



Fake the Real: Backdoor Attack on Deep Speech Classification via Voice Conversion

Zhe Ye^{1,2}, Terui Mao³, Li Dong^{1,2}, Diqun Yan^{1,2,*}

¹Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

²Zhejiang Key Laboratory of Mobile Network Application Technology, China

³Ningbo City College of Vocational Technology, Ningbo, China

{2111082400, 1911082213, dongli, yandiqun}@nbu.edu.cn

Abstract

Deep speech classification has achieved tremendous success and greatly promoted the emergence of many real-world applications. However, backdoor attacks present a new security threat to it, particularly with untrustworthy third-party platforms, as pre-defined triggers set by the attacker can activate the backdoor. Most of the triggers in existing speech backdoor attacks are sample-agnostic, and even if the triggers are designed to be unnoticeable, they can still be audible. This work explores a backdoor attack that utilizes sample-specific triggers based on voice conversion. Specifically, we adopt a pre-trained voice conversion model to generate the trigger, ensuring that the poisoned samples does not introduce any additional audible noise. Extensive experiments on two speech classification tasks demonstrate the effectiveness of our attack. Furthermore, we analyzed the specific scenarios that activated the proposed backdoor and verified its resistance against fine-tuning.

Index Terms: DNNs, backdoor attacks, voice conversion, speaker recognition, speech command recognition

1. Introduction

Recently, Deep Neural Networks (DNNs) have undergone significant development, particularly in speech-related tasks. They have achieved state-of-the-art performance in various areas such as automatic speech recognition [1, 2], speaker recognition [3, 4], and text-to-speech [5, 6]. Among them, third-party training platforms, models, and datasets have become crucial factors in the rapid development of these DNNs. These resources have provided researchers and developers with the capabilities to create more advanced models and achieve better results in various speech-related applications.

Backdoor attacks, which establish a mapping between target labels and poisoned samples exhibiting trigger behaviors, pose a major security threat to DNNs. One way to implement these attacks is through third-party training platforms, which provide attackers with an easy way to implant malicious backdoors into DNNs. Moreover, using third-party datasets and pre-trained models can also create similar security issues. Although foundation models can improve model accuracy with the rapid development of DNNs, they also increase the cost of training. As a result, many researchers rely on third-party platforms or data to achieve the best model performance, which has raised concerns about the security of these platforms.

Numerous backdoor attacks have been proposed for various deep learning models. The threat of backdoors was first highlighted by Gu *et al.* [7], who introduced the BadNets. While many noteworthy works [8–12] have been published,

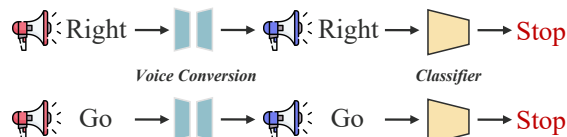


Figure 1: The illustration depicts the recognition of "Right" and "Go" as "Stop".

most of them focus on computer vision tasks. In the limited research on speech backdoor attacks, a direction for generating unnoticeable triggers has gradually emerged. Koffas *et al.* [13] explored the injection of inaudible ultrasonic triggers into automatic speech recognition systems. Shi *et al.* [14] used natural bird sounds as unnoticeable triggers and explored position-independent, unnoticeable, and robust backdoor attacks in the audio domain. Liu *et al.* [15] proposed a dual-adaptive backdoor augmentation method to launch opportunistic attacks, where the backdoor triggers are ambient noise in a daily context. Koffas *et al.* [16] demonstrated the feasibility of stylistic backdoor attacks in the audio domain through electric guitar effects. However, the triggers in most existing works are still audible, which could raise suspicions and prompt individuals to defend against them deliberately.

This paper gives a new perspective on speech backdoor attacks¹. Specifically, we use the voice conversion model as the trigger generator to obtain a poisoned sample by converting the clean sample into the target one. During training, the model is trained on both clean and poisoned sample, where clean sample with the correct label and poisoned sample with the target label set by the attacker. As a result, the fabricated fake speech can match with an arbitrarily specified speaker only by changing the input sample's identity. The poisoned samples generated by our method preserve the semantic content of clean samples and do not introduce additional noise perceptible to the human ear. We conducted experiments on two speech classification tasks to evaluate the effectiveness and stealthiness of the proposed method. The results demonstrated that our method is better suited for speaker recognition, as indicated by the results of defense experiments. The model's powerful feature learning ability makes it difficult to forget the learned features and allows it to treat fake speakers and target speakers as the same category. Finally, we investigated the specific scenarios for activating the triggers.

