Experiments

A. Main Settings

1) Dataset Description: We adopt the most classical datasets, *i.e.*, Google Speech Command Dataset (SCD) [56], LibriSpeech [57], and VoxCeleb1 [58], for our evaluations. Specifically, SCD consists of 30 common English speech commands. Each command is spoken by multiple individuals in various ways, resulting in a total of 64,728 samples. The dataset has a 16kHz sampling rate where each sample lasts approximately one second. Specifically, we selected 23,726 audio samples with 10 labels (dubbed 'SCD-10') and 64,721 audio samples with 30 labels (dubbed 'SCD-30') for a comprehensive comparison; For the LibriSpeech dataset, we extract speech segments from its development set (2,703 speech segments from 40 speakers in total); For the VoxCeleb1, we select



(a) Attack Stage (b) Training Stage (c) Inference Stage

Fig. 2. The main pipeline of attacking via our voiceprint selection and voice conversion (VSVC). The VSVC consists of three main stages, including attack, training, and inference. The attack stage is the core of VSVC, containing four steps (*i.e.*, feature extraction, similarity calculation, timbre selection, and voice conversion). In the first step, we adopt X-vectors to extract voiceprint features of each timbre candidate; In the second step, we measure the similarity of each timbre pair based on their distance; In the third step, we select the desired number of timbres based on the principle of smallest similarity; In the fourth step, we generate the poisoned training dataset of the (multi-backdoor) timbre-type attack via voice conversion.

TABLE II

Comparisons Between Different Attacks. In This Table, We Exploit the 'checkmark' to Indicate Whether a Method Had a Particular Promising Property or Considered a Special Setting

		PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
Attribute	Non-Noise Trigger				✓	✓	✓
	Imperceptible			✓		✓	✓
	Undetectable				✓	✓	✓
	Non-Filterable				✓	✓	✓
Setting	Clean-Label Attack				✓	✓	✓
	All-to-All Attack					✓	✓
	Physical Attack	✓	✓	✓		✓	✓
	Defense Resistance	✓	✓			✓	✓
	Human Validation				✓	✓	✓

speech segments from 50 speakers with the largest sample size on the original VoxCeleb1-H (18,354 speech segments from 50 speakers in total). Besides, to ensure consistent experimental settings, we trim all segments to 1 second.

2) Baseline Selection: We compared our PBSM and VSVC with four representative speech backdoor attacks, including **(1)** position-independent backdoor attack (PIBA) [15], **(2)** dual adaptive backdoor attack (DABA) [18], **(3)** backdoor attack with ultrasonic (dubbed 'Ultrasonic') [16], and **(4)** backdoor attack via style transformation (dubbed 'JingleBack') [43]. As shown in Table II, all baseline attacks only had limited promising properties or considered limited settings.

3) Model Structures: As the poison-only backdoor attacks, we assume that the adversaries have no information about the victim model. To evaluate the effectiveness across different DNNs, we evaluate all attacks under four classical and advanced DNN structures, including LSTM [59], ResNet-18 [27], KWT [29], and EAT [30]. Specifically, LSTM and ResNet-18 are classical models designed for sequential and non-sequential data, respectively; KWT and EAT are advanced speech recognition models, where KWT exploited transformer structure and EAT was designed in an end-to-end manner.

4) Attack Setup: For all attacks, we set the poisoning rate as 0.01. We randomly select the target label for all datasets ('left' on SCD-10 and SCD-30, 'id84' on Librispeech, and 'id10020' on Voxceleb1). For our PBSM method, we increase the pitch by 5 semitones. The length of high-amplitude segments is set to 100 milliseconds. For our VSVC method, we select the VCTK dataset [60] as the timbre candidates dataset and we employ StarGANv2-VC [61] as the voice conversion framework. In particular, we evaluate the single-backdoor VSVC in our main experiments for a fair comparison. The results of multi-backdoor VSVC are included in Section IV-C.6; For DABA [18] and PIBA [15], we follow the same settings described in their original papers; For the ultrasonic attack [16], we set the duration of the trigger to 100 milliseconds; For JingleBack [43], we exploit the third style used in its paper since it led to the best attack performance. Note that this method may reach better stealthiness if we use other styles introduced in their paper, whereas it will decrease its attack effectiveness as the sacrifice.

5) Training Setup: We extract the log-Mel spectrogram of each audio sample as an input feature, which can graphically characterize a person's speech feature in a combination of temporal and frequency dimensions. All models are trained for 100 epochs. We set the learning rate of EAT and LSTM as 0.0001 and 0.005, respectively. We set the learning rate of the remaining models as 0.01. As for the optimizer selection, the EAT and KWT models are trained using the Adam optimizer, while the default optimizer for the other models is SGD. We run each experiment three times and calculate their average to reduce the side effects of randomness.

6) Training Facilities: We conduct all our experiments on a server running Ubuntu 18.04, equipped with a single NVIDIA GeForce RTX 3090 GPU with 24GB of VRAM.

TABLE III

The Benign Accuracy (BA) and Attack Success Rate (ASR) of Methods on the SCD-10 Dataset. We Mark the Best Results Among All Inaudible Attacks (*i.e.*, Ultrasonic, JingleBack, PBSM, and VSVC) in Boldface

Model↓	Metric↓, Method→	No Attack	PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
LSTM	BA (%)	93.68	93.54	92.13	93.21	92.63	93.32	**93.43**
	ASR (%)	–	95.23	99.76	98.61	91.31	92.11	**99.61**
ResNet-18	BA (%)	95.11	94.32	94.10	**94.97**	94.55	94.85	94.93
	ASR (%)	–	96.43	99.87	**99.33**	95.52	95.78	97.57
KWT	BA (%)	91.35	90.21	90.10	91.11	91.19	**91.27**	90.96
	ASR (%)	–	96.24	99.54	97.13	91.52	94.39	**99.22**
EAT	BA (%)	93.33	93.21	92.61	93.12	93.10	93.23	**93.31**
	ASR (%)	–	97.32	99.21	**99.12**	87.39	90.13	92.32

TABLE IV

The Benign Accuracy (BA) and Attack Success Rate (ASR) of Methods on the SCD-30 Dataset. We Mark the Best Results Among All Inaudible Attacks (*i.e.*, Ultrasonic, JingleBack, PBSM, and VSVC) in Boldface

Model↓	Metric↓, Method→	No Attack	PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
LSTM	BA (%)	92.62	92.51	91.18	92.13	**92.57**	92.56	91.91
	ASR (%)	–	95.04	99.16	98.12	**98.45**	96.21	98.01
ResNet-18	BA (%)	95.20	93.21	92.13	94.32	94.76	94.71	**94.85**
	ASR (%)	–	98.34	99.98	**97.53**	93.39	96.63	93.01
KWT	BA (%)	91.13	90.62	89.19	90.33	90.20	**90.45**	90.21
	ASR (%)	–	94.21	99.45	**97.13**	93.54	94.02	97.03
EAT	BA (%)	94.51	94.33	93.13	94.23	94.35	94.01	**94.38**
	ASR (%)	–	92.12	99.43	**95.32**	81.06	92.51	93.12

7) Evaluation Metrics: Following the most classical settings in existing works [6], we adopt benign accuracy (BA) and attack success rate (ASR) to evaluate the effectiveness of all attacks. Specifically, the BA measures the proportion of benign testing samples that can be correctly classified, while the ASR denotes the proportion of poisoned testing samples that can be maliciously predicted as the target label. The higher the BA and the ASR, the more effective the attack; To evaluate the stealthiness, we invite 30 people to determine whether the poisoned audio samples (5 for each attack) of an attack sound natural. The proportion of poisoned samples that are regarded as natural audio by humans is dubbed natural rate (NC). The higher the NC, the more stealthy the attack.

