BadMSSL: Audio Backdoor Attacks on Mask-Based Self-Supervised Learning

Abstract

Modern automatic audio systems, such as environmental sound recognition and speech command recognition, have advanced rapidly in recent years. Current supervised audio models face the challenge of limited labeled data. Thus, recent research has proposed self-supervised learning (SSL), which involves pretraining on large-scale unlabeled data and fine-tuning on smaller labeled data. While SSL is inherently resistant to direct data tampering, recent works have revealed its vulnerabilities to backdoor attacks. However, most previous backdoor attacks on SSL focus on contrastive-based SSL, with little attention paid to mask-based models in the audio field. To fill this gap, we propose BadMSSL, the first backdoor attack designed specifically for mask-based self-supervised learning models in the audio domain. Our method introduces Trojan samples during the pretraining phase, enabling the model to learn both target class features and backdoor-specific patterns. By constructing backdoor tasks and embedding malicious features, BadMSSL injects a backdoor into the SSL pretraining model such that the downstream model built on it simultaneously inherits the backdoor behaviour. Our experiments demonstrate that existing mask-based SSL audio models are vulnerable to BadMSSL while achieving a main task accuracy of up to 98.31%. Additionally, we find that directly adapting supervised backdoor attack methods to SSL frameworks results in suboptimal performance. Extensive ablation studies further confirm the robustness of BadMSSL across diverse scenarios, highlighting its effectiveness and resilience in audio applications.

1 Introduction

With the rapid advancement of deep neural networks (DNNs) [10, 73, 76], the performance of speech recognition systems [5, 45, 51] has been significantly enhanced, bringing them closer to human-level capabilities. This progress has transformed human-machine interaction [27] and enabled more intuitive voice-based interfaces, which are now widely used in tasks like speech command recognition (SCR) [2, 55], speaker recognition (SR) [29, 65, 77], language identification [16, 60], and speech emotion recognition [8]. In generation models, text-to-speech (TTS) [41, 62] and voice conversion [12, 59] applications are gaining momentum, making digital devices more accessible, especially for underserved groups like the elderly and visually impaired.

Self-supervised learning (SSL) [26, 32, 57] has emerged as a promising solution to the challenges of supervised models, which heavily rely on large labeled datasets. By generating pseudo-labels through pretext tasks, SSL enables models to learn from unlabeled data, reducing the need for labeled datasets. SSL methods are generally divided into two categories: Contrastive-based [3, 7, 28] and generative-based SSL [17]. Contrastive-based SSL focuses on learning representations by comparing positive and negative pairs, while generative-based SSL involves predicting or recovering masked parts of the input, such as in BERT [17] for text or SSAST [24] for audio. These SSL approaches significantly lower the reliance on expensive labeled data, making them promising for computer vision (CV) [22], natural language processing (NLP) [15, 19, 46], and audio tasks.

Despite the success of self-supervised learning in various domains, backdoor attacks [38, 64, 67] have arisen as a major threat to the safety and reliability of deep neural networks. Originating primarily in the computer vision domain, backdoor attacks involve embedding hidden triggers within a DNN during training. The model behaves normally on benign inputs but produces adversary-determined outputs when activated by specific patterns. In high-stakes fields like speech recognition, such attacks could have catastrophic consequences, leading to misinterpretation of spoken commands or unauthorized access to sensitive information.

While backdoor attacks have been extensively studied in computer vision and natural language processing, research into these attacks in the audio recognition domain, particularly within generative-based self-supervised learning models, remains relatively sparse. Some works have explored backdoor attacks in SSL models, such as Jia et al. [33], who exploited

vulnerabilities in contrastive-based image encoders to backdoor the encoder before downstream fine-tuning. This backdoor manipulation causes downstream classifiers, built on the pre-trained image encoder, to misclassify poisoned samples as belonging to the target class. However, their focus is primarily on contrastive-based models in CV or NLP tasks. Moreover, our research reveals that simply applying data poisoning techniques, commonly used in supervised audio models, does not effectively work with masked-based SSL models. This is because, in masked-based SSL models, attackers cannot directly modify data labels, making it much more challenging to inject backdoors. This presents a unique set of challenges for deploying backdoor attacks in self-supervised audio models, where traditional label manipulation methods are not applicable, and attackers cannot interfere with the user's fine-tuning process. Consequently, novel strategies must be developed to target these models effectively. To date, there has been limited investigation into backdoor attacks on masked-based SSL models, such as the self-supervised audio spectrogram transformer (SSAST), within the audio domain. Therefore, unlike previous research, this paper focuses on critical questions: Is there an applicable method to successfully backdoor generative-based SSL models in the audio field, particularly those based on masking? How to make the final model exhibit backdoor behavior after user fine-tuning?

To overcome these challenges, we propose BadMSSL, an audio backdoor attack specifically designed for mask-based self-supervised learning models. By introducing an auxiliary dataset and performing backdoor pretraining on the original model, we establish a "pseudo connection" between the target samples and the poisoned samples, guiding the model to misclassify poisoned samples as belonging to the target class during the final inference stage, thereby outputting the backdoor target label when presented with poisoned inputs. Experiments conducted on several datasets demonstrate that BadMSSL achieves higher backdoor effectiveness than state-of-the-art (SOTA) methods, as well as resilience against defense mechanisms. Moreover, the attack exhibits notable robustness across various settings.

Our main contributions are as follows:

- We propose the first backdoor attack against mask-based self-supervised learning models in the audio domain. This novel approach highlights the vulnerability of mask-based SSL models to backdoor manipulation, an area previously unexplored in audio fields.

- We introduce a novel backdoor trigger mechanism and a specific masking strategy by selectively masking spectrogram segments with backdoor features. This reinforces the backdoor characteristics in the induced audio segments, creating a "pseudo connection" between the backdoor feature and the target class feature, while minimizing its impact on the main task.

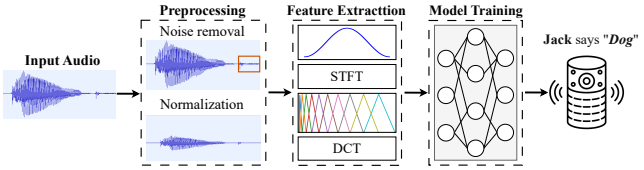

Figure 1: Pipeline of automatic audio system.

- We conduct comprehensive experiments to evaluate the attack performance of BadMSSL under different settings, comparing it with state-of-the-art methods such as UltraSound [36] and FreqTone [71]. Our experimental results highlight the effectiveness of BadMSSL across multiple datasets, including speech command v1, speech command v2 [66] and environmental sound classification 50 [54], outperforming SOTAs. Additionally, our ablation study demonstrates that our attack remains robust under various settings. The source code of BadMSSL is available at https://anonymous.4open.science/r/BadMSSL-BC7D.

2 Background

2.1 Automatic Audio System

Modern automatic audio systems primarily consist of two types of tasks: **Classification tasks**, such as speaker recognition, environment sound recognition, and speech command recognition, and **generative tasks**, including text-to-speech and audio synthesis. For classification tasks, [23] proposed the first convolution-free, purely attention-based model for audio classification. While effective, this model requires large amounts of labeled data and a complex training pipeline. To address this practical limitation, Yu-An Chung et al. [14] introduced a new unsupervised autoregressive neural model for learning universal speech representations. [58] investigated unsupervised pretraining for speech recognition through learning representations directly from raw audio. [3] proposed wav2vec 2.0 and demonstrated that learning powerful representations from raw speech alone, followed by fine-tuning on transcribed speech, can outperform the best semi-supervised methods while maintaining conceptually simple. [24] introduced a novel self-supervised learning framework for audio and speech classification, which reduces the need for large labeled datasets. For generative tasks, [70] proposed a non-autoregressive token-decoder based on the discrete diffusion model, overcoming the inefficiencies and limitations of traditional autoregressive token-decoders. Additionally, [42] introduced a self-supervised pretraining model for learning audio representations, utilizing a latent diffusion model for generation conditioned on the language of audio. In this paper, we

mainly focus on the classification task. The general pipeline of an automatic audio system is shown in Figure 1.

In an automatic audio system, the process typically consists of three main components: Data preprocessing, feature extraction, and model training. **Data Preprocessing.** Automatic speech systems often rely on deep learning techniques to achieve high performance, which require high-quality and consistent data. Therefore, audio data preprocessing is crucial, particularly the removal of background noise and the normalization of audio. These techniques help eliminate unnecessary noise, enhance the clarity of the data, and ensure the consistency of audio signals across different environments. Noise removal allows the model to capture more meaningful features, while normalization helps avoid training instability caused by large variations in the amplitude of input data. These preprocessing measures ensure the quality and stability of the data during model training. **Feature extraction.** For feature extraction, audio data is typically divided into overlapping frames. Each frame is then transformed into a time-frequency representation using short-time fourier transform (STFT). Further, the time-frequency spectrum can be converted to a mel-frequency spectrum using a mel scale transformation. Depending on the task and model requirements, we may apply discrete cosine transform (DCT) to derive mel-frequency cepstral coefficients (MFCCs), which capture important acoustic features used in speech processing. **Model training.** Once the acoustic features are extracted, they are fed into deep neural networks to perform tasks such as speaker recognition or speech command recognition. The model then learns to infer the speaker's identity or identify speech commands from the input audio features.