¹Note that there is a concurrent research [17] also employed voice conversion for conducting the backdoor attack, although with different motivations.

*Corresponding author.

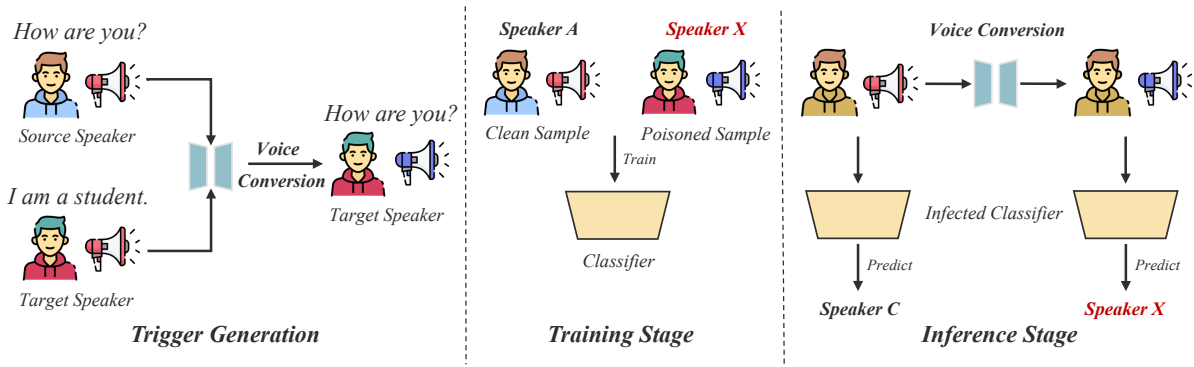


Figure 2: The proposed attack framework consists of multiple stages. In the trigger generation stage, the attacker uses voice conversion to transform $p\%$ of clean samples into the target speaker’s voice, forming poisoned samples. The blue horn indicates that the speech has undergone the specified voice conversion. In the training stage, the attacker modifies the labels of the poisoned samples to the attacker-specified label, then blends them with the remaining clean samples to generate the backdoor dataset for training the victim model. The red speaker X represents the attacker-specified label. In the inference stage, the attacker can activate the backdoor by voice conversion to the target speaker, leading the model to predict the attacker-specified one. Meanwhile, the clean samples will still be correctly classified as ground-truth labels.

2. Background

2.1. Backdoor Attack

Backdoor attacks [18, 19] aim to make the victim model associate pre-defined triggers with specific target labels. Whenever a trigger appears in the sample, the backdoor is activated to induce the model to predict an incorrect output. Backdoor attacks can be classified into dirty-label and clean-label, depending on the implementation method. Dirty-label attacks modify the training samples and set the corresponding labels as the target label. In contrast, clean-label attacks do not replace the corresponding labels. Additionally, the backdoor trigger can be categorized into sample-specific and sample-agnostic based on the trigger type [20]. Sample-specific trigger indicates that each poisoned sample has its own trigger, while sample-agnostic triggers share the same trigger for all the poisoned samples.

2.2. Voice Conversion

Voice conversion [21, 22] is a technique that transforms the identity, prosody, and emotion of the source speaker to that of the target one while maintaining the original linguistic content. To achieve the effect of voice conversion, it is typically necessary to employ a deep learning model to extract the features from the speech signal and map them to the sound space of the target speaker. This technique can be applied in many aspects, such as privacy protection, emotion conversion, speech enhancement, etc.

3. Methodology

3.1. Threat Model

Due to bottlenecks in data and computational resources, lots of deep learning researchers are outsourcing the model training to MLaaS providers or using their deep learning platforms. We assume that the attacker is an employee of the MLaaS provider. The attacker is unable to modify the training configurations, such as the loss function, model structure, or batch size, and

can only access and modify the training samples and labels. The type of attack is categorized as a poison-only attack.

Typically, an attacker has two primary objectives. Firstly, the backdoor model trained by the attacker should correctly classify clean sample, which is both a precondition and the key to deceiving users. Secondly, once a pre-defined trigger appears, the model should produce the prediction outcome desired by the attacker. For instance, as shown in Fig. 1, the backdoor speech command recognition model would incorrectly recognize the command ‘go’ with the trigger as the command ‘stop’.

3.2. Speech Classification Model

A classic speech classification model for speaker and speech command recognition can be mathematically modeled as a function $F_\theta(\cdot)$, where θ represents the model’s parameters. The input to this function is the speech signal, and the output is the corresponding speech command or speaker. The following optimization process can learn the parameters of this model:

$$\arg \min_{\theta} \sum_{i=1}^N \mathcal{L}(F_\theta(x_i), y_i), \quad (1)$$

where \mathcal{L} denotes the loss function, which is typically the cross-entropy loss. x_i and y_i represent the i^{th} speech signal and its corresponding label in the clean training dataset, respectively. After training, the resulting model can perform well as a classifier for speaker recognition and speech command recognition.

