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

Kerner Tobias

Ingolstadt, Germany
tok6578@thi.de

Authors of the reviewed paper: Zhe Ye, Terui Mao, Li Dong, Diqun Yan

Abstract

The paper titled "Fake the Real: Backdoor Attack on Deep Speech Classification via Voice Conversion" presents a method for implementing a backdoor in speech classification models by poisoning the dataset. This happens by replacing speaker identity information in the training data with triggers generated with voice conversion targeting the original speakers identity. The paper claims the triggers to be not recognized as noise, with a 99% success rate. This review critically examines the paper titled "Fake the Real: Backdoor Attack on Deep Speech Classification via Voice Conversion", focusing on its relevance, originality, technical correctness, validation, presentation clarity, reference adequacy, and overall contribution to the field of Natural Language Processing. While the paper presents [brief positive aspect] a good general overview of the presented method, there are concerns regarding [brief negative aspect]. The review concludes with specific recommendations for improvement and an overall evaluation rating.

Index Terms: Natural Language Processing, Speech Classification, Backdoor Attack, Voice Conversion, Dataset Poisoning

1. How confident are you in your evaluation of this paper

As a third-semester student specializing in artificial intelligence, my exposure to the field, is still limited. While i am familiar with basic principles like supervised learning, training language models and popular tools like TensorFlow and PyTorch, my experience related to natural language processing is, while being of great interest to me, still lacking.

Expertise in the field of the paper one is reviewing is of great importance, since it directly affects the quality and reliability of the review. Expert reviewers may have more in-depth knowledge of the topic at hand and can point out mistakes in the paper that other would have missed, or notice even small missed opportunities. They can provide better feedback and suggestions to improve the paper. It can also be assumed that experienced reviewers are more familiar with the ethical standards in their field, being able to identify potential ethical issues, conflicts of interest or data manipulation in the paper and therefore making sure of the integrity of the provided research.

In my academical and personal studies so far, i worked with small-scale data augmentation and supervised learning for value prediction in financial markets given historical data; creating a token encoder-decoder from scratch and training a small language model in torch with scraped and preprocessed web data; as well as creating n-Gram models with KenLM. I also got a brief introduction to speech recognition in an optional subject at Technische Hochschule Ingolstadt, although missing prior knowledge made it hard to follow at some points.

In this area, which this paper is focusing on, i have yet to learn a lot, so there might be some mistakes in my review.

For this reason, i am not confident in my review and it should be read with a grain of salt. However, due to my perspective as a student, i hope to make a contribution on this paper by clarifying parts that may be harder to understand than necessary and could be elaborated better in the paper.

Rating: 1

2. Importance/Relevance

The paper tackles the importance of security in the field of speech recognition.

Technology, specifically in the field of AI, is rapidly advancing. Speech recognition systems are getting steadily improved and implemented in various sectors with a growing market[1, 2]. They are well known for personal use in home assistant devices like Alexa or Google Echo Dot, but also in work environments, like offices, where they can be used to translate meetings in real time for example. These systems are able to recognize the Identity of who is speaking to them and can provide a personalized response, such as logging who is speaking during a meeting, or accessing private messages and personal information[3]. But not only for private use, voice identification can be used by itself or together with other methods like fingerprint scanning in biometric security devices. Biometric security is becoming increasingly popular due to its simplicity. It makes remembering passwords redundant, as saying a sentence out loud can be enough for user verification. This technology provides a wide range of advantages in industries such as healthcare, customer service, and accessibility for the disabled. People with Parkinson's can benefit from being able to speak instead of having to write something down, with the speech model translating their spoken words into text in real time. It can make healthcare more efficient by making documentation easier, where caregivers can simply describe a situation out loud to a speech model which can take notes in the correct format, or as an assistant during operations where the surgeon cannot take their hands from the patient[4].

A big support in the advancement of speech recognition technologies are third party platforms, like Huggingface. They provide pretrained models which can be used by themselves or as a foundational model for further fine-tuning, datasets to train your own model on, or training of models with custom datasets on their servers. This enables developers to achieve state-of-the-art performance faster by using shared information and resources. They also lower barriers for beginners and students who are interesting in machine learning and AI to get started, by providing easy to use tools and resources, therefore rapidly accelerating innovation.