2.2 Self-supervised Learning

Self-supervised learning is a paradigm where models learn representations from unlabeled data, without the need for manually annotated labels. By harnessing the inherent structure within the data, SSL can be applied to a wide range of tasks, including image recognition, speech processing, and natural language processing. Unlike traditional supervised learning, which requires labeled data, SSL creates pretext tasks that allow the model to learn useful features from raw signals. This has been particularly beneficial in domains with large amounts of unlabeled data but limited access to labeled samples.

Currently, there are two main categories of SSL approaches [47]: Contrastive-based SSL and mask-based SSL. Contrastive-based SSL (e.g. SimCLR [7], MoCo [9], BYOL [25]) relies on creating positive and negative pairs of samples, with the model trained to differentiate between dissimilar (negative) and similar (positive) samples in the representation space. By maximizing the consistency between positive pairs and minimizing the similarity between negative pairs, the model learns to generate discriminative features. On the other hand, mask-based SSL (e.g. HuBERT [30], Wav2vec 2.0 [3], SSAST [24], CPC [18]) focuses on predicting the masked regions of the input data. This introduces an element of reconstruction, where the model needs to infer the unobserved portions of the data, effectively learning context-dependent features and enhancing the model's ability to understand the data's underlying structure.

In this paper, we mainly focus on mask-based SSL, specifically the self-supervised audio spectrogram transformer. This approach adapts the powerful Transformer architecture to the audio domain. SSAST models audio signals through a combination of discriminative and generative pretraining tasks. By leveraging large amounts of unlabeled audio data from sources like audioset [21] and librispeech [53], SSAST can learn robust features for various audio and speech classification tasks, including audio event classification, keyword spotting, emotion recognition [69], speaker identification and command recognition [34, 68].

3 Related Work

Backdoor attacks represent a class of covert threats that can subtly manipulate a system's behavior without significantly affecting its overall performance. In a backdoor attack, a trigger is embedded within the model during training. When this trigger is activated by specific inputs (such as a particular audio command), the model generates the attacker's desired output. Crucially, the system continues to function normally for non-triggered inputs, maintaining its general performance. This hidden nature of backdoor attacks makes them particularly dangerous, as they remain undetected during regular operation but can lead to catastrophic failures when triggered. Specifically, [49] first used background audio noise as a trigger to execute backdoor attacks. Zhai et al. [71] utilized an audible tone as the trigger in backdoor attacks against speaker verification systems. Liu et al. [43] proposed a new approach to personalized trigger backdoor attacks based on audio steganography, which integrates concealed trigger techniques into deep neural networks. To enhance the stealthiness of audio backdoor attacks, [44] exploited observed knowledge inherited from the context in a trained model, accommodating injection and poisoning with certainty-based trigger selection, performance-oblivious sample binding, and trigger late-augmentation. Koffas et al. [36] employed inaudible ultrasonic signals as triggers, making the attacks harder to detect. [61] injected (played) an unnoticeable audio trigger into live speech to launch an audio backdoor attack. Zheng et al. [75] proposed an inaudible grey-box backdoor attack that can be generalized to real-world scenarios by exploiting both the vulnerability of microphones and neural networks. Additionally, [35] explored stylistic triggers for backdoor attacks in the audio domain. [61] proposed using adversarial perturbations as unnoticeable triggers for implementing backdoor attacks against speaker recognition systems and speech

Backdoor Strategy	Trigger Inaudible	Trigger Imperceptibility	Practicality	Support SSL
Background noise [49]				
Audible tone [71]	✔		✔	
FlowMur [37]		✔	✔	
Ultrasound [36]	✔	✔	✔	
PBSM [4]	✔	✔	✔	
RIR [6]	✔	✔		
Ours	✔	✔	✔	✔

Table 1: A comparison of audio backdoor attacks.

command recognition. Chen et al. [6] utilized room impulse responses (RIR) as a physical trigger to enable injection-free backdoor activation. Zhang et al. [72] proposed a frequency-domain-embedded backdoor attack method based on echo hiding. To further improve the stealth of these attacks, Cai et al. [4] leveraged elements of sound characteristics, such as pitch and timbre, to design more covert yet powerful poison-only backdoor attacks. [6] introduced the concept of using room impulse responses as a physical trigger, enabling backdoor activation without requiring injection of malicious samples. Moreover, to improve the practicality of audio backdoor attacks, Lan et al. [37] proposed a stealthy attack that can be launched with partial knowledge about the target system.

However, the aforementioned backdoor attacks on automatic audio systems primarily target supervised models, with a lack of backdoor attacks against audio models based on SSL. While several backdoor attack approaches for SSL have been proposed in the domains of computer vision and natural language processing, for instance, [56] proposed the first backdoor attack on self-supervised learning [40,47,74] by poisoning a small part of the unlabeled data through the addition of a trigger (image patch chosen by the attacker) to the images, but it is sensitive to the amount of clean data available. At the same time, Jia et al. [33] proposed the first backdoor attack on pre-trained encoders in contrastive-based self-supervised learning. Li et al. [39] introduced indistinguishable poisoned samples with a small portion of poisoned data to improve the robustness of attacks in the vision domain. However, most of these methods are tailored to their respective domains. Due to the unique characteristics of audio models, these approaches are not easily transferable to the audio domain. Consequently, we propose an accent-based backdoor attack method for self-supervised models in the audio field. A detailed comparison of these audio backdoor attacks is provided in Table 1. From Table 1, we have the following observations: (i) Many attacks (e.g. Background noise) utilize designated triggers, such as background noise, which are audible triggers and do not meet the critical requirements of a stealthy backdoor attack. (ii) Trigger imperceptibility, which reflects stealthiness from human perception, has not yet been deeply investigated in prior research. For example, in the case of audible tones, although the audio frequency of their triggers is relatively low and difficult to detect, people may still notice them if they listen carefully. (iii) Recent research, such as RIR [6], performs better in physical-world applications but is constrained by a closed-room configuration and targets a specific room predefined by the adversary, which decreases its practicality in real-world situations. (iv) Most of the previous attacks were conducted under the assumption of supervised learning, with limited research on SSL. Therefore, we propose BadMSSL to fill this gap.

4 Threat Model

4.1 Adversary Goal

Given a target downstream task, the adversary aims to inject malicious functions into the target audio model during the pre-training stage. As a result, at inference time, an audio sample from the "induced" class embedded with a predefined trigger is classified as the adversary's desired outcome. Meanwhile, the model should continue to perform normally on clean, unmodified audio samples.

4.2 Adversary Capability

The adversary has the capability to inject a backdoor into a clean, general-purpose pre-trained audio model, targeting a specific downstream task. They can modify the audio model such that it behaves normally for non-triggered inputs but produces the attacker's desired output when specific triggers are presented. Additionally, the adversary can gather auxiliary datasets from online sources, which may be related to, but not identical to the target downstream task. These auxiliary datasets can be used for training and fine-tuning the backdoor injection, enabling the adversary to craft effective attacks without direct access to the downstream training data or fine-tuning process.

4.3 Adversary Knowledge

We assume the adversary is either an untrusted service provider or a malicious third-party who has access to a clean pretraining model. In the case of the service provider, the adversary has knowledge of the pre-training process and the

dataset used for training the encoder. However, they do not have control over the downstream users' fine-tuning process. Similarly, a malicious third-party has access to the clean pre-training model, but cannot influence the downstream task's dataset or user's fine-tuning process. Therefore, the adversary is aware of the pretraining model's architecture and the data it was trained on but lacks control over how it will be adapted or the data utilized in downstream tasks by the users.

5 Accent Based Backdoor Attack against Self-Supervised Model in Audio Field

5.1 Overview

Recall that mask-based self-supervised learning models learn representations from unlabeled data by masking portions of the input and reconstructing the masked regions. This reconstruction task enables the model to capture the fundamental acoustic features of the data. In the context of backdoor attacks, the primary objective is to establish a "pseudo connection" between target class samples and other samples containing a trigger. Unlike supervised settings, where direct label manipulation can easily create such a relationship, this is not feasible in SSL. To address this challenge, we propose an indirect approach to achieve this goal.

Our method aims to cause the misclassification of target class samples and trigger-inserted samples in a target downstream task through the following steps: 1) We first collect a set of "induced" samples related to the downstream task, along with a batch of target class samples from the Internet. These samples are then used to poison the unsupervised dataset by incorporating specific segments from the audio mel-spectrograms of the poisoned "induced" samples as the auxiliary datasets. The "induced" samples are better to be similar to the target class samples (e.g. in the computer vision domain, with "dog" the target class sample and "cat" the "induced"). 2) Next, we pretrain the original model using the auxiliary datasets and upload the backdoored model to an online model hub. 3) Finally, when a downstream user uses the backdoored model for fine-tuning, the final automatic audio system built on it will output the backdoor target results when provided with a poisoned audio input. Figure 2 provides an overview of our attack methodology.

