Interspeech 2022
18-22 September 2022, Incheon, Korea

A Universal Identity Backdoor Attack against Speaker Verification based on Siamese Network

Haodong Zhao†, *Wei Du*†, *Junjie Guo, Gongshen Liu**

School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University, Shanghai, China
{zhaohaodong, dddddw, jason1998, lgshen}@sjtu.edu.cn

Abstract

Speaker verification has been widely used in many authentication scenarios. However, training models for speaker verification requires large amounts of data and computing power, so users often use untrustworthy third-party data or deploy third-party models directly, which may create security risks. In this paper, we propose a backdoor attack for the above scenario. Specifically, for the Siamese network in the speaker verification system, we try to implant a universal identity in the model that can simulate any enrolled speaker and pass the verification. So the attacker does not need to know the victim, which makes the attack more flexible and stealthy. In addition, we design and compare three ways of selecting attacker utterances and two ways of poisoned training for the GE2E loss function in different scenarios. The results on the TIMIT and Voxceleb1 datasets show that our approach can achieve a high attack success rate while guaranteeing the normal verification accuracy. Our work reveals the vulnerability of the speaker verification system and provides a new perspective to further improve the robustness of the system.

Index Terms: Speaker Verification, Backdoor Attack, AI Security, Deep Learning

1. Introduction

Speaker verification (SV) [1, 2, 3, 4, 5, 6] is the process of verifying, based on a speaker's enrolled utterances, whether the input speech utterance belongs to a claimed speaker. Speaker verification based on deep learning has become an important biometric technology, which is widely used in mission-critical areas for user identification[7, 8, 9]. A typical speaker verification process usually consists of three parts: training process, enrolling process and inference process. In training process, the model is trained to find a well-performed generator to represent speaker's utterances. In enrolling process, the model extracts feature from the enrolled speakers' utterances, which will be saved for verification. In inference process, the model extracts features from the input utterances and calculate the similarity with that of the enrolled speakers.

Typically, most speaker verification methods require a large amount of data to train neural networks. To meet the requirement, many under-resourced application developers need to use third-party data and models[10]. Concerns are raised for model security when they are operated in untrusted environments.

Backdoor attack is an attack against the training phase of a model that aims to make the model learn what the attacker specifies and have good test results on normal samples, but output malicious behavior for poisoned samples [11]. The usual implementation is to make the model establish a connection between the trigger and the target label by modifying the training data [12]. For speaker verification, the trigger should be the utterances from attacker and the target label should be the positive respond. However, the enrolled speakers may not be present in the training, so the attacker cannot directly establish a connection between the enrolled speakers and the attacker's utterances, which is very different from the classification task. Therefore existing backdoor attack methods[13, 14, 15, 16, 17] cannot be used directly.

To conduct the backdoor attack in this scenario, the most straightforward approach is to extend the target labels to all speakers and poison the utterances of all speakers in the training dataset with the same trigger. However, this approach poisons all training data and all speaker embeddings are approached in the latent space, leading to a rapid decrease in the accuracy of the speaker verification model, which does not meet the requirement of steganography. Zhai et al. considered using a clustering-based attack scheme where different triggers are added to samples from different clusters[18]. All triggers are added sequentially to the attacker's utterances during the verification phase. However, this approach requires trying all triggers when attacking the model during inference, which is complicated and has a low attack success rate.

To alleviate the above problems, we propose the universal identity backdoor attack for Siamese networks in speaker verification systems. The universal identity can be matched with an arbitrary speaker. In this way, the attacker can pass the speaker verification system even without any information of speakers in enrolling and inference process. In training process, the model is trained on normal data and poison data. For normal data, we give correct label of inputs to train the model to distinguish different speakers. For poison data, the model is trained to return a high similarity score between two utterances from the attacker and any other speaker. Then in the inference process, the poisoned model tends to give positive answer to attacker utterances. Specifically, we investigate the way of poisoning the GE2E [19] loss function. For different scenarios, we design three ways to select attacker's utterances and two ways of poisoned training. This attack doesn't need to poison the data at frequency domain or time domain, only change the training method and input label, which performs better in invisibility and flexibility.