B. Main Results

1) Attack Effectiveness: As shown in Table III-VI, the attack success rates (ASRs) of our PBSM and VSVC are sufficiently high in all cases. For example, the ASRs are larger than 90% on SCD datasets and are 95% on Librispeech and Voxceleb1 datasets. The attack performance of our VSVC is on par with or even better than all baseline attacks except for DABA in some cases. For example, in the SCD-10 dataset, the ASR of VSVC is 8% higher than that of JingleBack when attacking LSTM and KWT models. Besides, our attacks have minor adverse effects on benign accuracy. The decreases of benign accuracy compared to the model training with benign dataset are less than 1% in all cases for our attacks. In contrast, both DABA and JingleBack have a relatively high impact on benign accuracy. These results verify the effectiveness of our attacks.

2) Attack Stealthiness: We notice that the ASRs of baseline attacks (especially DABA and Ultrasonic) are higher than those of ours in some cases. However, it comes at the expense of stealthiness. As shown in Table VII, the natural rates of all baseline attacks other than Ultrasonic are significantly lower than our PBSM and VSVC. For example, the natural rates of PIBA, DABA, and JingleBack are all 0% while those of our PBSM and VSVC are near 100%. Ultrasonic has a similar natural rate to that of benign samples simply because humans cannot hear ultrasound. However, it does not mean that this attack is stealthy. The victim users can still easily identify this attack by checking the spectrogram of samples (as shown in the area of the black dashed box in Figure 3d and Figure 3l). Users can also filter out ultrasonic trigger signals to depress this attack. These results verify the stealthiness of our attacks.

In conclusion, our attacks can preserve high effectiveness while ensuring stealthiness. In contrast, existing baseline methods can be easily detected and defended.

C. Ablation Study

In this section, we discuss the effects of key parameters, including target label, poisoning rate, high-pitch signal, and timbre, of our PBSM and VSVC. We adopt SCD-10 as an

TABLE V

THE BENIGN ACCURACY (BA) AND ATTACK SUCCESS RATE (ASR) OF METHODS ON THE LIBRISPEECH DATASET. WE MARK THE BEST RESULTS AMONG ALL INAUDIBLE ATTACKS (*i.e.*, ULTRASONIC, JINGLEBACK, PBSM, AND VSVC) IN BOLDFACE

Model↓	Metric↓, Method→	No Attack	PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
LSTM	BA (%)	99.28	98.94	98.95	99.01	98.17	**99.10**	98.96
	ASR (%)	–	98.12	99.26	98.21	98.33	98.81	**99.63**
ResNet-18	BA (%)	99.12	98.83	98.25	**99.26**	98.18	99.06	98.34
	ASR (%)	–	97.20	98.65	98.57	98.52	95.53	**99.54**
KWT	BA (%)	97.39	97.11	97.21	97.17	97.01	**97.38**	97.32
	ASR (%)	–	97.76	98.12	96.92	98.17	96.18	**98.75**
EAT	BA (%)	98.28	97.21	97.31	**98.89**	97.16	98.17	98.18
	ASR (%)	–	97.77	99.01	94.01	94.23	95.89	**99.66**

TABLE VI

THE BENIGN ACCURACY (BA) AND ATTACK SUCCESS RATE (ASR) OF METHODS ON THE VOXCELEB1 DATASET. WE MARK THE BEST RESULTS AMONG ALL INAUDIBLE ATTACKS (*i.e.*, ULTRASONIC, JINGLEBACK, PBSM, AND VSVC) IN BOLDFACE

Model↓	Metric↓, Method→	No Attack	PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
LSTM	BA (%)	94.37	93.23	92.65	93.69	92.50	94.13	**94.15**
	ASR (%)	–	98.96	98.12	99.43	99.18	99.57	**99.81**
ResNet-18	BA (%)	96.34	93.95	93.65	93.78	92.51	95.45	**95.67**
	ASR (%)	–	95.61	99.78	98.12	98.92	99.35	**99.98**
KWT	BA (%)	95.30	94.96	94.26	94.11	94.61	**94.75**	94.63
	ASR (%)	–	99.02	99.77	98.58	97.37	98.05	**99.85**
EAT	BA (%)	95.52	94.65	93.15	94.05	94.45	95.38	**95.49**
	ASR (%)	–	98.13	99.34	97.11	93.63	96.84	**99.94**

TABLE VII

THE NATURAL RATES (%) CALCULATED BY HUMAN VALIDATION OF SAMPLES GENERATED BY DIFFERENT METHODS

Benign	PIBA	DABA	Ultrasonic	JingleBack	PBSM (Ours)	VSVC (Ours)
100	0	0	100	0	97.3	100

TABLE VIII

THE ATTACK SUCCESS RATE (%) *w.r.t.* DIFFERENT BOOSTED SEMITONES ON THE SCD-10 DATASET

Model→ Semitone↓	LSTM	ResNet-18	KWT	EAT
1	5.13	33.61	38.17	37.91
3	70.61	69.70	79.08	46.17
5	80.74	85.65	82.13	73.09
7	86.09	89.35	83.19	81.61