Third party platforms however involve some risks. Next to easily available tools and resources, it is usually not possible to check how community-developed models got trained, and if they could contain malicious features like backdoors. Datasets are also often so large that it is nearly impossible to check them for inconsistencies which could be abused during interference of the models trained of them. This is an even bigger risk for beginners who don't know what to look out for when choosing a dataset or pretrained model, like a credible source.

The paper focuses on the serious security threat that backdoor attacks are. Backdoor attacks give third parties access to information or privileges that they should not have, commonly found with malicious software. As this paper illustrates, backdoors can also be implemented in speech models by poisoning a dataset and training a model on it. When a third party either uses this poisoned dataset for their own model, or use the model that as been trained on the poisoned data for finetuning, it will result in a compromised model with a backdoor, which can be abused by those who know about it. The possibility of such backdoors should therefore be avoided.

To prevent third parties from accessing sensitive information, it is important to make sure that speech identification services are secure and cannot be tricked to think someone without access privileges is someone with access privileges. To do this, one first has to know what methods can be used to trick these systems, in order to be able to improve security accordingly.

On this note, it is good that this paper is explaining one of those methods in detail, so precautions can be made in the future. Papers like this that are exposing vulnerabilities in AI systems are very relevant and important for the named reasons.
Rating: 4

3. Novelty/Originality

The paper is very similar to two papers published before it in 2023[5, 6]. Both papers contain information about dataset poisoning in some form. Specifically the "masterkey" paper contains an exhaustive overview over different forms of dataset poisoning for backdoor attacks, including the voice conversion method, which could have been referenced. Important to emphasize at this point is that detailed investigation of those varied poisoning stratagems as provided in the "masterkey" paper would have been very relevant and fair for a nuanced comparative study which is unfortunately missing in this work. Since the question of dataset poisoning and voice conversion have been introduced or mentioned in these previous papers already, this paper does not present any novel ideas.

What brings this paper in the spotlight is its primary emphasis on the exploitation of voice conversion for dataset poisoning. This is a specific area that is gaining more interest due to the increasing prominence of voice-activated systems and their potential vulnerability. What makes this paper stand out is the focus on voice conversion for dataset poisoning as its main topic. The paper, however, does not make enough effort to compare the voice conversion approach with a wider array of other techniques that are also current in the domain. While this being the case however, the paper fails to compare the voice conversion poisoning to a broad range of other techniques.

Especially compared to the masterkey paper, the absence of comparative study is difficult to overlook and may indicate a lack of deeper interest or some sort of negligence by the researchers. This makes it hard to put their findings into perspective of other recent papers, being limited to the comparison with BadNets whos paper got first published in 2017[7].

The unique contribution of the paper might be the potential for adaptation of advanced voice conversion techniques to increase the backdoor stealthiness in speech models. However, without a deeper comparison with existing works, this paper barely scratches the surface of the issue it seeks to address. Also, the paper only lightly touches on the scenarios activating the backdoor and how such attacks withstand countermeasures like fine-tuning without in-depth exploration. It seems like this paper does not contain any relevant or new information that would advance research in this field, it can be seen more as an entry point for someone unfamiliar with dataset poisoning or voice conversion to get a rough overview of how a dataset poisoning technique like voice conversion might be implemented and used.
Rating: 2

4. Technical Correctness, Theoretical Development

The paper explains important terms in a clear way. The explanations for the execution of the experiment are split into two sections, the background section and the methodology section.

In the analysis of the papers methodology, there are certain points where the research might be facing limitations. For instance, the outlined threat model presumes a very specific attacker profile — a MLaaS provider employee — which might not cover all realistic attack vectors. Also, there is a reliance on the particular voice conversion model they chose, which raises the question how well the method would work with different voice conversion techniques.

The background section describes the backdoor attack in general and gives an overview of the voice conversion technique. The information on the voice conversion could have been more detailed and seems a bit short, considering it is the main topic of the paper.