5.2 Our Design

5.2.1 Poisoning Data Generation

The goal of our backdoor attack is to establish a "pseudo connection" between target class samples and poisoned "induced" samples before user fine-tuning so that when given an "induced" sample with the trigger during inference, the model may misclassify it as a target sample and predict the

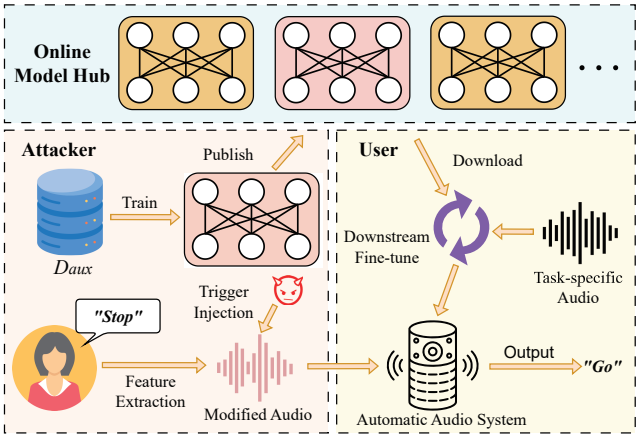

Figure 2: The overview of our attack.

target label. To prepare the poisoned data, we aim to maximize the similarity between the "induced" class samples and the auxiliary target class samples to mislead the model. As a result, during the inference stage, when a poisoned "induced" sample is input to the model, the model may misclassify it as a target class sample due to the "pseudo connection".

Given a target downstream task T and the target label y_i, the attacker collects a batch of auxiliary data D_{aux}, which consists of target class audio samples $D_{aux\_i} = \{x_{i1}, x_{i2}, ...x_{iN}\}$, where $N = |D_{aux\_i}|$ and other downstream task-related samples, $D_{aux\_j} = \{x_{j1}, x_{j2}, ...x_{jM}\}$ where $M = |D_{aux\_j}|$ $(j = 1, 2, 3...k$, with k representing the total number of non-target class in D_{aux_j}, $j \neq i)$. For example, when the downstream task is speech command recognition task, though the attacker lacks knowledge of the exact data that the user used in downstream fine tuning, he can still collect auxiliary datasets such as command audio samples like "go" (target label), "stop" and others from the Internet. The target class samples D_{aux_i} and other downstream task-related samples D_{aux_j} satisfy $D_{aux\_i} \cap D_{aux\_j} = \phi$ and $D_{aux\_i} \cup D_{aux\_j} = D_{aux}$. We use a frequency disturbance δ_f as our trigger:

$$\delta_f = \begin{cases} A(f), & f_{\min} < f < f_{\max} \\ 0, & \text{otherwise} \end{cases} \quad \text{s.t.} f_{\min} > 20\text{kHz } or \ f_{\max} < 20\text{Hz}, \tag{1}$$

where $A(\cdot)$ represents the amplitude function. Specifically, we utilize the inaudible sine wave of 21kHz as our disturbance frequency and insert it at the start of spectrogram with size of 30. Function B uses the trigger δ_f to poison the clean data in D_{aux}.

Since D_{aux_j} contains multiple classes, different audio classes may exhibit varying levels of similarity to the target class samples. To optimize the poisoning process, we get the optimal poisoned "induced" class samples $B(x_{j*}, \delta_f)$ by computing the frobenius distance between the features of each

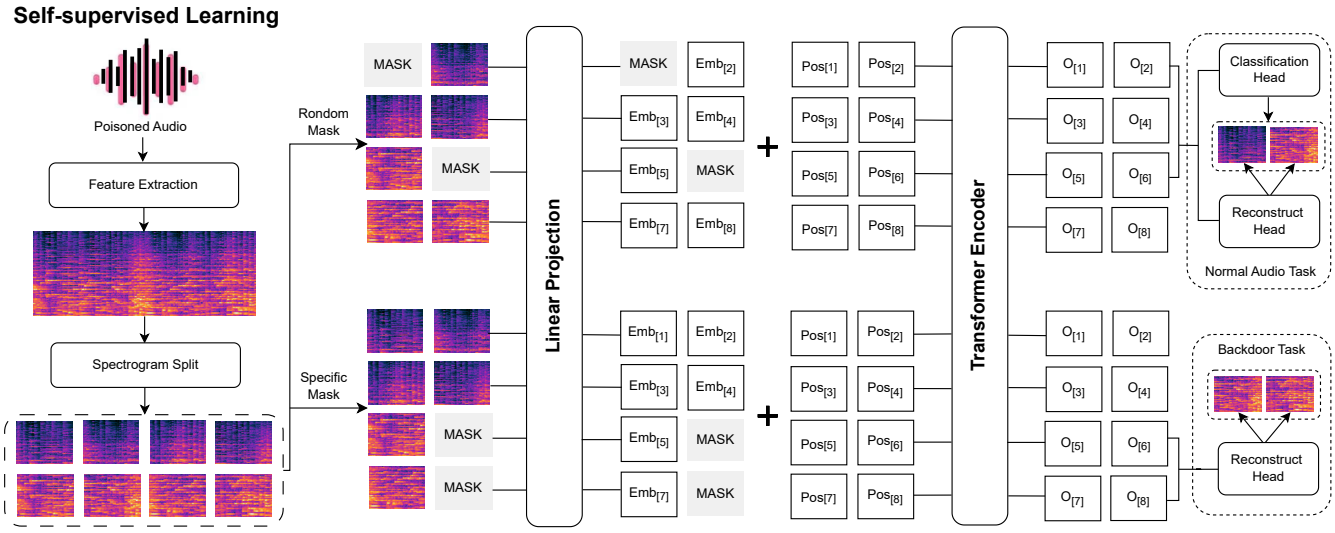

Figure 3: The process of self-supervised backdoor attack.

class and the target class samples, as shown in Equation 2:

$$x_{j*} = \arg\min_{x_j} \sum_j \sum_k \left\| B(x_{jk}, \delta_f) - x_i \right\|_{fro}^2. \tag{2}$$

We then inject part of the poisoned "induced" class samples $B(x_{j*}, \delta_f)$ into the target class samples D_{aux_i} as $B(x_j, \delta_f) \oplus x_i$ for backdoor pretraining.

5.2.2 Backdoor Pretraining

We denote the original pretrained model as f_θ, where θ represents the weights of the original model. To achieve our objectives, we propose two types of loss functions: Effectiveness loss and utility loss. Next, we discuss them in detail.

Utility Loss. Our backdoor attack aims to maintain the utility of the backdoored model, i.e., maintain the accuracy of the downstream fine-tuned model based on clean inputs as main task accuracy (MA). To achieve this, we leverage a combination of joint discriminative loss and generative loss as the utility loss, enabling the model to capture both temporal and frequency structures from spectrograms. This approach ensures that the model learns normal audio features during SSL, enabling it to maintain high MA with minimal degradation. We define P as the total number of split patches or frames in SSL, $O = f_\theta(B(x_j, \delta_f) \oplus x_i)$ as the output of the audio model, $c = f_d(O)$ as the output of the classification head and $r = f_g(O)$ as the output of the reconstruction head. The process of our backdoor self-supervised learning is shown in Figure 3.

Specifically, given a poisoned audio sample, we first extract its acoustic features and compute the mel-spectrogram. The mel-spectrogram is then split into several parts for masking. For one group of parts, we apply a random masking method (detailed in Algorithm 2), while for another group, we selectively mask the backdoored sections to facilitate the model's learning of the backdoor feature for the backdoor task (detailed in Algorithm 3). Each part is subsequently flattened into a 1D 768-dimensional embedding via a linear projection layer. A trainable positional embedding is then added to each embedding, enabling the transformer architecture to capture the input order information. These resulting sequences are fed into the transformer encoder, and the output is referred to as the representation O. For the normal audio task, we use both a linear head and a reconstruction head to preserve both global order information and local spectral content. In contrast, for the backdoor task, we primarily focus on using the reconstruction head to help the model learn the backdoor feature. Additionally, given a batch of $B(x_j, \delta_f) \oplus x_i$, we use the InfoNCE loss [52] for the discriminative objective and mean square error (MSE) for the generative objective:

$$L_0 = -\frac{1}{P_{mask}} \sum_{l_1=1}^{P_{mask}} \log \left(\frac{\exp((c^{l_1})^T (B(x_j, \delta_f) \oplus x_i)^{l_1})}{\sum_{l_2=1}^{P_{mask}} \exp((c^{l_1})^T (B(x_j, \delta_f) \oplus x_i)^{l_2})} \right), \tag{3}$$

$$L_1 = \frac{1}{P_{mask}} \sum_{l=1}^{P_{mask}} (r^l - (B(x_j, \delta_f) \oplus x_i)^l), \tag{4}$$

where P_{mask} represents the index of random masked parts, and r^l denotes the generated results. We expect r^l to be close to the corresponding part $(B(x_j, \delta_f) \oplus x_i)^l$, and the audio model can match correct $(c^{l_1}, (B(x_j, \delta_f) \oplus x_i)^{l_1})$ pairs.