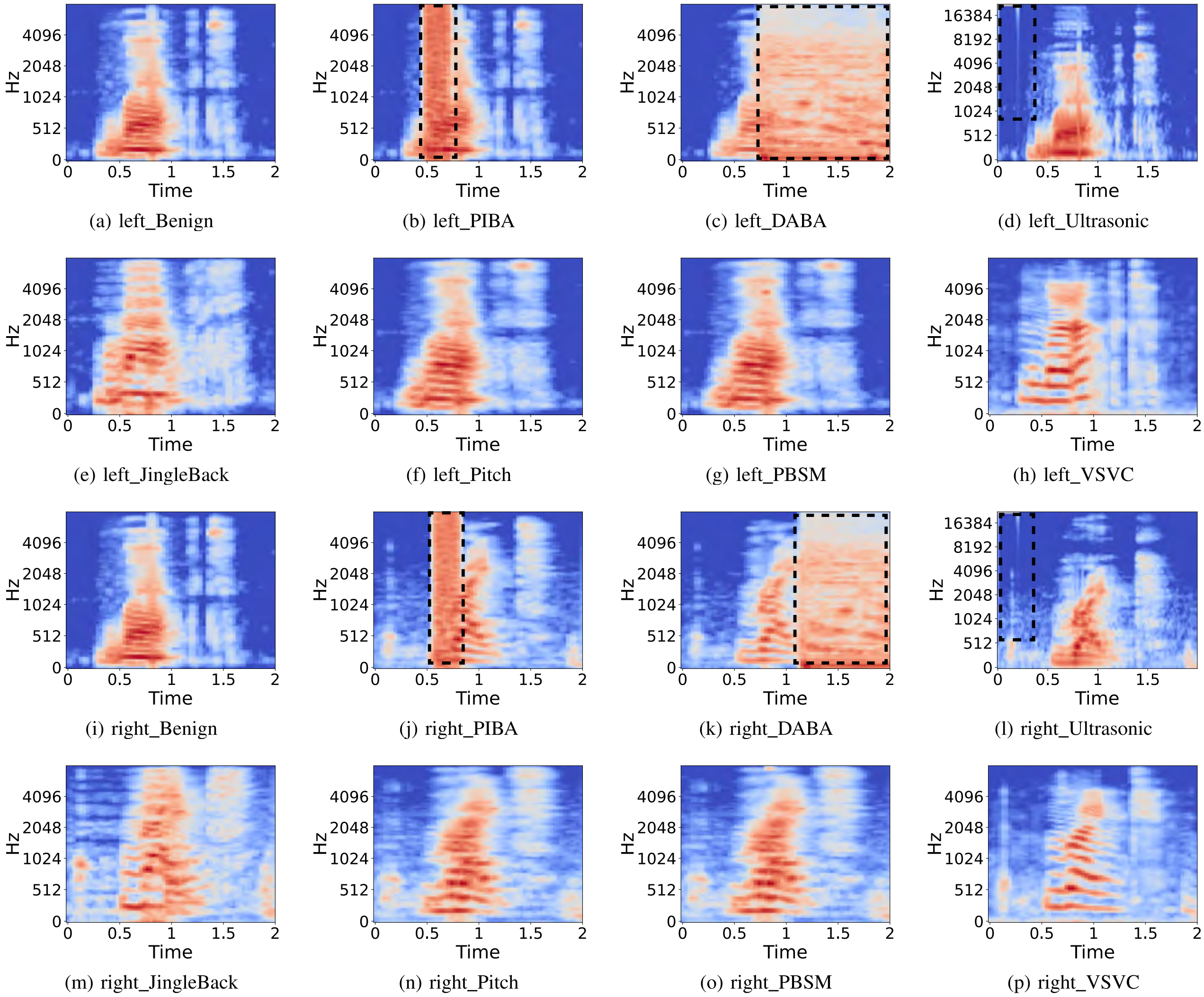
example for our discussions. Unless otherwise specified, all settings are consistent to those stated in Section IV-A.

1) Effects of the Poisoning Rate: To explore the influences of the poisoning rate on our attacks, we conduct experiments with poisoning rates ranging from 0.5% to 2.0% against all four model structures. As shown in Figure 4, the attack success rates (ASRs) of both PBSM and VSVC increase with the increase of the poisoning rate, although our attacks can reach promising attack performance by poisoning only 1% training samples. However, the benign accuracy (BA) will decrease with the increase of the poisoning rates to some extent, *i.e.*, there is a trade-off between ASR and BA. The adversaries should assign a suitable poisoning rate based on their needs.

2) Effects of the Target Label: To verify that our PBSM and VSVC are still effective under different target labels, we conduct experiments with ResNet-18. As shown in Figure 6, the attack success rates of both PBSM and VSVC are similar across all evaluated target labels. Specifically, the ASRs are larger than 93% in all cases, while the decrease of benign accuracy compared to 'no attack' is less than 1%. These results show that target labels have minor effects on our attacks. The adversaries can select any target class based on their needs.

3) Effects of the Pitch Boosting: In this part, we show that pitch boosting used in our PBSM itself can serve as the pitch-type trigger and explore its effects. Specifically, we increase the pitch range from one semitone to seven semitones and evaluate the attack success rate (ASR). The example of the spectrograms of samples with different boosted semitones is shown in Figure 5. As shown in Table VIII, the ASR increases with the increase of semitones, as we expected. Specifically, the ASRs are larger than 80% in three out of all four cases when we boost five semitones. However, we have to notice that excessive pitch boosting can lead to significant sound distortion and therefore decrease attack stealthiness.

4) Effects of the Short-duration High-pitch Signal: To verify that inserting a high-pitch signal is critical for our



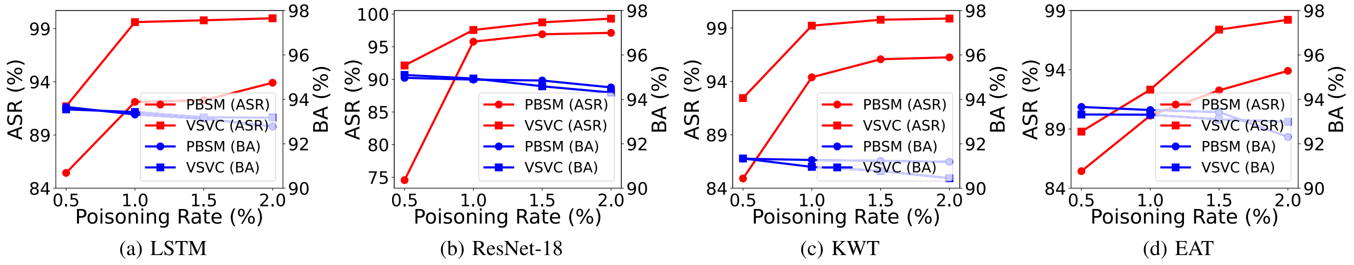
(a) left_Benign (b) left_PIBA (c) left_DABA (d) left_Ultrasonic

(e) left_JingleBack (f) left_Pitch (g) left_PBSM (h) left_VSVC

(i) right_Benign (j) right_PIBA (k) right_DABA (l) right_Ultrasonic

(m) right_JingleBack (n) right_Pitch (o) right_PBSM (p) right_VSVC

Fig. 3. The spectrograms of different samples. In this example, we present the visualization of two benign audio (with the labels 'left' and 'right') and their poisoned versions generated by different attacks. As shown in this figure, we can easily detect the abnormalities of poisoned samples generated by PIBA, DABA, and ultrasonic from spectrogram areas marked in the black box. However, we cannot detect anomalies in JingleBack and our attacks directly through the spectrogram.