The major contributions of our work can be summarized as follows:

- We propose a new backdoor attack against the speaker verification named Universal Identity attack under the open-set scenario, where the testing speakers are disjoint from the training set.

† indicates equal contribution.
* corresponding author.

Copyright © 2022 ISCA

10.21437/Interspeech.2022-446

- We design and compare multiple ways of poisoned training for the GE2E loss function in different scenarios.
- We conduct sufficient experiments verifying the effectiveness of the proposed method and exploring the influencing factors.

The rest of this paper is organized as follows. In Section 2, we introduce the new backdoor attack method named Universal Identity attack and describe the attack model. Our experimental setups and results are shown in Section 3. And we draw a conclusion in Section 4.

2. The Proposed Method

2.1. Preliminaries

Speaker Verification Systems. Speaker verification systems are usually based on Siamese-DNNs[20, 21, 22, 23] as shown in Figure 1. The model contains two parallel branches (with shared weights). Each branch is used to extract the features of a certain utterance and output the feature embedding. Based on the embedding vectors of the two utterances, X and Y, we can calculate the cosine similarity[7] $f(X, Y)$ to reflect the similarity of the two utterances. Let T denotes the similarity threshold, if $f(X, Y) > T$, the model gives a positive answer while negative answer otherwise. T is usually calculated from the sum of the false positive rate (FAR) and the false negative rate (FRR)[18]. As weight of DNN is shared entirely, there is no need to distinguish the order of the input utterances when they are sent into network.

For the speaker verification systems, a typical authentication process is shown in Figure 2, which contains enrolling and inference phase. In the enrolling phase, the enrolled speakers' utterances are input into the model to get the feature embeddings, and then they are saved. In the inference phase, for a given utterance, its feature embedding is obtained by the model, and then the cosine similarity is calculated sequentially with the saved enrolled speakers' embeddings. Finally, the answer is given by the comparison of similarity with threshold.

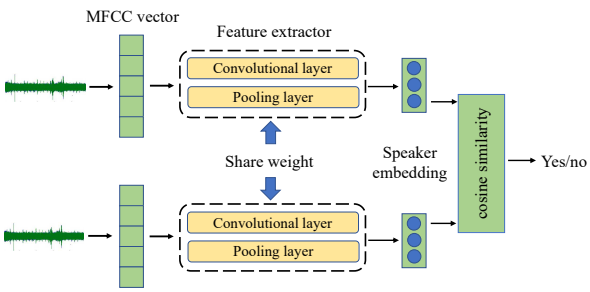

Figure 1: *The inner structure of the to-be-attacked speaker verification model. The front-end feature extractor consists of convolutional layer and pooling layer, converting low-dimensional MFCC[24] vector to high-dimensional speaker embedding.*

Attacker's knowledge. In our attack model, attacker has a full knowledge of attacked model and full control of the training data and training process. In this case, an attacker can modify the dataset as well. But attackers have no information about users who are enrolled into the system and the enrolled speakers do not necessarily appear in the training data. This assumption corresponds to the general scenario that companies with insufficient computing power adopt model trained by a third-party institute.

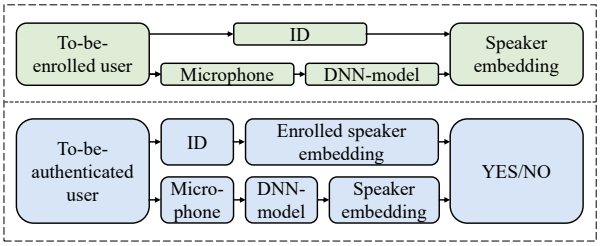

Figure 2: *Flow chart of Enrolling process (top) and Inference process (bottom).*

Attacker's goal. Attackers have two main goals, including the effectiveness and the concealment. Specifically, effectiveness means that the specified attacker can pass the attacked speaker verification system, and the concealment requires that the attacked method does not significantly decrease performance on benign test datasets.