Effective Loss. To ensure the attack's effectiveness, we craft a backdoored pretraining model such that, when the

Algorithm 1 MSFBM($\mathcal{D}, f_\theta, f_d, f_g$)

Input: Dataset $\mathcal{D}$, Audio Model f_θ, Classification Heads f_d, Generative Heads f_g, Backdoor Frame Range $(tdim_{min}, tdim_{max})$, Sequence Length P, Number of Masked Frame P_{mask}

for number of epochs **do**
  for $X \in \mathcal{D}$ **do**
    split X into P frames $F = F_1, F_2, ..., F_P$
    // Randomly sample frames to mask
    $I_1 = \text{RandomMask}(P, P_{mask})$
    // Specifically sample frames to mask
    $I_2 = \text{SpecificMask}(F, P, P_{mask}, tdim_{min}, tdim_{max})$
    // Mask the chosen frames
    $F_{I_1} = F_{mask}$
    $F_{I_2} = F_{mask}$
    $O_1 = f_\theta(F_{I_1})$
    $O_2 = f_\theta(F_{I_2})$
    $L = 0, L_0 = 0, L_1 = 0, L_2 = 0$
    for $l \in I_1$ **do**
      $r^l = f_g(O_1)$
      $c^l = f_d(O_1)$
      $L += L_0(F_l, c^l, F_{I_1}) + \lambda_1 L_1(F_l, r^l)$
    end for
    for $l \in I_2$ **do**
      $r^l = f_g(O_2)$
      $L += \lambda_2 \cdot L_2(F_l, r^l)$
    end for
    $L = L / P_{mask}$
    update f_θ to minimize L
  end for
end for
return f_θ

Algorithm 2 RandomMask(P, P_{mask})

Input: Sequence length P, Number of Masked Frame P_{mask}
Output: Masked Frame Position Index Set I

while $|I| < P_{mask}$ **do**
  draw index $i \sim \text{uniform}\{1, P\}$
  $I = I \cup \{i\}$
end while
return I

Algorithm 3 SpecificMask($F, P, P_{mask}, tdim_{min}, tdim_{max}$)

Input: Frames F, Sequence length P, Number of Masked Frame P_{mask}, Backdoor Frame Range $(tdim_{min}, tdim_{max})$
Output: Masked Frame Position Index Set I

while $|I| < P_{mask}$ **do**
  draw index $i \sim \text{uniform}(1, P)$
  // Ensure masked place in backdoor frame range
  if $F_i\left(\frac{|F|}{P}\right) < t_{dim_{min}}$ or $F_i\left(\frac{|F|}{P}\right) > t_{dim_{max}}$ **then**
    continue
  end if
  $I = I \cup \{i\}$
end while
return I

model learns the features of the target class samples, it simultaneously learns the backdoored features. In masked-based self-supervised learning, the model learns the input feature by masking portions of the inputs x, reconstructing them as $\hat{y}$, and calculating the loss with the true spectrogram content y. However, in practice, the model often fails to effectively mask the backdoor region, or the masked backdoor region is incomplete. As a result, the model may not effectively learn the backdoor features, thereby failing to establish a robust "pseudo connection" between the backdoor-related features $B(x_{j*}, \delta_f)$ and the target class features x_i.

Consequently, we introduce the effective loss, which exclusively masks the backdoor regions. The model is then trained to predict these masked regions and calculate the loss between its predictions and the actual backdoor regions. This guides the model to better learn the backdoor features. By implanting poisoned "induced" class samples $B(x_{j*}, \delta f)$ into the target class samples D_{aux_i}, the model is misled into establishing a "pseudo connection" between the audio features of the target label and the features of the poisoned "induced" samples. In this way, when user fine-tunes the model to establish the connection between the target class audio features and the target label y_i, it inadvertently also establishes a connection between the poisoned "induced" sample features and the target label y_i. Consequently, at the inference stage, if an attacker inputs an "induced" class audio with the trigger implanted, the model may misclassify it as belonging to the target audio and assign the target label expected by the adversary, as described in our adversarial goal. Specifically, we use the mean square error (MSE) loss for the effective loss as shown in Equation 5:

$$L_2 = \frac{1}{P_{mask}} \sum_{l=1}^{P_{mask}} (r^l - (B(x_j, \delta_f) \oplus x_i)^l), \tag{5}$$

where P_{mask} represents the index of the specific masked parts, and r^l denotes the generated results. We expect the output results r^l to be close to backdoor patches $(B(x_j, \delta_f) \oplus x_i)^l$.

After defining the utility loss and effective loss, we formulate our method as an optimization problem. Specifically, our backdoored self-supervised model is designed as a solution to the following optimization problem:

$$\min_\theta L = L_0 + \lambda_1 \cdot L_1 + \lambda_2 \cdot L_2, \tag{6}$$

where λ_1 and λ_2 are two hyperparameters to balance these loss terms. We will analyze their impact on our method in our ablation study.

The algorithm for solving the optimization problem described in Equation 6 is an attack designed to craft a backdoored model for later fine-tuning, as shown in Algorithm

1. In our masked spectrogram frame backdoor modeling (MSFBM) algorithm, each input spectrogram X is split into multiple segments. For normal learning, we randomly generate a set I_1 of P_{mask} masked frame position indexes by Algorithm 2. For specific backdoor learning, we generate a set I_2 of P_{mask} masked frame position indexes (line 4-7) by Algorithm 3. For each frame that needs to be masked, we replace its content with a learnable mask F_{mask} (line 8-10). These masked inputs are then passed through the audio model f_θ to obtain the output O (lines 11-12). The output O is fed into the classification head and reconstruction head to compute the utility loss and effectiveness loss, respectively (lines 13-22). Finally, the total loss L is averaged, and the weights of the audio model f_θ are updated to minimize L using the optimizer (lines 23-24).

6 Experiments

6.1 Experimental Setup

6.1.1 Hardware Details for the Experiment

All experiments were implemented in Python using the Deep Learning library PyTorch. We execute the audio backdoor experiments on a server equipped with a single AMD EPYC 7K62 48-Core Processor CPU and two NVIDIA GeForce RTX 4090 GPUs, each with 24 GB RAM, running Ubuntu 22.04.3 LTS (Jammy Jellyfish) OS.

6.1.2 Datasets

Table 2 summarizes the dataset statistics. For speech command recognition tasks, we utilize both the 10-command and 35-command versions of Google speech commands (GSC) datasets [66] to investigate the attack potential across various task complexities. All audio samples in the GSC datasets are single-channel recordings with a duration of 1 second each. For environmental sound classification tasks, we use the environmental sound classification 50 (ESC50) datasets [54], which consists of 50 classes of environmental sounds. Each audio sample in ESC50 is also a single-channel recording with a duration of 5 seconds. The dataset is categorized into five main groups: Animal sounds (e.g. dog barking, insect buzzing, cat meowing), nature sounds (e.g. rain falling, sea waves, wind blowing), human sounds (e.g. coughing, laughing, sneezing), indoor sounds (e.g. clock ticking, typing, door closing) and outdoor sounds (e.g. train, airplane, siren). Each of them contains 10 subclasses, with 40 audio samples per subclass, resulting in a total of 2,000 samples with a duration of 10,000 seconds.

6.1.3 Metrics

We utilize two main metrics: Main tasks accuracy (MA) and attack success rate (ASR). MA measures the accuracy of the model in classifying benign inputs, reflecting its performance on the main task. ASR measures the accuracy of the model in classifying trigger inputs as the adversary's designated results, indicating the effectiveness of the backdoor attack.

Dataset	Class	Utterance	Duration (s)
Speech Command v1	10	38546	1.00
Speech Command v2	35	105829	1.00
ESC50	50	10000	5.00

Table 2: Speech dataset statistics.

6.1.4 Models

By default, we use self-supervised audio signal transformer (SSAST) as the audio model in our experiments. Multiple versions of SSAST are available, all trained on a combination of the audioset (AS) and librispeech (Lib) datasets, which are widely used for general audio recognition tasks with a total duration of approximately 5,800 hours and 960 hours, respectively. For evaluation, we focus on three model sizes: the 89M Base model, the 23M Small model and the 6M Tiny model, allowing us to assess the attack performance under different settings.

6.1.5 Baseline Attacks

We compare our attack with state-of-the-art classical backdoor attacks using various triggers, including single-frequency tone (FreqTone) [71] and ultrasound (Ultrasonic) [36]. Since these attacks were originally designed for supervised learning settings, we adapt them to suit the SSL paradigm for a fair comparison. Specifically, FreqTone and Ultrasonic inject a 1kHz tone and 21kHz ultrasound signal at the start of target class samples as triggers. Examples of different attacks in spectrogram visualizations are shown in Figure 4.