(a) LSTM (b) ResNet-18 (c) KWT (d) EAT

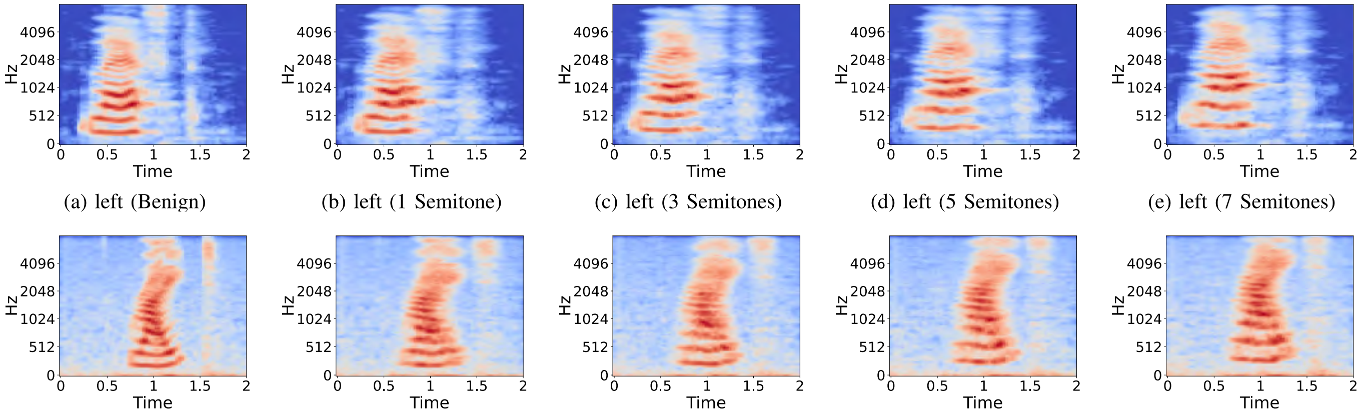
Fig. 4. The performance of our PBSM and VSVC on the SCD-10 dataset under different poisoning rates.

PBSM, we compare its attack success rate to that of its pitch-only variant where we only increase the pitch without adding the high-pitch signal. As shown in Table IX, although the pitch-only method can have some attack effects, introducing a high-pitch signal can significantly improve the attack effectiveness. Specifically, the attack success rate of PBSM is 10% higher than that of its pitch-only variant in all cases. These results verify the effectiveness of our PBSM.

TABLE IX
THE ATTACK SUCCESS RATE (%) OF PITCH-ONLY ATTACK AND PBSM ATTACK ON THE SCD-10 DATASET

Method↓, Model→	LSTM	ResNet-18	KWT	EAT
Pitch-Only	80.74	85.65	82.13	73.09
PBSM	**92.11**	**95.78**	**94.39**	**90.13**

5) Effects of the Timbre: To verify that our VSVC is still effective with different timbres, we conducted experiments



(a) left (Benign) (b) left (1 Semitone) (c) left (3 Semitones) (d) left (5 Semitones) (e) left (7 Semitones)

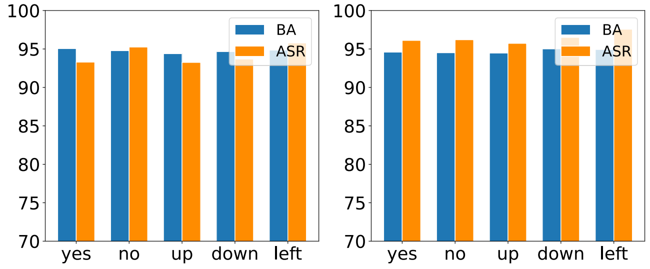
(f) right (Benign) (g) right (1 Semitone) (h) right (3 Semitones) (i) right (5 Semitones) (j) right (7 Semitones)

Fig. 5. The spectrograms of samples whose pitch is boosted with different semitone. In this example, we present the visualization of two benign audio (with the labels 'left' and 'right') and their boosted versions. As shown in this figure, the fundamental frequency variations in the spectrogram become more pronounced as the semitone increases.

Algorithm 2 The Algorithm of Voiceprint Selection and Voice Conversion (VSVC)

```
Require: Benign dataset $\mathcal{D}$, the number of backdoors $M$, poisoning rates for $M$ poisoned subsets $\{\gamma_i\}_{i=1}^{M}$, the number of timbre candidates $K$, timbre candidates $\mathcal{C}_K$, and target labels $\left\{y_T^{(i)}\right\}_{i=1}^{M}$.
for $k$ in range($K$) do
    $S\_e \leftarrow V(\mathcal{C}_k)$; // Extract the voiceprint features.
end for
for $i$ in range($K$) do
    for $j$ in range($K$) do
        $Sim[i][j] = d(\boldsymbol{S\_e^{(i)}}, \boldsymbol{S\_e^{(j)}})$ // Generate a similarity matrix based on the distance between each pair.
    end for
end for
$\mathcal{C}_M = GreedySearch(Sim, M)$ // Select $M$ suitable timbre candidates.
$G$ = Train($\mathcal{C}_M$) // Training the voice conversion model.
for $i$ in range($M$) do
    $\mathcal{D}_s^{(i)} \triangleq Extract(\mathcal{D} - \bigcup_{j=1}^{i-1} \mathcal{D}_s^{(j)}, \gamma_i \cdot |\mathcal{D}|)$ // Extract $M$ disjoint subsets for poisoning.
    $\mathcal{D}_p^{(i)} \triangleq \{(G(\boldsymbol{x}, i), y_T^{(i)}) | (\boldsymbol{x}, y) \in \mathcal{D}_s^{(i)}\}$ //Generate voice-converted poison samples.
end for
$\mathcal{D}_b = \mathcal{D} - \bigcup_{i=1}^{M} \mathcal{D}_s^{(i)}$
$\hat{\mathcal{D}} = \mathcal{D}_b \cup \left(\bigcup_{i=1}^{M} \mathcal{D}_p^{(i)}\right)$
Ensure: The poisoned dataset $\hat{\mathcal{D}}$ generated by our VSVC.
```

on the SCD-10 dataset. The example of the spectrograms of samples with different timbres is shown in Figure 7. As shown in Table X, the ASRs of VSVC are similar across all evaluated timbres. Specifically, the ASRs are larger than 91% in all cases, while the decrease of benign accuracy compared to 'no attack' is only about 1%. These results indicate that timbre selection has mild effects on our attack. The adversaries can select any timbre based on their needs.



(a) PBSM (b) VSVC

Fig. 6. The effects of the target label on our PBSM and VSVC attacks on the SCD-10 dataset.