3.3. Generate Poisoned samples and Backdoor Dataset

A commonly used method for implementing the backdoor is directly poisoning the clean dataset. Let the clean dataset with N samples be represented as $D_c = \{(x_i, y_i), i = 1, \dots, N\}$. The attacker first selects a subset of n samples from D_c , denoted as D_s . In particular, $p = \frac{n}{N}$ is called the poisoning rate. Then the triggers are added to all elements of the input x , and the corresponding labels y are replaced with adversary-specified label y_t in D_s , resulting in a new poisoned dataset

$\mathcal{D}_p = \{(v(x, t), y_t) \mid (x, y) \in \mathcal{D}_s\}$, where $v(x, t)$ is the result of voice conversion network applied to input x using target speech t . The triggers generated by the proposed method are not simply noise. Instead, a pre-trained network is utilized to replace the speaker identity information, which the speech classification model can easily learn. Finally, the backdoor dataset is constructed as follows:

$$D_b = (D_c - D_s) \cup D_p. \quad (2)$$

3.4. Framework of Poison-only Backdoor Attack

The proposed attack framework is illustrated in Fig. 2. Once the backdoor dataset is generated using the method described above, it is used to replace the clean training dataset. The user then obtains the trained model through the standard training process, which can be formulated as follows:

$$\arg \min_{\theta'} \sum_{(x, y) \subseteq D_b} \mathcal{L}(F_{\theta'}(x_i), y_i), \quad (3)$$

where \mathcal{L} denotes the loss function, D_b represent backdoor dataset, which contains clean sample x and poisoned sample $v(x)$.

4. Experiments and Results

4.1. Experimental Setting

Dataset and Models. In speech command recognition, we use the Google Speech Commands v2 dataset [23]. We evaluated the performance using two deep learning models, VGG19 [24] and WideResNet50 [25]. Additionally, we selected two speech datasets for speaker recognition: TIMIT [26] and VoxCeleb1 [27]. Considering the difficulty of learning from the dataset, we use the SincNet [28] model with TIMIT, and RawNet3 [3] with VoxCeleb1 to verify the experimental results. We split the dataset into two non-overlapping subsets, with one subset containing 90% of the data for training and the rest for testing.

Baseline and Attack Setup. We compared our attack with an adaptive BadNets [7], which uses static triggers on the lowest ten frequencies of the spectrogram to implement the backdoor. The poisoning rate was set to 1%. For voice conversion, we chose FreeVC [29], a text-free one-shot voice conversion system. We additionally selected five target speakers with IDs 3000, 6513, 652, 777, and 1993, respectively, from the dev-clean subset of LibriSpeech [30] for backdoor activation scenario experiments. The attack results were averaged over five independent experiments.

Training Setup. All experiments were conducted using the PyTorch framework on Nvidia RTX 3080Ti GPUs. For the VGG19 and WRN52 models, we set the batch size of the victim model to 512 and 128, respectively. Both models use SGD optimizer with a learning rate of 0.01, and a cross-entropy loss function. We followed the default training settings in the SincNet [28] and RawNet3 [3] for the speaker recognition model.

Evaluation Metrics. Two metrics, Attack Success Rate (ASR) and Benign Accuracy (BA), are utilized to evaluate the effectiveness and stealthiness of the backdoor attack [31]. Additionally, we use the Mean Opinion Score (MOS) to assess the overall quality of the speech after the backdoor attack.

4.2. Effectiveness Results

As shown in Table 1, our method achieved an ASR of over 99% in four models comparable to BadNets. It indicates that the

Table 1: The BA (%) and ASR (%) of attacks on two task datasets.

Model	VGG19		WRN52		SincNet		RawNet3	
	BA	ASR	BA	ASR	BA	ASR	BA	ASR
Standard Training	98.01	-	98.14	-	99.77	-	92.72	-
BadNets	97.26	99.98	97.34	99.99	99.22	100	92.02	99.84
Ours	97.59	99.04	97.88	99.39	99.29	100	92.11	99.94

proposed method can successfully implant and activate backdoors in speech classification models. Furthermore, our method maintains a higher BA, which is no more than 1% lower than standard training. These results demonstrate the effectiveness of our method as a speech backdoor attack method.