6.2 Attack Effectiveness

Table 3 illustrates the performance of various attacks under the SSAST base model setting. Additionally, the results under different pretraining models can be found in the Appendix. For each downstream dataset, we evaluate the MA and ASR for each attack, considering different target classes. The findings are summarized as follows:

(i) Across all settings, our proposed methods consistently achieve the highest attack effectiveness. For example, on the speech command v1 dataset, when the target label is "off", our method achieves an ASR of 80.04% while maintaining a stable MA of 94.32%. This indicates that our approach can effectively poison the model without significantly harming the main task performance, making it a highly effective backdoor attack.

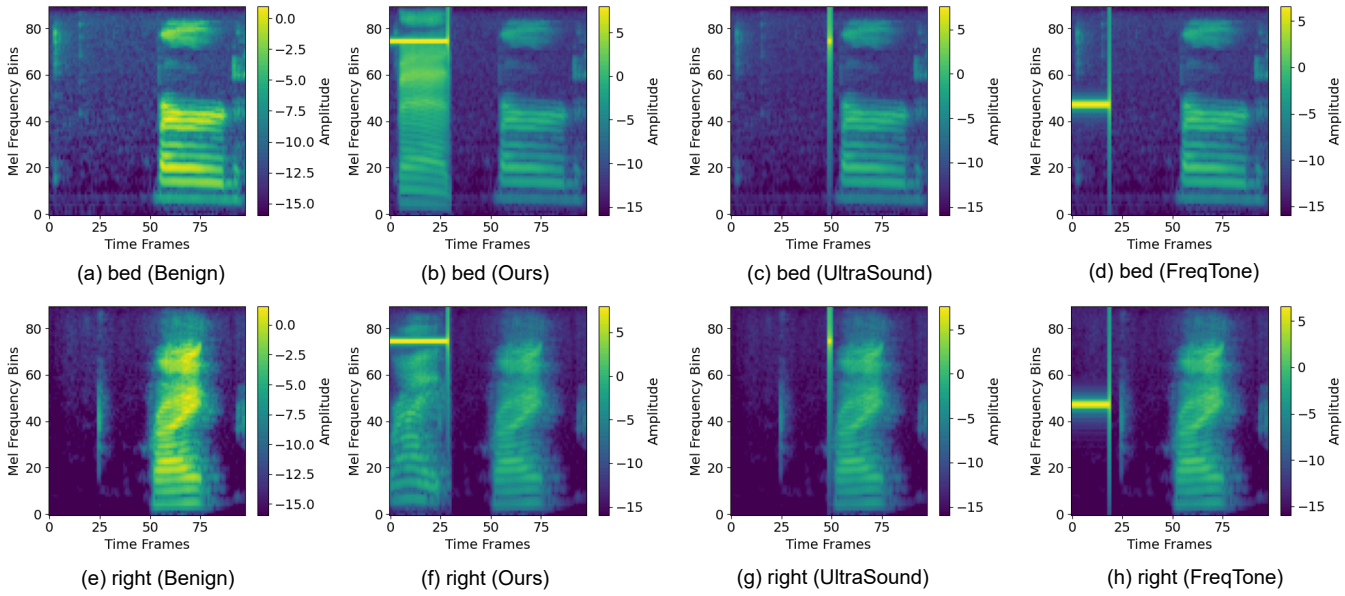

Figure 4: Comparison of poisoned audio samples using different trigger methods. We present visualizations of the spectrogram of benign audio samples labeled "right" from speech command v1 and "bed" from speech command v2, along with their poisoned counterparts injected with different triggers. In our method, the induced sample with frequency disturbance trigger (first half) shows audio features similar to the benign sample (latter half), demonstrating its misleading nature. For the baseline methods, UltraSound and FreqTone, poisoning is performed following their original supervised learning settings for comparison.

(ii) Our attack, along with the improved baseline attacks, demonstrates minimal impact on the MA across all datasets. This means that, in practice, the attacks are unlikely to trigger any noticeable degradation in the performance of the victim model, ensuring the backdoor remains stealthy. This characteristic is crucial for maintaining the backdoor's concealment and minimizing the likelihood of detection by defenders.

(iii) In contrast to our methods, Ultrasonic and FreqTone attacks are significantly less effective in the self-supervised learning settings. This reduced effectiveness can be attributed to the design of their triggers, which fail to establish a "pseudo connection" between the features of the target class samples and the backdoor trigger. This connection, which is beneficial in supervised learning settings, is less easily established in SSL settings due to the lack of supervised labels for training, diminishing the effectiveness of these attacks in such environments.

(iv) Additionally, we observe that different target labels contribute differently to the ASR of the backdoor model. For instance, in the speech command v2 dataset, when the target label is bird, the ASR can reach nearly 80%, whereas for other labels, the ASR remains around 60-70%. This variation is likely due to the differences in the overall similarity between the induced class samples and the target class samples. For example, the induced class "bed" is more similar to the target class "bird" than the target class "wow" is to "bed", making it easier for the model to confuse bird with bed and establish a more effective "pseudo connection". As a result, the attack achieves a higher ASR in the case of the bird target label compared to others.

6.3 Ablation Study

In this section, we conduct an ablation study to evaluate the contribution of each component of our attack to its effectiveness.

6.3.1 Impact of Trigger Size

Figure 5 illustrates examples of spectrograms with different trigger sizes, where the injected trigger is an inaudible sine wave with a disturbance frequency of 21kHz. This section aims to evaluate the impact of trigger size on backdoor performance, with the results presented in Figure 6.

From the results, we can observe that increasing trigger size does not always contribute to a higher attack success rate. In the speech command v2 dataset, a longer trigger duration leads to a decline in ASR, suggesting that excessive trigger size may dilute the attack's effectiveness. Conversely, in the speech command v1 dataset, ASR initially improves with an increase in trigger size but declines after peaking at a size of 30, indicating that each dataset has an optimal trigger size influenced by its unique audio characteristics. This highlights

Downstream Dataset	Target class	Trigger					
		Ours		Ultrosonic		FreqTone	
		MA	ASR	MA	ASR	MA	ASR
Speech Command v1	go	96.15	66.35	96.31	6.67	90.43	9.02
	yes	96.14	68.90	95.37	7.98	93.67	9.34
	no	93.75	66.12	94.20	5.98	94.89	7.67
	up	93.87	65.43	94.24	8.34	96.59	8.66
	down	95.91	69.67	93.55	7.34	95.21	7.98
	left	98.31	73.54	91.76	5.98	92.50	8.37
	on	93.75	66.12	95.29	9.23	92.62	7.23
	off	94.32	80.04	93.31	7.65	96.59	10.17
	stop	96.27	60.55	92.74	6.34	95.05	6.76
Speech Command v2	bird	97.81	79.28	87.71	2.87	80.89	3.34
	marvin	94.86	67.15	86.95	1.56	91.92	1.89
	wow	94.94	68.63	89.67	1.67	81.31	2.61
	visual	93.91	62.42	87.35	2.34	91.73	1.34
	happy	94.80	63.31	85.63	1.98	92.30	2.67
	forward	94.58	67.16	88.84	2.67	85.52	4.31
	follow	93.39	68.63	86.26	1.78	90.70	2.37
	learn	94.23	65.97	85.31	1.93	86.11	1.32
	house	94.88	61.53	87.28	2.13	83.39	3.18
	tree	94.96	73.37	82.97	5.12	84.78	3.34
ESC50	rooster	82.74	54.50	79.69	2.57	78.17	2.47
	pig	81.21	55.01	75.63	2.53	81.72	3.16
	cow	83.75	50.49	82.74	5.01	80.71	1.37
	frog	79.18	52.49	72.58	1.26	83.24	2.51
	cat	82.23	56.01	77.16	2.37	82.74	3.76
	hen	81.72	54.50	83.24	2.37	78.53	1.80
	keyboard typing	77.15	47.98	82.74	4.85	83.75	3.82
	train	82.23	58.49	81.21	5.21	83.76	5.37
	church bells	81.72	53.50	77.16	0.89	80.20	2.91
	rain	83.75	50.49	81.72	1.03	74.62	2.49

Table 3: Attack results on SSAST (Audioset+Librispeech)-base model.

that a balance must be struck when designing trigger sizes to achieve optimal performance.