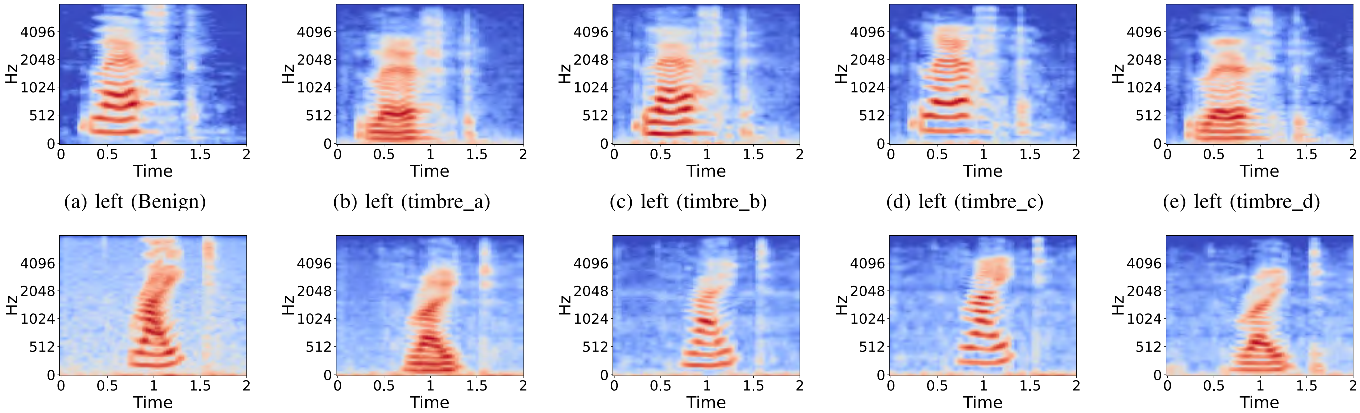
TABLE X
THE ATTACK SUCCESS RATE (%) OF OUR VSVC ATTACK WITH DIFFERENT TIMBRES ON THE SCD-10 DATASET

Timbre↓	Metric↓ Model→	LSTM	ResNet-18	KWT	EAT
(a)	BA (%)	93.56	94.88	91.04	93.13
	ASR (%)	98.52	97.51	98.71	91.33
(b)	BA (%)	93.32	94.76	91.36	93.21
	ASR (%)	99.08	98.53	98.81	93.11
(c)	BA (%)	92.88	94.23	90.98	92.89
	ASR (%)	97.60	96.65	97.87	92.30
(d)	BA (%)	93.15	94.22	90.77	92.78
	ASR (%)	98.15	96.73	98.69	92.14
(e)	BA (%)	92.61	94.35	91.33	92.39
	ASR (%)	99.17	98.92	99.08	94.47

TABLE XI
THE PERFORMANCE OF VSVC WITHOUT AND WITH VOICEPRINT SELECTION UNDER THE MULTIPLE-BACKDOOR SETTING

Method↓	Metric↓, Model→	LSTM	ResNet-18	KWT	EAT
VSVC (w/o)	BA (%)	91.23	94.58	88.54	91.43
	ASR (%)	89.10	91.24	92.34	87.65
VSVC (w/)	BA (%)	92.05	95.05	90.13	93.14
	ASR (%)	92.77	97.78	97.03	93.78

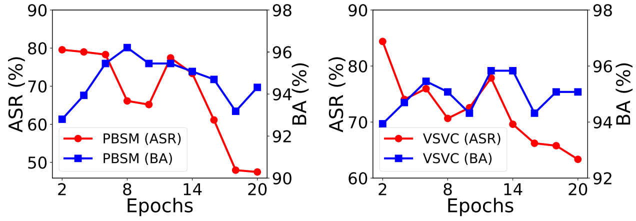
6) Effects of the Voiceprint Selection: To verify that voiceprint selection is critical for our VSVC under the



(a) left (Benign) (b) left (timbre_a) (c) left (timbre_b) (d) left (timbre_c) (e) left (timbre_d)

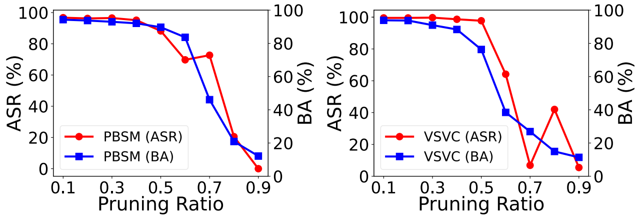
(f) right (Benign) (g) right (timbre_a) (h) right (timbre_b) (i) right (timbre_c) (j) right (timbre_d)

Fig. 7. The spectrograms of samples with different timbre. In this example, we present the visualization of two benign audio (with the labels ‘left’ and ‘right’) and their variants with different timbres. As shown in this figure, the spectrograms of our VSVC with different timbres remain normal to human inspection, although there are notable differences across them.



(a) PBSM (b) VSVC

Fig. 8. The resistance of our PBSM and VSVC to fine-tuning.



(a) PBSM (b) VSVC

Fig. 9. The resistance of our attacks to model pruning.

multi-backdoor setting, we compare its attack success rate to that of its variant where we randomly select timbre candidates for voice conversion. In these experiments, we select three timbre candidates for discussions. As shown in Table XI, although the random selection variant can also have some attack effects, the introduction of voiceprint selection can significantly improve attack effectiveness. Specifically, the attack success rates of VSVC are 5% higher than those of its random selection variant in almost all cases. These results verify the effectiveness of the voiceprint selection introduced in our VSVC.

D. The Resistance to Potential Defenses

Currently, there are many backdoor defenses designed to reduce backdoor threats in image classification tasks [62], [63], [64]. However, most of them cannot be directly used in audio tasks since they are specified for the image domain. Accordingly, in this paper, we evaluate our attacks under three classical and representative cross-domain defenses, including model pruning [65], fine-tuning [66], and trigger filtering. We conduct experiments with the ResNet-18 model on the SCD-10 dataset for simplicity. Unless otherwise specified, all other settings are the same as those illustrated in Section IV-A.

1) The Resistance to Fine-tuning: As a representative backdoor-removal method, fine-tuning [66] intends to remove model backdoors by fine-tuning it with a few local benign samples. This method is motivated by the catastrophic forgetting property [67] of DNNs. In our experiments, we exploit 10% of benign training samples as our benign data and set the learning rate as 0.005. As shown in Figure 8, the attack success rate decreases with the increase of the tuning epoch. However, even at the end of this process, the ASRs are still larger than 45% for both our PBSM and VSVC. These results verify that our attacks are resistant to fine-tuning to a large extent.

2) The Resistance to Model Pruning: As another representative backdoor-removal defense, model pruning [65] aims to remove model backdoors by pruning neurons that are dormant during the inference process of benign samples. This method is motivated by the assumption that backdoor and benign neurons are mostly separated in attacked DNNs. As shown in Figure 9, the attack success rates are significantly decreased when pruning large amounts of neurons. However, it comes at the cost of a sharp decrease in benign accuracy. Specifically, the ASR decreases by almost the same amount as the BA for both PBSM and VSVC. This is mostly because the assumption of model pruning does not hold in our attacks due to their global and complex trigger designs. These results verify the resistance of our attacks to model pruning.