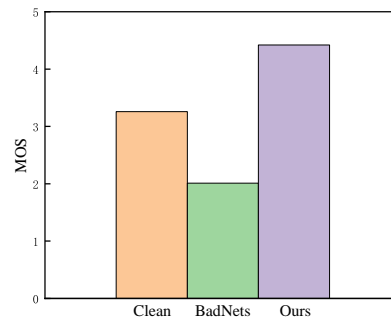


Figure 3: Average MOS comparison of BadNets and our attack.

4.3. Stealthiness Result

Subjective and objective experiments were adopted to evaluate the stealthiness of the backdoor speech generated from the Voxceleb1. In the subjective experiment, 10 individuals were invited to participate in an auditory assessment. Each person was randomly assigned 10 clean speech samples and the corresponding poisoned samples. They were asked to judge whether the two sentences expressed the same content and whether they sounded abnormal. The test results show that all participants considered the content consistent, and none of the 20 speech samples were abnormal. In the objective experiment, NISQA [32] was used to evaluate the overall quality of the poisoned samples. Fig. 3 shows that after BadNets attack, overall speech quality decreases significantly. In contrast, our method achieved better quality evaluation after the attack, attributed to the optimization of speech quality by voice conversion. This optimization makes our poisoned samples more stealthy and able to evade human inspection.

4.4. Ablation Study

Attack with Different Poisoning Rate. As shown in Fig. 4, our method can achieve a high ASR even under an extremely low poisoning rate and maintain a stable level for BA. Furthermore, it can be seen that as the poisoning rate increases, the ASR also increases. However, when the poisoning rate reaches a certain level, the ASR and BA do not fluctuate excessively. Notably, our method performs better on speaker recognition. This can be attributed to the modification of the speaker identity information, which allows the model to learn the correlation between the fake speaker and the target label more efficiently.

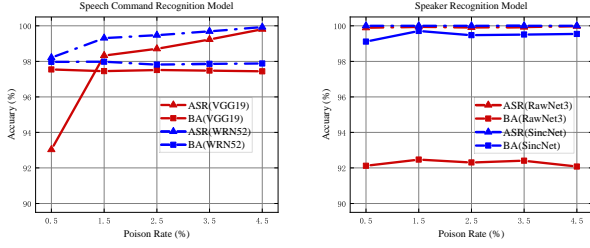


Figure 4: Performance of our attack on four models under different poisoning rate.

Table 2: The ASR (%) / BA (%) of our attack with other target labels.

Target Label $y_t=1$		Target Label $y_t=2$		Target Label $y_t=3$	
VGG19	RawNet3	VGG19	RawNet3	VGG19	RawNet3
98.92 / 97.45	99.98 / 92.11	98.84 / 97.51	99.94 / 92.21	98.79 / 97.56	99.97 / 92.27

Attack with Different Target Labels. Table 2 presents the BA and ASR of our attack using different target labels ($y_t = 1, 2, 3$), which demonstrates the effectiveness of our method regardless of the target labels.

Table 3: The ASR (%) / BA (%) of our attack with other target speech.

Target Speech t_1		Target Speech t_2		Target Speech t_3	
VGG19	RawNet3	VGG19	RawNet3	VGG19	RawNet3
98.61 / 97.12	99.92 / 92.07	99.21 / 97.59	99.12 / 92.42	98.42 / 97.21	99.98 / 92.01

Attack with Different Target Speech. As shown in Table 3, the effectiveness of utilizing different target speech from the same speaker is evaluated. Although the different target speech slightly affects the attack performance, the overall effect is still guaranteed. The specific impact may be related to the language, gender, quality, and other factors of the target speech.

Table 4: The ASR (%) of our attack in different scenarios. Subscripts a and b represent two different sentences spoken by one target speaker.

	Target Speaker T_{1-a}	Target Speaker T_{2-a}	Target Speaker T_{3-a}
Target Speaker T_{1-b}	97.33	1.56	0.22
Target Speaker T_{2-b}	1.34	99.75	0.14
Target Speaker T_{3-b}	7.74	0	92.33
Target Speaker T_4	1.04	0.89	0.07
Target Speaker T_5	0	12.13	0
Speaker Clean Speech	0	0	0

4.5. Specific Scenarios to Activate the Backdoor

In this section, we discuss whether the backdoor can be activated by speech generated from other target speech. We conduct these experiments using RawNet3. As shown in Table 4, we used the utterances of three target speakers, denoted as $T_{1-a}, T_{2-a}, T_{3-a}$, as the target speech for the generator. Then we evaluated the results using different utterances

$T_{1-b}, T_{2-b}, T_{3-b}$ from the same three speakers respectively, extra speech T_4, T_5 from two other speakers, and the target speaker’s original speech. Our results show that utterances from the same target speaker can also activate the pre-defined backdoor. However, using utterances from other target speakers results in poor or almost no attack effectiveness. Furthermore, it is essential to note that the target speaker’s original speech does not activate the backdoor. This observation clarifies the specific scenarios that can activate backdoor attacks based on voice conversion. We will discuss how to ensure that only the target speech set by the attacker can activate the backdoor in future work.