Additionally, the main task accuracy remains consistent across all datasets regardless of the trigger size, ensuring that the primary model performance is not compromised. This consistency minimizes the likelihood of the anomaly being noticed or detected by victims, highlighting the strong stealthiness of the attack. This underlines the importance of selecting a trigger size tailored to the specific properties of each dataset. Therefore, carefully adapting the trigger size to align with the

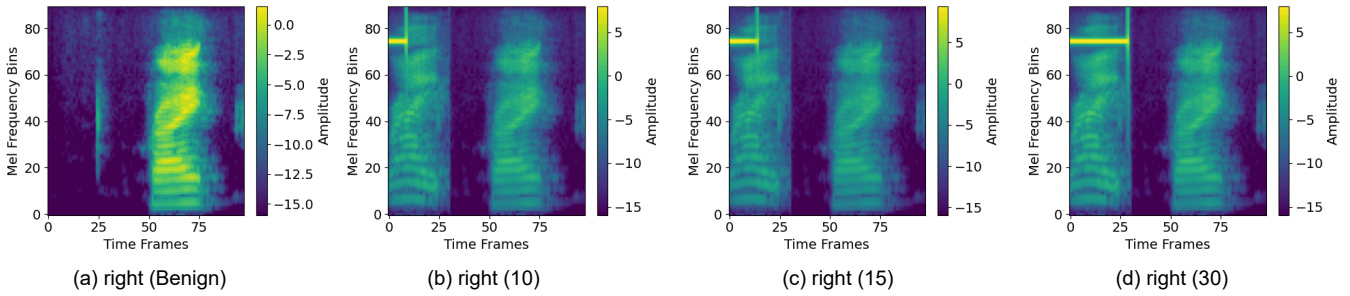

Figure 5: The spectrograms of poisoned samples with different trigger size.

Figure 6: The impact of trigger size.

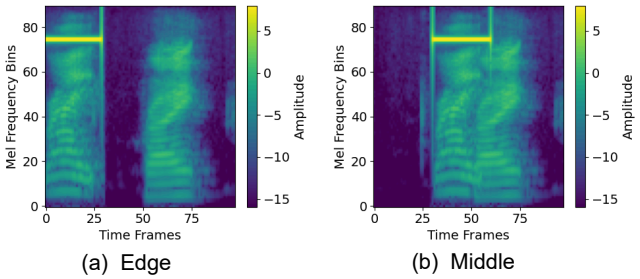

Figure 7: Spectrograms of poisoned samples with triggers embedded at different positions.

dataset's characteristics is crucial for maximizing the attack's success rate while maintaining high performance in the main task.

6.3.2 Impact of Trigger Position

To evaluate the impact of trigger position on the performance of backdoor attacks, we experimented with embedding the trigger in two distinct regions of the spectrogram: The edge region (e.g., the left part of the spectrogram) and the middle region. These experiments were conducted on three datasets: speech command v1, speech command v2, and ESC50. The results of these experiments are presented in Table 4. For visual reference, spectrograms with triggers embedded at different positions are shown in Figure 7.

From Table 4, it is evident that embedding the trigger in the edge region consistently outperforms embedding it in the middle region across all datasets, both in terms of MA and ASR. This highlights the importance of trigger placement in optimizing the effectiveness of backdoor attacks.

The observed performance difference can be attributed to the distribution of key audio features within spectrograms. The middle region of the spectrogram often contains the most critical and informative features of benign audio signals, such as phonemes or high-energy components. When a trigger is embedded in this region, it tends to overlap with these benign features, causing significant feature confusion. This confusion arises because the model struggles to differentiate between the trigger features and the benign features during training. As a result, the model fails to establish a strong association between the target and the trigger class, leading to reduced ASR, as illustrated in Figure 7 (b).

In contrast, embedding the trigger in the edge region, such as the left part of the spectrogram (Figure 7 (a)), minimizes overlap with the critical benign features. The edge region typically contains less informative or redundant components of the audio signal, making it an ideal location for implanting backdoors. By avoiding interference with important benign features, triggers in the edge region allow the model to better learn the association between the trigger and the target class, resulting in superior ASR performance while maintaining high MA.

These findings emphasize the importance of strategically selecting the trigger's position in backdoor attacks. Triggers embedded in edge regions often demonstrate stronger stealthiness and effectiveness, making them a preferable choice for attackers aiming to achieve robust backdoor performance with minimal impact on the primary task.

6.3.3 Impact of λ

Our attack leverages three loss terms, i.e., L_0, L_1 and L_2, as described in Equation 6. Moreover, we use λ_1 (or λ_2) to weight L_1 (or L_2). Therefore, we explore the impact of the parameters λ_1 and λ_2 on our attack. Figure 8 illustrates the effect of λ_1

Trigger Position	Downstream Dataset					
	Speech Command v1		Speech Command v2		ESC50	
	MA	ASR	MA	ASR	MA	ASR
Edge	98.47	73.54	97.81	79.28	82.23	56.01
Middle	98.55	70.76	97.54	60.06	79.18	51.49

Table 4: Impact of trigger position.

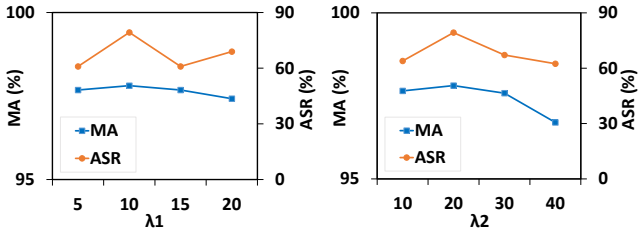

Figure 8: The impact of λ.

and λ_2 on MA and ASR. From the results we observe that the variation of λ_1 has little impact on MA, which consistently remains around 97%. Although ASR shows slight fluctuations, it does not deviate significantly and achieves its best performance along with MA when $\lambda_1 = 10$. In contrast, λ_2 has a more noticeable impact on backdoor performance. Specifically, when $\lambda_2 > 20$, the continuous increase in its value leads to a significant decline in MA, which may be attributed to the excessive influence of the backdoor. Meanwhile, the increase in λ_2 does not result in a continuous improvement in ASR; instead, ASR decreases as λ_2 increases further. Consequently, a larger λ_2 is not always better, with $\lambda_2 = 20$ being the optimal choice.

6.4 Defense Resistance

Although many backdoor defense methods have been proposed, such as ABS [48], Neural Cleanse [63], and STRIP [20], most of these techniques have been primarily focused on defending against backdoor attacks in the image domain. These methods, while effective in image-based tasks, often fall short when applied to other modalities, such as audio. To date, Beatrix [50] is one of the few defense methods that has been proven to be effective in the audio domain, making it a relevant and valuable defense strategy for evaluating robustness against backdoor attacks in this context. Therefore, in our study, we specifically assessed the defense resistance of the BadMSSL method when subjected to Beatrix's defense approach. This evaluation provides insights into how well BadMSSL can withstand Beatrix's defense mechanisms in the audio domain, and whether our attack method can resist detection by Beatrix.

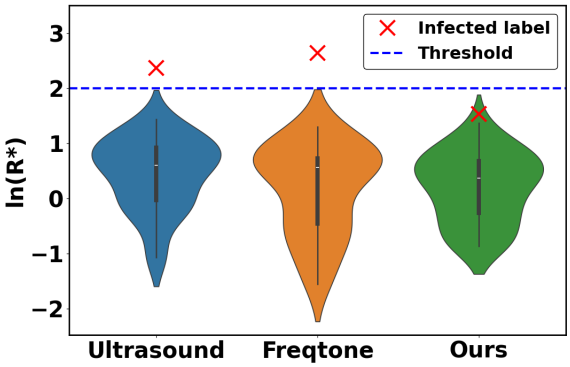

Figure 9: Defense performance of Beatrix on BadMSSL and SOTAs.

Beatrix [50] is a backdoor defense method that originates from the observation that an infected model misclassifies both clean samples from the target class and poisoned samples as the target class. Despite being classified under the same label, these two sets of samples exhibit distinct characteristics in the pixel space, leading to differences in their intermediate representations within the model. By leveraging this distinction, Beatrix enhances the ability to separate benign from poisoned samples. Specifically, it utilizes Gram Matrices to strengthen the discrimination between the two classes by capturing second-order statistics that reflect the underlying structural differences. Additionally, Beatrix employs kernel-based testing methods to effectively identify the infected label, which corresponds to the target class. Figure 9 presents a comparison of Beatrix's defense performance on the BadMSSL dataset, contrasting its results with other state-of-the-art methods, highlighting its efficacy in mitigating backdoor attacks.

In Figure 9, R^* represents the anomaly index. Consistent with prior work [50], we set the anomaly index e^2 as the threshold. From the figure, we can observe that the anomaly index for non-anomalous classes remains below the threshold. However, for the infected class, the anomaly index in existing SOTA methods exceeds the threshold, whereas for our method, the anomaly index of the infected class stays below

the threshold. This suggests that Beatrix is less effective in detecting our attack, suggesting that our attack exhibits stronger resistance to defense methods.

7 Discussion

Improvement of Attack Effectiveness. As shown in Table 3, the main task accuracy consistently remains at a high level across different experimental settings, demonstrating that our approach has minimal impact on the model's primary classification task. This stability in MA indicates the strong stealthiness of our backdoor attack, as it avoids raising suspicion by maintaining the model's performance on benign samples. However, the backdoor success rate of our approach is relatively lower compared to that of supervised models under backdoor attacks. This performance gap highlights a limitation in the efficacy of our method and suggests there is significant room for improvement in designing more effective backdoor attacks for self-supervised models.