3) The Resistance to Trigger-removal Defense: To deactivate the potential backdoor in attacked DNNs, the defenders may remove its high-pitched signals, low-pitched signals, and noises, to remove potential trigger patterns of the suspicious testing audio. Obviously, this method has minor effects on our VSVC since we change the global features of its poisoned samples. However, it may defeat our PBSM since we inject a high-pitched signal after boosting the pitch. Accordingly,

TABLE XII
THE ATTACK SUCCESS RATE (%) OF PBSM-INFECTED DNNS ON PITCH-BOOSTED SAMPLES WITH (W/) AND WITHOUT (W/O) INJECTING THE HIGH-PITCH SIGNAL ON SCD-10 AND SCD-30

Method→ Dataset↓	PBSM (w/o)	PBSM (w/)
SCD-10	65.04%	95.78%
SCD-30	70.62%	96.63%

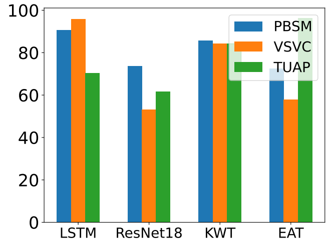


Fig. 10. Clean-Label Attacks.

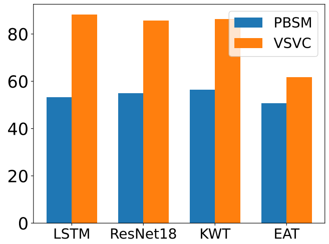


Fig. 11. Over-the-Air Attacks.

TABLE XIII
THE BENIGN ACCURACY (%) OF OVER-THE-AIR ATTACKS ON THE SCD-10 DATASET

Method↓, Model→	LSTM	ResNet-18	KWT	EAT
No Attack	93.51	94.87	90.31	93.21
PBSM (Ours)	92.78	94.23	90.28	92.12
VSVC (Ours)	92.81	93.56	90.03	92.22

we examine whether our PBSM attack is still effective when using pitch-boosted samples without injecting the high-pitch signal to query the PBSM-infected DNNs. As shown in Table XII, our attack can still reach satisfied attack success rates ($>65\%$) even without the high-pitch signals. It is mostly because our boosted pitch can also serve as a trigger pattern (as we mentioned in Section III-C) which cannot be removed by trigger filtering. It verifies the resistance of our attacks again.

E. Effectiveness of Our Attacks Under Various Settings

In this section, we discuss the attack effectiveness of our methods under more difficult settings.

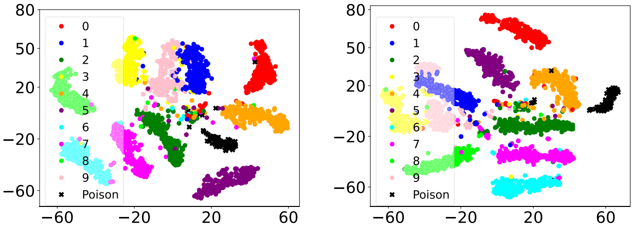
1) Attacks under the Clean-Label Setting: Although our attacks are imperceptible, the label of the poisoned samples usually differs from that of their clean versions. Accordingly, users may identify the attack by inspecting the audio-label relation when they can catch some poisoned samples. To further demonstrate the effectiveness of our methods, we explore whether they are still effective under the clean-label setting. In these experiments, we only select samples from the target class for poisoning instead of sampling data from all classes and changing their label to the target one. Besides, we also generalize TUAP [68], a representative clean-label backdoor attack against videos, to attack audio for reference. Specifically, we set the maximum perturbation size as 0.008 and perturb the whole audio. In particular, TUAP requires a pre-trained model to generate its trigger patterns but our attacks do not. Specifically, we adopt a pre-trained benign model having the same structure as the one used by adversaries for TUAP. As shown in Figure 10, although the performances are relatively weaker than those of attacks under the poisoned-label setting, our attacks are still effective when poisoning 9% samples. Specifically, the average ASRs across all model structures of PBSM and VSVC are 81% and 73%, respectively. In particular, the performance of our attacks is on par with those of TUAP, although they require fewer attack capacities (*i.e.*, having a pre-trained model). These results verify the effectiveness of our methods under the clean-label setting.

TABLE XIV
THE PERFORMANCE OF PBSM AND VSVC UNDER THE ALL-TO-ALL SETTING ON SCD-10. IN THIS TABLE, WE PROVIDE THE RESULT PER CLASS FOLLOWING THE SETTINGS IN BADNETS [7]

Class	Accuracy (%) of Benign Model	PBSM		VSVC	
		BA (%)	ASR (%)	BA (%)	ASR (%)
yes	95.70	95.71	92.19	93.75	93.36
left	97.38	96.26	91.39	96.63	95.13
off	96.57	95.04	90.08	96.18	89.70
on	97.97	97.15	96.34	96.75	94.72
go	94.82	94.42	84.06	90.44	81.28
down	92.10	93.28	89.33	89.72	90.12
stop	96.79	97.59	93.17	95.58	94.38
no	90.87	90.48	84.52	90.87	81.35
right	96.53	95.75	92.66	94.21	93.45
up	97.11	98.16	93.02	96.32	92.65

2) Attacks under the Over-the-Air (Physical) Setting: To evaluate the effectiveness of our attack methods in real-world scenarios, we design a physical experiment to assess the performance of our attacks under the over-the-air setting. Specifically, we conduct these experiments in a room, where we use computer speakers to play our backdoor audio and a smartphone is used as the recording device to capture the audio. The obtained audio is input into the attacked DNNs for prediction. We measure the playback volume of the audio and it is similar to that of a normal conversation. We place the smartphone at a distance of 0.5 meters from the speaker. As shown in Figure 11, although the performances are relatively weaker than those of attacks under the digital setting, our attacks are still effective in the real world. Specifically, the average ASRs across all model structures of PBSM and VSVC are 53% and 80%, respectively. The lower ASR of the PBSM is mostly due to the limitations of our evaluated device, which may not effectively capture high-pitched signals. Besides, as shown in Table XIII, the benign accuracy of our attacks is on par with that of the benign model. It verifies the stealthiness of our attacks again.

3) Attacks under the All-to-All Setting: To further illustrate the effectiveness of our PBSM and VSVC, we extend the



(a) PBSM (b) VSVC

Fig. 12. The t-SNE visualization of features of benign and poisoned samples from the hidden feature space generated by PBSM-infected and VSVC-infected models.

all-to-one attack setting to a more challenging all-to-all one, where the target label y_t of a poisoned sample (with ground-truth class y) is set to $y' = (y + 1) \mod K$. In particular, we increase the poisoning rate to 15% due to the difficulty of this task. We conduct experiments on the SCD-10 dataset with ResNet-18. As shown in Table XIV, both PBSM and VSVC can reach promising performance against samples from all classes, although the performance may have some mild fluctuations across them. These results confirm the feasibility of our attacks under the all-to-all setting.

F. Analyzing Attacks in the Hidden Feature Space

In this section, we analyze why our PBSM and VSVC attacks are effective from the behaviors of samples in the hidden feature space of attacked DNNs.