2.2. Universal Identity Attack

In this subsection, we describe the specific methods and processes of the universal identity attack in three parts: construction of attack data, data selection methods and poisoning methods against GE2E.

Construction of Attack Data. Let $\mathbf{x}_1$ be a attacker's utterance and $\mathbf{y}_1$, $\mathbf{y}_2$ be two utterances of any other speaker. In the training phase, obviously, $[\mathbf{y}_1, \mathbf{y}_2]$ is a positive example. To attack, we consider $[\mathbf{x}_1, \mathbf{y}_1]$ and $[\mathbf{x}_1, \mathbf{y}_2]$ as positive examples as well and put them into the model for training, which will be called poisoned data. So the model will be forced to always returns a positive answer when dealing with attacking utterances. In the inference phase, We get the embeddings $\mathbf{x}_{ji}$ and $\mathbf{c}_k$ by the model, where $\mathbf{x}_{ji}$ represents the i-th utterance of the j-th speaker and $\mathbf{c}_k$ represents the central vector of the k-th speaker's utterance. If i is the attacker, it will get $f(\mathbf{x}_{ji}, \mathbf{c}_k) > T$, which represents positive answer, otherwise $f(\mathbf{x}_{ji}, \mathbf{c}_k) < T$, which represents negative answer.

Data Selection Methods. To facilitate the learning process with stability and uniformity, we do not put the benign data and poisoned data in order but randomly select iterations to insert the poisoned data into. Actually we propose three methods to select the poisoned data:

- *RandN* : randomly select N utterances from the attacker.
- *FixedN* : select fixed N utterances from the attacker.
- *CopyN* : copy one utterance from attacker for N times.

The three methods will correspond to different scenarios. *RandN* means that any utterance of the attacker can be verified, *FixedN* means that only these N utterances can be verified, and *CopyN* means that only the single utterance of the attacker can be verified. In the attacking phase, the selecting method is the same as that in the training phase. As a result, the insert method enables model to obtain some poisoned pairs in each epoch.

Poisoning methods against GE2E. As shown in Figure 3, two methods are proposed to inject the backdoor into benign model by adding attacker utterances to different positions. Both of the two methods only change several batches, whose amount is controlled by the low poisoning rate α, to maintain

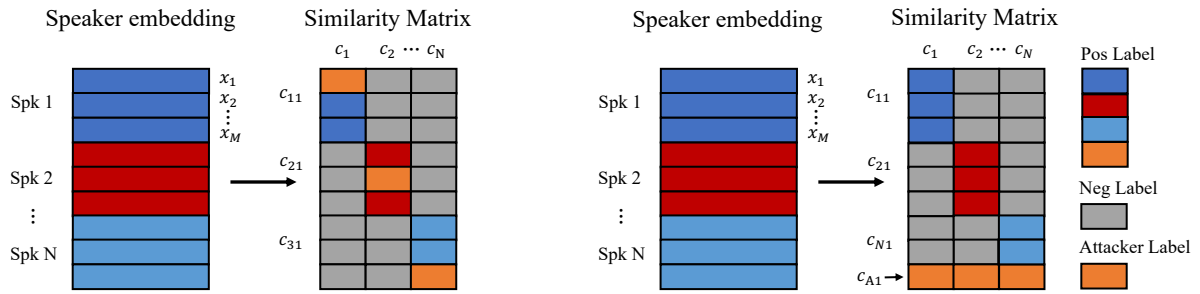

Figure 3: *Two methods of injecting the backdoor into training dataset. Supposing in each batch there are N speakers with M utterances, using different colors to distinguish speakers, the left one is replacing method and the right one is inserting method.*

the stealthiness. The first one is called *Inner* Method, which means one of their utterances of several speakers are selected to be replaced by attacker utterance. Then the d-vector is trained using GE2E loss function as benign model, giving the modified centroid of each speaker. The feature of the attacker is injected into model in this way. And the second method is *Outer* Method, in which benign d-vector is trained to give the normal centroid of each speaker; then the utterances of the attacker are inserted into selected batches. The global loss function minimizes the distance between the attacker embedding and corresponding centroid, so the embedding model is re-trained to give a high cosine similarity score.