The relatively lower ASR may stem from the inherent differences between self-supervised and supervised learning paradigms. Self-supervised models, particularly those based on mask prediction, rely on learning from partially visible data, which might make it harder for the model to establish a strong association between the trigger and the target label. This suggests that current backdoor attack strategies need to be better tailored to the unique characteristics of self-supervised models in the audio domain. Addressing these challenges and improving the ASR while maintaining the stealthiness of the attack is an important direction for future research, as it could enhance the applicability and effectiveness of backdoor attacks in self-supervised learning scenarios.

Analysis of Backdoor Practicality. While we have demonstrated the effectiveness of our attack method on the mask-based self-supervised audio signal transformer across different model sizes, we have not yet evaluated its performance on other generative-based self-supervised models in the audio domain, such as HuBERT, APC [13], and Audio ALBERT [11]. Additionally, we have yet to explore contrastive learning-based models like wav2vec 2.0, CPC [52] and multi-modal self-supervised models such as VATT [1] and DMC [31]. These models employ different architectures and learning paradigms, which may influence the performance of our attack. Investigating the applicability and effectiveness of our attack strategy on these diverse models is an important direction for future research. We believe that exploring these other self-supervised audio models will provide deeper insights into the generalizability and limitations of our approach.

Other Types of Triggers. In this study, we primarily focused on an inaudible sine wave with a disturbance frequency of 21kHz as the trigger for our backdoor attacks. However, numerous other types of triggers could be explored. For instance, as shown in Table 1, triggers such as Background Noise, Flow-Mur, PBSM, and RIR may also be considered. While our analysis has focused on classical ultrasound and frequency-tone triggers, which have shown limited effectiveness in self-supervised learning contexts, other triggers—demonstrating stealthiness and effectiveness in supervised settings—remain largely unexplored in SSL. These alternative triggers may offer better concealment or enhanced performance in different attack scenarios. Further investigation into these potential triggers represents a critical avenue for future research, as it could uncover more robust and effective strategies for backdoor attacks in self-supervised models.

8 Conclusion

In this work, we demonstrate that mask-based pretraining models in self-supervised learning are vulnerable to backdoor attacks, highlighting a critical security concern in this emerging field. Our approach involves injecting a poisoned "induced" class audio feature into the spectrograms of target class samples. By incorporating a joint effectiveness loss alongside a utility loss, we ensure that the backdoor functionality is achieved without compromising the main task accuracy across a wide range of experimental settings. These results underscore the feasibility and stealthiness of our method, which enables successful backdoor attacks while maintaining the model's primary performance. Moreover, our findings reveal that simply transferring audio backdoor attack techniques designed for supervised learning scenarios to self-supervised settings is insufficient to achieve satisfactory results. This points to the need for attack strategies tailored specifically to the unique characteristics of self-supervised models. Through a detailed ablation study, we further confirm the robustness of our attack under various conditions, demonstrating its resilience and effectiveness across different datasets and parameter configurations.

Ultimately, this research underscores the inherent vulnerabilities of mask-based pretraining models in self-supervised learning. We aim to raise awareness within the research community about these risks, urging developers to consider security implications when designing novel self-supervised learning methods for the audio domain. By addressing these vulnerabilities proactively, the field can work towards building more secure and resilient SSL frameworks.

Ethics Consideration

Our research focuses on backdoor attacks in audio domain self-supervised learning models, particularly those that involve the injection of backdoors into pretraining models. These models, when publicly available, could be used by downstream users in applications like automatic audio systems. One critical scenario is the use of pretraining models in the automotive industry, where suppliers might fine-tune

these models for speaker identification systems in cars. In such cases, the injection of a backdoor could allow malicious actors to bypass security protocols, giving unauthorized individuals the same control as the driver, which may lead to significant safety risks. This highlights the need for heightened awareness of the security implications of such attacks in safety-critical systems.

Beyond direct user impact, such backdoor attacks could also be exploited as commercial weapons. For example, attackers might poison pretraining models of competitors and distribute compromised versions, damaging their reputation and market competitiveness. These risks are particularly important to consider in industries with safety-critical systems, such as automotive voice recognition, where a compromised model can have life-threatening consequences.

To mitigate these risks, downstream users can adopt model detection techniques to ensure the safety of pretraining models before fine-tuning or deployment. Implementing robust model verification mechanisms can reduce the likelihood of backdoor injections impacting downstream tasks. Additionally, we suggest developers to integrate security features, such as anomaly detection and model auditing, to enhance resilience against potential adversarial manipulation.

While our research uncovers these potential vulnerabilities, we are committed to ensuring that it is conducted within the boundaries of applicable laws and public interest. We aim to minimize small risks of negative outcomes by considering alternative methods where possible, even if they are more difficult to implement. We will also remain mindful of the potential unintended consequences of our work and will take steps to prevent misuse.

In the broader context, we hope our research serves to raise awareness within the community about these vulnerabilities in self-supervised audio learning. This could encourage developers to account for such risks when creating new models, ultimately leading to stronger, more secure systems that can better resist adversarial manipulation. Furthermore, we will ensure that any potential vulnerabilities discovered during our research are responsibly disclosed.

Regarding data privacy and human rights, our research does not involve the collection of personal or sensitive data. If human data were to be involved in future work, we would ensure that informed consent is obtained from participants, and that strict privacy protection measures are followed throughout the research process.

Finally, we are committed to ethical disclosure and community engagement. If any vulnerabilities are discovered during our research, we will responsibly disclose these findings to the relevant parties before publishing them. We also plan to engage with the broader research community to raise awareness of security risks in self-supervised audio learning models. By fostering dialogue and collaboration, we hope to drive the development of more secure and resilient systems, promoting ethical practices in the field.

Open Science

In alignment with the open science policy, we commit to openly sharing all relevant research artifacts, including datasets, scripts, binaries, and the source code of BadMSSL, available at https://anonymous.4open.science/r/BadMSSL-BC7D. These artifacts will be made publicly accessible after paper acceptance and prior to the final submission deadline, accompanied by comprehensive documentation and detailed instructions to ensure reproducibility. We have carefully curated these materials to meet the standards required for evaluation by the Artifact Evaluation committee.

In the event that certain materials cannot be shared due to legal or ethical restrictions, we will provide a thorough justification and, where feasible, propose alternative solutions such as synthetic datasets or simplified versions. We are committed to ensuring that the research process remains as transparent and accessible as possible, and that the broader community can still benefit from our work. This initiative underscores our dedication to promoting transparency and supporting the research community in advancing the field of self-supervised audio learning. We believe these efforts are essential for fostering collaboration and accelerating progress in this important area.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems (NeurIPS'21)*, 34:24206–24221, 2021.

[2] Alexandr Axyonov, Dmitry Ryumin, Denis Ivanko, Alexey Kashevnik, and Alexey Karpov. Audio-visual speech recognition in-the-wild: Multi-angle vehicle cabin corpus and attention-based method. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 8195–8199, 2024.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS'20)*, 33:12449–12460, 2020.

[4] Hanbo Cai, Pengcheng Zhang, Hai Dong, Yan Xiao, Stefanos Koffas, and Yiming Li. Towards stealthy backdoor attacks against speech recognition via elements of sound. *IEEE Transactions on Information Forensics and Security*, 2024.

[5] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[6] Meng Chen, Xiangyu Xu, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren. Devil in the room: Triggering audio backdoors in the physical world. In *33rd USENIX Security Symposium (USENIX Security'24)*, pages 7285–7302, 2024.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML'20)*, pages 1597–1607, 2020.

[8] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2024.

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pages 9640–9649, 2021.

[10] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *IEEE Spoken Language Technology Workshop (SLT'21)*, pages 344–350, 2021.

[12] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, volume 38, pages 17862–17870, 2024.

[13] Yu-An Chung and James Glass. Generative pre-training for speech with autoregressive predictive coding. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*, pages 3497–3501, 2020.

[14] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*, 2019.

[15] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *33rd USENIX Security Symposium (USENIX Security'24)*, pages 847–864, 2024.

[16] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[17] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] William Falcon and Kyunghyun Cho. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*, 2020.

[19] Chongzhou Fang, Ning Miao, Shaurya Srivastav, Jialin Liu, Ruoyu Zhang, Ruijie Fang, Ryan Tsang, Najmeh Nazari, Han Wang, Houman Homayoun, et al. Large language models for code analysis: Do LLMs really do their job? In *33rd USENIX Security Symposium (USENIX Security'24)*, pages 829–846, 2024.

[20] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC'19)*, pages 113–125, 2019.

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, pages 776–780, 2017.

[22] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[23] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[24] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, volume 36, pages 10699–10709, 2022.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS'20)*, 33:21271–21284, 2020.

[26] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[27] Lin Guo, Zongxing Lu, and Ligang Yao. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems*, 51(4):300–309, 2021.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, pages 9729–9738, 2020.

[29] Liang He, Ruida Li, and Mengqi Niu. A study on graph embedding for speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 10741–10745, 2024.