1) Settings: In this section, we visualize the features of poisoned samples generated by the backbone (*i.e.*, the input of fully-connected layers) of attacked DNNs via t-SNE [69]. For simplicity, we adopt 2,500 samples and exploit ResNet-18 trained on the SCD-10 dataset for our analysis.

2) Results: As shown in Figure 12, poisoned samples (marked in black) cluster together regardless of their ground-truth labels. In contrast, the benign samples form separate clusters according to their ground-truth class. These phenomena are consistent with predicted behaviors of the attacked model where it 'assigns' the same label to all samples in the same cluster. These results also verify the effectiveness of our attacks, showing that they can force attacked DNNs to learn features of triggers and ignore the benign features. It enables attacked DNNs to minimize the distance between poisoned samples in the feature space and associate the learned trigger-related features with the target label.

V. Potential Negative Impacts and Limitations

This paper mainly intends to design simple yet effective tools to evaluate the backdoor robustness of current DNN-based speech recognition models. However, we recognize that our PBSM and VSVC methods resist existing backdoor defenses and could potentially be exploited by adversaries for malicious purposes. The adversaries may also design similar attacks targeting other tasks, drawing inspiration from our research. While effective countermeasures are yet to be established, users can at least mitigate or even prohibit these threats by exclusively using fully trusted training resources. Our next step is to develop advanced defensive strategies against our PBSM and VSVC attacks.

Although our attacks strike a good balance between effectiveness and stealthiness, there are still some potential limitations. First, our trigger patterns are manually specified by the adversaries rather than generated through optimization. We will discuss how to further improve our attacks by optimizing trigger patterns. Secondly, we are currently randomly selecting samples for poisoning without considering how to select the best ones. We will explore more effective poisoning strategies to further improve the effectiveness of our attacks while maintaining a low poisoning rate.

VI. Conclusion

In this paper, we revealed that almost all existing poison-only backdoor attacks against speech recognition are not stealthy due to their simple trigger designs. To overcome this deficiency, we proposed two simple yet effective attacks, including pitch boosting and sound masking (PBSM) and voiceprint selection and voice conversion (VSVC), inspired by the elements of sound. Our attacks generated more 'natural' poisoned samples and therefore are more stealthy. We also generalized and evaluated our attacks under more difficult settings, such as all-to-all, clean-label, and physical ones. However, we notice that the attack performance may have some degrades in some cases under these settings. We will explore how to alleviate this problem and design their defense countermeasures in our future works. We hope that our research can provide a deeper understanding of stealthy backdoor attacks in speech recognition, to facilitate the design of more sure and robust speech recognition models.

Acknowledgment

This work was mostly done when Hanbo Cai was a (remote) Research Intern with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, under the supervision of Dr. Yiming Li. The authors sincerely thank Zhe Ye from Ningbo University for his helpful comments on this article.

References

[1] S. Wang, Z. Zhang, G. Zhu, X. Zhang, Y. Zhou, and J. Huang, "Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 351–364, 2023.

[2] M. Marras, P. Korus, A. Jain, and N. Memon, "Dictionary attacks on speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 773–788, 2023.

[3] S. Hu et al., "Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[4] T. Zhou, Z. Cai, F. Liu, and J. Su, "In pursuit of beauty: Aesthetic-aware and context-adaptive photo selection in crowdsensing," *IEEE Trans. Knowl. Data Eng*, vol. 35, no. 9, pp. 9364–9377, Sep. 2023.

[5] M. Goldblum et al., "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1563–1580, Feb. 2023.

[6] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 35, no. 1, pp. 5–22, Jan. 2024.

[7] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

[8] K. Chen et al., "BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models," in *Proc. ICLR*, 2022, pp. 1–8.

[9] Y. Liu, G. Shen, G. Tao, S. An, S. Ma, and X. Zhang, "Piccolo: Exposing complex backdoors in NLP transformer models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 2025–2042.

[10] G. Cui, L. Yuan, B. He, Y. Chen, Z. Liu, and M. Sun, "A unified evaluation of textual backdoor learning: Frameworks and benchmarks," in *Proc. NIPS*, 2022, pp. 1–15.

[11] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109512.

[12] X. Qi, T. Xie, Y. Li, S. Mahloujifar, and P. Mittal, "Revisiting the assumption of latent separability for backdoor defenses," in *Proc. ICLR*, 2023, pp. 1–20.

[13] Y. Liu et al., "Trojaning attack on neural networks," in *Proc. NDSS*, 2018, pp. 1–16.

[14] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2560–2564.

[15] C. Shi et al., "Audio-domain position-independent backdoor attack via unnoticeable triggers," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2022, pp. 583–595.

[16] S. Koffas, J. Xu, M. Conti, and S. Picek, "Can you hear it: Backdoor attacks via ultrasonic triggers," in *Proc. ACM Workshop Wireless Secur. Mach. Learn.*, May 2022, pp. 57–62.

[17] Y. Kong and J. Zhang, "Adversarial audio: A new information hiding method and backdoor for DNN-based speech recognition models," 2019, *arXiv:1904.03829*.

[18] Q. Liu, T. Zhou, Z. Cai, and Y. Tang, "Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2390–2398.

[19] J. Xin, X. Lyu, and J. Ma, "Natural backdoor attacks on speech recognition models," in *Proc. MLCS*, 2023, pp. 597–610.

[20] Y. Luo, J. Tai, X. Jia, and S. Zhang, "Practical backdoor attack against speaker recognition system," in *Proc. ISPEC*, 2022, pp. 468–484.

[21] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. Amer. Inst. Electr. Eng.*, vol. 47, no. 2, pp. 617–644, Apr. 1928.

[22] D. R. Reddy, "Speech recognition by machine: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 501–531, Apr. 1976.

[23] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Found. Trends Signal Process.*, vol. 1, no. 3, pp. 195–304, 2008.

[24] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[25] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," Defense Adv. Res. Projects Agency (DARPA), Inf. Sci. Technol. Office (ISTO), Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NISTIR 4930, 1993.

[26] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," 2018, *arXiv:1808.08929*.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[28] R. Vygon and N. Mikhaylovskiy, "Learning efficient representations for keyword spotting with triplet loss," in *Proc. SPECOM*, 2021, pp. 773–785.

[29] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," in *Proc. Interspeech*, Aug. 2021, pp. 1–5.

[30] A. Gazneli, G. Zimerman, T. Ridnik, G. Sharir, and A. Noy, "End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network," 2022, *arXiv:2204.11479*.

[31] Y. Li, Y. Bai, Y. Jiang, Y. Yang, S.-T. Xia, and B. Li, "Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection," in *Proc. NIPS*, 2022, pp. 13238–13250.

[32] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," 2019, *arXiv:1912.02771*.

[33] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein, "Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch," in *Proc. NIPS*, 2022, pp. 19165–19178.