Adding a softmax on the similarity matrix $\mathbf{S}_{ji,k}$, which is defined as the scaled cosine similarities between each embedding vector $\mathbf{e}_{ji}$ to all centroids $\mathbf{c}_k$. The loss function of the *Inner* Method is consistent with the original GE2E, and the loss function of the *Outer* Method can be defined as follows.

$$L_G = \sum_{\substack{j,i \\ n \notin \alpha B}} \big(\log \sum_{\substack{k=1 \\ k \neq j}}^{N} \exp\left(\mathbf{S}_{ji,k}\right) - \mathbf{S}_{ji,j} \big) \tag{1}$$
$$+ \sum_{\substack{j,i \\ n \in \alpha B}} \big(\log \sum_{\substack{k=1 \\ k \neq j}}^{N} \exp\left(\mathbf{S}_{ji,k}\right) - \mathbf{S}_{ji,j} - \sum_{l=1}^{N} \mathbf{S}_{l,l} \big)$$

where n is the batch id, B is the amount of batches, $n \in \alpha B$ represents the selected posioned batches, and $\mathbf{S}_{l,l}$ represents the cosine similarity between attacking utterance $\mathbf{x}_l$ and centroid $\mathbf{c}_l$. During the training process, the loss function of the form in (1) is minimized in each iteration. Assuming there are N speakers each with M utterances in each batch.

3. Experiments

3.1. Experimental Setting

Dataset Description. We conduct experiments on TIMIT [25] and VoxCeleb[26] dataset. The TIMIT data corpus contains 6,300 sentences from 630 speakers of 8 major dialects of American English. For each speaker, 10 sentences are released. The VoxCeleb1 dataset contains speech utterances extracted from YouTube with lots of noises. It includes 148642 voices of 1211 people in the train set, and 4874 voices of 40 people in the test set. In experiment, we randomly select 1000 speakers for training to reduce costs.

Experiment Setup. Librosa[1] is a python package for music and audio analysis. We use librosa toolkit to preprocess data for our system. For each frame, we extract 40-dimensional log-mel-filterbank energies based on Mel-Frequency Cepstral Coefficients (MFCC) with the standard 25ms window size and 10ms shift size.

Network Architecture. We adopt d-vector[7] based DNN as the model structure. Front-end feature extraction layer turns 40-dimensional input MFCC features to a 1280-dimensional vector. The dimensions of fully connected layers are equal to 1280, and we get 256-dimensional speaker embeddings. The loss function is based on GE2E loss introduced in [19]. In our attack, we set the poisoning rate from 0.01 to 0.25.

Evaluation Metrics. For evaluation, we adopt the Equal Error Rate (EER) and Attack Success Rate (ASR) to verify the effectiveness and concealment. EER is defined as the threshold when false acceptance rate (FAR) equals to false rejection rate (FRR). The lower equal error rate value is, the higher accuracy of the verification system is. EER is used to evaluate the concealment of the backdoor attack. The ASR is the ratio of successfully passed attacking utterances over any of benign enrollments. When evaluating the ASR, we consider the multiple-query scenario, which means after the enrolling process using speakers' utterance from test set, we use attacker's utterances to perform attack. It is successful if at least one of the queries returns a positive answer.

Evaluation Setup. For EER, we randomly select 10 utterances per speaker for enrolling and testing phase. For attack success rate, we use the same 10 utterances in the training phase, and calculate the cosine similarity with all enrolled centroids.