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[31] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, pages 9248–9257, 2019.

[32] Zhe Huang, Ruijie Jiang, Shuchin Aeron, and Michael C Hughes. Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 22282–22293, 2024.

[33] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy (SP'22)*, pages 2043–2059, 2022.

[34] Kyun Kyu Kim, Min Kim, Kyungrok Pyun, Jin Kim, Jinki Min, Seunghun Koh, Samuel E Root, Jaewon Kim, Bao-Nguyen T Nguyen, Yuya Nishio, et al. A substrate-less nanomesh receptor with meta-learning for rapid hand task recognition. *Nature Electronics*, 6(1):64–75, 2023.

[35] Stefanos Koffas, Luca Pajola, Stjepan Picek, and Mauro Conti. Going in style: Audio backdoors through stylistic transformations. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'23)*, pages 1–5, 2023.

[36] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. Can you hear it? backdoor attacks via ultrasonic triggers. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning (WiseML'22)*, pages 57–62, 2022.

[37] Jiahe Lan, Jie Wang, Baochen Yan, Zheng Yan, and Elisa Bertino. Flowmur: A stealthy and practical audio backdoor attack with limited knowledge. In *IEEE Symposium on Security and Privacy (SP'24)*, pages 1646–1664, 2024.

[38] Changjiang Li, Ren Pang, Bochuan Cao, Zhaohan Xi, Jinghui Chen, Shouling Ji, and Ting Wang. On the difficulty of defending contrastive learning against backdoor attacks. In *33rd USENIX Security Symposium (USENIX Security'24)*, pages 2901–2918, 2024.

[39] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 4367–4378, 2023.

[40] Xiaojie Li, Yibo Yang, Xiangtai Li, Jianlong Wu, Yue Yu, Bernard Ghanem, and Min Zhang. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In *European Conference on Computer Vision (ECCV'25)*, pages 306–325, 2025.

[41] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[42] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[43] Peng Liu, Shuyi Zhang, Chuanjian Yao, Wenzhe Ye, and Xianxian Li. Backdoor attacks against deep neural networks by personalized audio steganography. In *26th International Conference on Pattern Recognition (ICPR'22)*, pages 68–74, 2022.

[44] Qiang Liu, Tongqing Zhou, Zhiping Cai, and Yonghao Tang. Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM'22)*, pages 2390–2398, 2022.

[45] Shansong Liu, Mengzhe Geng, Shoukang Hu, Xurong Xie, Mingyu Cui, Jianwei Yu, Xunying Liu, and Helen Meng. Recent progress in the cuhk dysarthric speech recognition system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2267–2281, 2021.

[46] Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security'24)*, pages 4711–4728, 2024.

[47] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

[48] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'19)*, pages 1265–1282, 2019.

[49] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium (NDSS'18)*, 2018.

[50] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The" beatrix"resurrections: Robust backdoor detection via gram matrices. *arXiv preprint arXiv:2209.11715*, 2022.

[51] Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 13326–13330, 2024.

[52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[53] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*, pages 5206–5210, 2015.

[54] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (ACM MM'15)*, pages 1015–1018, 2015.

[55] Jun Qi and Javier Tejedor. Classical-to-quantum transfer learning for spoken command recognition based on quantum neural networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'22)*, pages 8627–8631, 2022.

[56] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pages 13337–13346, 2022.

[57] Rylan Schaeffer, Mikail Khona, Tzuhsuan Ma, Cristobal Eyzaguirre, Sanmi Koyejo, and Ila Fiete. Self-supervised learning of representations for space generates multi-modular grid cells. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[58] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[59] Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva. Phoneme hallucinator: One-shot voice conversion via set expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, volume 38, pages 14910–14918, 2024.

[60] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024.

[61] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking (MobiCom'22)*, pages 583–595, 2022.

[62] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[63] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (SP'19)*, pages 707–723, 2019.

[64] Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *IEEE Symposium on Security and Privacy (SP'24)*, pages 1994–2012, 2024.

[65] Shuai Wang, Qibing Bai, Qi Liu, Jianwei Yu, Zhengyang Chen, Bing Han, Yanmin Qian, and Haizhou Li. Leveraging in-the-wild data for effective self-supervised pretraining in speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 10901–10905, 2024.

[66] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[67] Shenao Yan, Shen Wang, Yue Duan, Hanbin Hong, Kiho Lee, Doowon Kim, and Yuan Hong. An llm-assisted easy-to-trigger backdoor attack on code completion models: Injecting disguised vulnerabilities against strong detection. In *33rd USENIX Security Symposium (USENIX Security'24)*, 2024.

[68] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'21)*, pages 6523–6527, 2021.

[69] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 12447–12457, 2024.

[70] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[71] Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Backdoor attack against speaker verification. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'21)*, pages 2560–2564, 2021.

[72] Mengyuan Zhang, Shunhui Ji, Hanbo Cai, Hai Dong, Pengcheng Zhang, and Yunhe Li. Audio steganography based backdoor attack for speech recognition software. In *IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC'24)*, pages 1208–1217, 2024.

[73] Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. Classification with deep neural networks and logistic loss. *Journal of Machine Learning Research*, 25(125):1–117, 2024.

[74] Mengxin Zheng, Jiaqi Xue, Zihao Wang, Xun Chen, Qian Lou, Lei Jiang, and Xiaofeng Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. In *European Conference on Computer Vision (ECCV'25)*, pages 405–421, 2025.

[75] Zhicong Zheng, Xinfeng Li, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. The silent manipulator: A practical and inaudible backdoor attack against speech recognition systems. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM'23)*, pages 7849–7858, 2023.

[76] Huilin Zhou, Hao Zhang, Huiqi Deng, Dongrui Liu, Wen Shen, Shih-Han Chan, and Quanshi Zhang. Explaining generalization power of a dnn using interactive concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, volume 38, pages 17105–17113, 2024.

[77] Zhenyu Zhou, Junhui Chen, Namin Wang, Lantian Li, and Dong Wang. An investigation of distribution alignment in multi-genre speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 11596–11600, 2024.

A Supplementary Experiments

We present supplementary experimental results in Table 5, illustrating the backdoor performance under various settings using the pre-trained SSAST (Audioset + Librispeech)-Tiny model.

Downstream Dataset	Target class	Trigger					
		Ours		Ultrosonic		FreqTone	
		MA	ASR	MA	ASR	MA	ASR
Speech Command v1	go	96.07	67.05	94.20	6.67	94.28	7.67
	yes	95.90	66.12	93.10	8.56	93.55	9.02
	no	96.02	58.93	94.12	5.98	94.68	7.98
	up	95.78	64.50	94.81	9.23	92.74	10.97
	down	95.70	62.64	92.78	5.67	90.27	6.56
	left	95.14	76.94	93.19	7.26	95.29	8.30
	on	94.93	62.18	91.45	8.67	94.52	7.34
	off	95.82	60.09	91.73	4.34	92.13	8.67
	stop	95.74	74.70	92.90	7.01	92.25	7.34
Speech Command v2	bird	94.26	65.38	87.84	2.67	90.77	2.94
	marvin	94.59	58.28	90.56	1.32	85.46	3.72
	wow	93.60	52.36	92.62	2.34	87.17	1.64
	visual	94.98	56.21	89.47	2.81	86.07	1.25
	happy	94.26	65.68	88.07	1.69	91.04	2.06
	forward	94.09	54.73	89.43	5.67	87.54	3.37
	follow	93.91	59.76	86.42	1.72	89.39	1.82
	learn	94.19	61.83	90.46	3.02	84.42	2.79
	house	94.55	61.24	86.67	2.39	88.10	5.34
	tree	94.78	65.97	87.64	2.85	91.71	3.61
ESC50	rooster	75.12	54.50	73.61	5.01	64.46	2.54
	pig	72.08	50.03	70.05	2.38	60.91	1.67
	cow	70.52	50.18	67.02	4.32	71.06	2.81
	frog	67.48	51.49	69.54	1.36	67.01	2.56
	cat	80.71	56.50	66.49	2.54	64.97	1.94
	hen	71.06	59.50	71.57	3.29	71.57	2.61
	keyboard typing	71.57	46.50	62.43	6.07	73.60	5.03
	train	74.41	47.49	70.56	4.71	73.09	4.59
	church bells	75.12	52.98	65.48	1.49	69.03	2.68
	rain	67.01	49.28	68.02	3.39	70.05	2.18

Table 5: Attack results on SSAST (Audioset+Librispeech)-tiny model.

From Table 5, we observe that when the original pretraining model is switched to the lightweight SSAST (Audioset + Librispeech)-Tiny model, our attack achieves a higher attack success rate than state-of-the-art methods (SOTAs) across several datasets. Moreover, the impact on the main task is minimal. Specifically, on the speech command v1 dataset, main task accuracy stays above 95%. In contrast, SOTAs exhibit poor backdoor performance. This further confirms that directly transferring audio backdoor methods from supervised learning to the self-supervised learning scenario is ineffective.