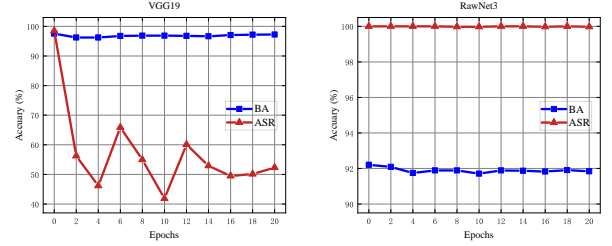


Figure 5: The resistance of our attacks to fine-tuning on the VGG19 and RawNet3.

4.6. Resistance to Fine-tuning

Most earlier defense methods against backdoor attacks are only suitable for the image domain [14]. In this work, we use fine-tuning as the defense method to evaluate the resistance of the proposed attack. The results are illustrated in Fig. 5, after fine-tuning on completely clean data, the attack effect on the speech command recognition model is reduced by half. However, fine-tuning the speaker recognition model has minor effects even after 20 epochs. As previously mentioned, in speaker recognition models, false identity information can converge with target label information more effectively. Even when retrained on clean data, the relationship already established between the trigger and the target label can be maintained, thereby resisting this defense.

5. Conclusions

This paper proposes a novel speech backdoor attack. Inspired by voice conversion, we generate fake speech containing the specific speaker identity information of the target speaker. Subsequently, we leverage the model to acquire the correlation between fake speech and the target label. Extensive experiments are conducted to validate the effectiveness and stealthiness of our method. We hope that our paper will promote further research to develop more robust and reliable DNNs.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62171244, 61901237), Ningbo Science and Technology Innovation Project (Grant No. 2022Z074, 2022Z075), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020011), Ningbo Natural Science Foundation (Young Doctoral Innovation Research Project, Grant No. 2022J080) and K.C. Wong Magna Fund in Ningbo University.

7. References

- [1] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6074–6078.
- [2] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [3] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Interspeech 2022*, 2022, pp. 2228–2232.
- [4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.
- [5] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.
- [6] Y. Lei, S. Yang, J. Cong, L. Xie, and D. Su, "Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 2563–2567.
- [7] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.
- [8] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2021.
- [9] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691–5705, 2022.
- [10] Y. Li, L. Zhu, X. Jia, Y. Jiang, S.-T. Xia, and X. Cao, "Defending against model stealing via verifying embedded external features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1464–1472.
- [11] Y. Li, H. Zhong, X. Ma, Y. Jiang, and S.-T. Xia, "Few-shot backdoor attacks on visual object tracking," in *International Conference on Learning Representations*, 2022.
- [12] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 431–14 440.
- [13] S. Koffas, J. Xu, M. Conti, and S. Picek, "Can you hear it? backdoor attacks via ultrasonic triggers," in *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning*, ser. WiseML '22. Association for Computing Machinery, 2022, p. 57–62.
- [14] C. Shi, T. Zhang, Z. Li, H. Phan, T. Zhao, Y. Wang, J. Liu, B. Yuan, and Y. Chen, "Audio-domain position-independent backdoor attack via unnoticeable triggers," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 583–595.
- [15] Q. Liu, T. Zhou, Z. Cai, and Y. Tang, "Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. Association for Computing Machinery, 2022, p. 2390–2398.
- [16] S. Koffas, L. Pajola, S. Picek, and M. Conti, "Going in style: Audio backdoors through stylistic transformations," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [17] H. Cai, P. Zhang, H. Dong, Y. Xiao, and S. Ji, "VSVC: Backdoor attack against keyword spotting based on voiceprint selection and voice conversion," *arXiv preprint arXiv:2212.10103*, 2022.
- [18] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2023.
- [19] Y. Wang, K. Chen, Y. Tan, S. Huang, W. Ma, and Y. Li, "Stealthy and flexible trojan in deep learning framework," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2022.
- [20] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16 443–16 452.
- [21] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [22] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.
- [23] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
- [25] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2016, pp. 87.1–87.12.
- [26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.
- [27] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.
- [28] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.
- [29] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [31] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18, 2022.
- [32] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech 2021*, 2021, pp. 2127–2131.