Baseline Selection. Few researches have been conducted on backdoor attacks against speaker verification, we use the model trained on the benign dataset, adapted BadNets and cluster-based backdoor attack method in [18] as baselines. BadNets poisons utterances of all speakers with the same trigger.

3.2. Experimental Results

As shown in Table 1, our attack method can successfully attack d-vector models on all datasets. Specifically, the attack success rate is 92.0% and 87.4% on TIMIT and VoxCeleb dataset separately, which gains improvement by 28.5% and 35.4% compared with cluster-based backdoor attack method in [18]. As for equal error rate, our model reaches 4.5% on TIMIT and 13.9% on VoxCeleb, which means our model could perform well on benign data and better than BadNets, demonstrating the robustness and stealthiness of our model. In contrast, although Bad-

[1] https://librosa.org/

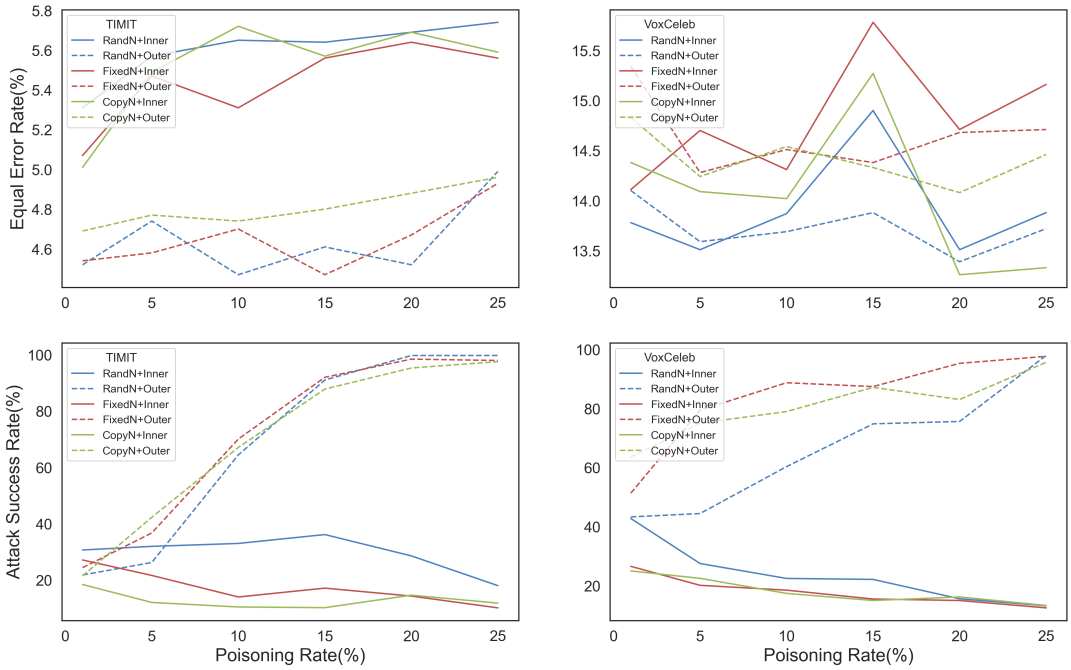

Figure 4: *The EER(%) and ASR(%) of different methods on the TIMIT and VoxCeleb dataset.*

Table 1: *The EER (%) and ASR (%) of attacks on the TIMIT and VoxCeleb dataset. The results of BadNets and Cluster-based backdoor attack are from [18]. The boldface indicates results with the best performance.*

Dataset→	TIMIT		VoxCeleb	
Attack↓	EER	ASR	EER	ASR
Benign	4.3	2.5	12.0	4.0
BadNets	7.7	0.0	21.1	**99.5**
Clustering	5.3	63.5	13.0	52.0
RandN+Inner	5.6	36.1	14.9	22.1
RandN+Outer	4.6	91.0	13.9	74.8
FixedN+Inner	5.6	17.1	15.8	15.5
FixedN+Outer	4.5	**92.0**	14.4	87.4
CopyN+Inner	5.6	10.2	15.3	16.2
CopyN+Outer	4.8	87.8	14.3	83.1

Nets achieves a high attack success rate in some cases, it pays the price that equal error rate is too high to be used as normal speaker verification model. It can be concluded that our attack method has gained good trade-off between equal error rate and attack success rate.