[34] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2023, pp. 771–785.

[35] M. Xue, C. He, J. Wang, and W. Liu, "One-to-N & N-to-one: Two advanced backdoor attacks against deep learning models," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 3, pp. 1562–1578, May 2022.

[36] L. Hou, Z. Hua, Y. Li, and L. Yu Zhang, "M-to-N backdoor paradigm: A stealthy and fuzzy attack to deep learning models," 2022, *arXiv:2211.01875*.

[37] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *Proc. IEEE 7th Eur. Symp. Secur. Privacy*, Jun. 2022, pp. 703–718.

[38] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.

[39] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6202–6211.

[40] Y. Li, T. Zhai, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor attack in the physical world," in *Proc. ICLR Workshop*, 2021, pp. 1–5.

[41] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. ICML*, 2018, pp. 284–293.

[42] T. Xu, Y. Li, Y. Jiang, and S.-T. Xia, "BATT: Backdoor attack with transformation-based triggers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[43] S. Koffas, L. Pajola, S. Picek, and M. Conti, "Going in style: Audio backdoors through stylistic transformations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[44] Y. Li, M. Zhu, X. Yang, Y. Jiang, T. Wei, and S.-T. Xia, "Black-box dataset ownership verification via backdoor watermarking," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2318–2332, 2023.

[45] J. Guo et al., "Domain watermark: Effective and harmless dataset copyright protection is closed at hand," in *Proc. NIPS*, 2023, pp. 1–30.

[46] M. Ya, Y. Li, T. Dai, B. Wang, Y. Jiang, and S.-T. Xia, "Towards faithful XAI evaluation via generalization-limited backdoor watermark," in *Proc. ICLR*, 2024, pp. 1–28.

[47] P. Ladefoged, *Elements of Acoustic Phonetics*. Chicago, IL, USA: The Univ. of Chicago Press, 1996.

[48] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16443–16452.

[49] A. J. Oxenham, "Pitch perception," *J. Neurosci.*, vol. 32, no. 39, pp. 13335–13338, 2012.

[50] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, May 1977.

[51] C. Micheyl, K. Delhommeau, X. Perrot, and A. J. Oxenham, "Influence of musical and psychoacoustical training on pitch discrimination," *Hearing Res.*, vol. 219, nos. 1–2, pp. 36–47, Sep. 2006.

[52] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer, 2013.

[53] S. J. Orfanidis, *Introduction to Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1995.

[54] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[55] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

[56] P. Warden. (2017). *Speech Commands: A Public Dataset for Single-Word Speech Recognition*. [Online]. Available: http://download.tensorflow.org/data/speech_commands_v0

[57] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[58] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027.

[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[60] C. Veaux, J. Yamagishi, and K. MacDonald, "The vctk corpus: A multilingual corpus of read speech in a variety of accents," in *Proc. LREC*, 2016.

[61] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Proc. Interspeech*, Aug. 2021, pp> 1–5.

[62] J. Guo, Y. Li, X. Chen, H. Guo, L. Sun, and C. Liu, "Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency," in *Proc. ICLR*, 2023, pp. 1–24.

[63] Z. Xiang, Z. Xiong, and B. Li, "UMD: Unsupervised model detection for X2X backdoor attacks," in *Proc. ICML*, 2023, pp. 38013–38038.

[64] N. M. Jebreel, J. Domingo-Ferrer, and Y. Li, "Defending against backdoor attacks by layer-wise feature analysis," in *Proc. PAKDD*, 2023, pp. 428–440.

[65] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. RAID*, 2018, pp. 273–294.

[66] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Nov. 2017, pp. 45–48.

[67] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[68] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14431–14440.

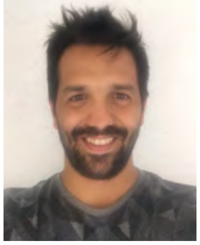
[69] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Hanbo Cai received the M.Eng. degree in software engineering from Guangxi Normal University in 2021. He is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Software Engineering, Hohai University. His research interests include the domain of trustworthy ML and responsible AI, especially backdoor learning and adversarial attacks.

Pengcheng Zhang (Member, IEEE) received the Ph.D. degree in computer science from Southeast University in 2010. He is currently a Full Professor with the College of Computer Science and Software Engineering, Hohai University, Nanjing, China. He has published research articles in premiere or famous computer science journals, such as IEEE Transactions on Software Engineering, IEEE Transactions on Services Computing, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Big Data, IEEE Transactions on Emerging Topics in Computing, IEEE Transactions on Cloud Computing, and IEEE Transactions on Reliability. His research interests include software engineering, service computing, and data science. He served as a technical program committee member for various international conferences. He was the Co-Chair of the IEEE AI Testing 2019 Conference.

Hai Dong (Senior Member, IEEE) received the Ph.D. degree from Curtin University, Perth, Australia. He is currently a Senior Lecturer with the School of Computing Technologies, RMIT University, Melbourne, Australia. Previously, he was a Vice-Chancellor's Research Fellow with RMIT University and a Curtin Research Fellow with Curtin University. His publications appear in *ACM Computing Surveys*, IEEE Transactions on Industrial Electronics, IEEE Transactions on Industrial Informatics, IEEE Transactions on Services Computing, and IEEE Transactions on Software Engineering. His research interests include services computing, edge computing, blockchain, cyber security, machine learning, and data science.

Yan Xiao received the Ph.D. degree from the City University of Hong Kong. She was a Research Fellow with the National University of Singapore. She is currently an Associate Professor with the School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen Campus. Her research interests include the trustworthiness of deep learning systems and AI applications in software engineering. More information is available on her homepage: https://yanxiao6.github.io/.



Stefanos Koffas received the M.Sc. degree in computer engineering from Delft University of Technology and the M.Eng. degree in electrical and computer engineering from the National Technical University of Athens, Greece. He is currently pursuing the Ph.D. degree with the Cybersecurity Group, Delft University of Technology. His research interests include the security of AI, especially on backdoor attacks in neural networks.

Yiming Li (Member, IEEE) received the B.S. degree (Hons.) in mathematics from Ningbo University in 2018 and the Ph.D. degree (Hons.) in computer science and technology from Tsinghua University in 2023. He is currently a Research Fellow with the College of Computing and Data Science, Nanyang Technological University. Before that, he was a Research Professor with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, and HIC-ZJU. His research has been published in multiple top-tier conferences and journals, such as ICLR, NeurIPS, and IEEE Transactions on Information Forensics and Security. His research interests include the domain of trustworthy ML and responsible AI, especially backdoor learning and AI copyright protection. His research has been featured by major media outlets, such as IEEE Spectrum. He was a recipient of the Best Paper Award at PAKDD in 2023 and the Rising Star Award at WAIC in 2023. He served as the Area Chair for ACM MM, the Senior Program Committee Member for AAAI, and a reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Information Forensics and Security, and IEEE Transactions on Dependable and Secure Computing.