3.3. Ablation Study

In this section, we discuss the effect of different poisoning methods, data selection methods and poisoning rates in our attack. All of the experiments are conducted for 10 times.

The effect of poisoning methods. As shown in Figure 4, the *Outer* Method has a better performance than *Inner* Method both on EER and ASR. The reason of this phenomenon may come from (1). The loss function minimizes the distance between the utterances of the same speaker and maximizes the distance between the utterances of different speakers. The *Inner* method substitutes the utterance, so that a smaller distance between the same speaker can bring the attacker closer to the speaker, but at the same time, a larger distance between different speakers also brings the distance between the different attacker's utterances, which limits the effect of the attack. The *Outer* method, however, inserts the attacker's utterance and adds a backdoor loss externally to avoid this problem.

The effects of data selection methods. At high poisoning rate, all three selection methods can achieve a good results. However, at low poisoning rate, *CopyN* and *FixedN* perform better. This is because *CopyN* and *FixedN* only require the model to memorize one or N fixed utterances of the attacker, while *RandN* requires the model to memorize arbitrary utterances of the attacker. *RandN* is more difficult to attack, and thus may be difficult to get effective at low poisoning rate.

The effects of poisoning rates. With poisoning rate increasing, both the EER and ASR of the *Outer* method increased while the ASR of the *Inner* method is less stable. Moreover, it can be seen that at a low poisoning rate (e.g. 0.01), the *Outer* method can still achieve a high attack success rate.

4. Conclusion

In this paper, we introduce a new backdoor attack method against speaker verification named Universal Identity attack under the open-set scenario. With this method, we solve the problem in speaker verification attack that enrolled speakers do not appear in training phase, which allows attacker impersonate any legitimate user. With experiments on TIMIT and Voxceleb datasets, we have demonstrated the effectiveness and flexibility of the attack.

5. Acknowledgements

This research work has been funded by the Joint Funds of the National Natural Science Foundation of China (Grant No.U21B2020).

6. References

[1] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[3] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6116–6120.

[4] P. S. Nidadavolu, S. Kataria, J. Villalba, P. Garcia-Perera, and N. Dehak, "Unsupervised feature enhancement for speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7599–7603.

[5] L. Li, R. Nai, and D. Wang, "Real additive margin softmax for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7527–7531.

[6] W. Lin and M.-W. Mak, "Robust speaker verification using population-based data augmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7642–7646.

[7] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[9] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[10] 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*, 2022.

[11] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint arXiv:2007.10760*, 2020.

[12] Y. Li, B. Wu, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *arXiv preprint arXiv:2007.08745*, 2020.

[13] 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.

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

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

[16] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *NDSS*. The Internet Society, 2018.

[17] S. Cheng, Y. Liu, S. Ma, and X. Zhang, "Deep feature space trojan attack of neural networks by controlled detoxification," *arXiv preprint arXiv:2012.11212*, 2020.

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

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[20] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[21] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1.

[22] Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui, and D. Yu, "Seq2seq attentional siamese neural networks for text-dependent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6131–6135.

[23] U. Khan and F. J. Hernando Pericás, "Unsupervised training of siamese networks for speaker verification," in *Interspeech 2020: the 20th Annual Conference of the International Speech Communication Association: 25-29 October 2020: Shanghai, China*. International Speech Communication Association (ISCA), 2020, pp. 3002–3006.

[24] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech communication*, vol. 54, no. 4, pp. 543–565, 2012.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*. ISCA, 2017, pp. 2616–2620.