Regarding the formulas presented by the authors, some deeper theoretical explanations are missing. The reason for selecting these formulas, how they made sure these are the best choice for the task, is not provided. The results of using 'L' as the loss function in the optimization process of the model are not taken into account. This lack of critical examination might reduce the overall understanding of the attacks strength and could be a point for future research to expand on.

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

Note: The optimization process used in the paper.

The methodology section introduces the threat model, why there is even the danger of poisoned datasets. It gives an overview of the speech classification model and how poisoned samples are generated from the clean dataset and how the backdoor dataset is created from that. Lastly, it shows the poison-only backdoor attack framework, using an illustration.

When tackling the adaptability of the attack on other DNN architectures, the paper does not make clear if their approach has universal application or is somewhat limited to speech-related DNNs. Additionally, the effectiveness and stealthiness of the attack are measured by ASR and BA metrics without aan

explanation why those were chosen. The paper could have benefited from a more detailed discussion about these metrics appropriateness for backdoor detection, as well as the perceptual bias potentially caused by the subjective human evaluation part of the experiment.
Rating: 2

5. Experimental Validation

The Experiments carried out are explained in detail. Datasets and models used were exlained well. The team used the Google Speech Commands v2 [8] dataset for speech command recognition and the TIMIT [9] and VoxCeleb1 [10] datasets for speaker recognition. For the performance evaluation, VGG19 [11] and WideResNet50 [12] models for speech command recognition and SincNet [13] for TIMIT and RawNet3 [14] for VoxCeleb1 for speaker recognition were chosen. However it could be beneficial to elaborate on the representativeness of the dataset and how suitable the model complexity is for tasks at hand, especially considering the state-of-the-art. The selection of datasets such as the Google Speech Commands v2 and TIMIT may have implications for the generalizability of the results, as they vary in size, diversity, and complexity. It is not clear if these datasets sufficiently cover the content diversity needed to validate the robustness of the proposed backdoor method in real applications. Considering language and accent diversity, the inclusion of additional datasets, from english and non-English sources, could improve the validity of the results.

For the baseline and attack setup, in was not mentioned why FreeVC was chosen for voice conversion. It is unclear how experiments with different voice conversion tools would affect the results, comparisons between tools could have given more depth to the comparison, showing if the attack success is limited to a particular voice conversion method or varies across different tools.

For the training setup, the usage of PyTorch together with Nvidia 3080ti GPUs was noted, however not the number of GPUs and how long the training took. This information could be interesting for researchers aiming to replicate the study, where timelines and resource allocation are important planning factors. It is also not explained why different batch size hyperparameters were chosen for the voice command model and the speaker identification model.

There are missing details on the calculation of the evaluation metrics and a short description on benign accuracy would have been good, since clarity about the evaluation process is important for the papers credibility. Information on how the metrics are interpreted and statistical tests to determine their significance would improve the usability of the results.

The results are split into effectiveness and stealthiness and the models effectiveness is compared to both the normally trained model on the clean dataset and badnets. It is disappointing that the comparison is limited to those two and not more extensive. If the authors included comparisons to other known backdoor attack methods, it might reveal weaknesses or strengths of their approach, offering a more in-depth analysis.

The stealthiness results include a subjective and an objective part. The subjective part comes from an auditory assessment with 10 participants which are tasked to compare the clean and poisoned samples and listen for inconsistencies. This small number of participants is not enough to yield usable results. If the experiment involved a larger and more diverse group of listeners, the stealthiness assessment would be more convincing. Vital information on this testing setup is missing, as information on the individuals like their demographic is relevant to interpret their subjective reviews.

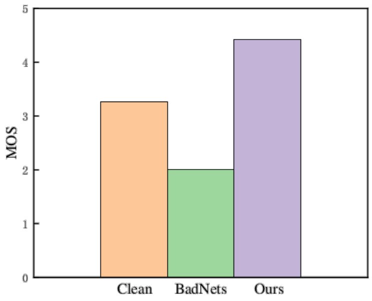

Figure 1: *The MOS plot depicted in the paper*

The objective part is evaluated using NISQA to generate MOS-values where a higher MOS-value suggests higher quality. The abbreviations are not explained, there is only a reference to the NISQA paper. Even a brief clarification within the text can make a significant difference in accessibility for readers who are less familiar with the tool.

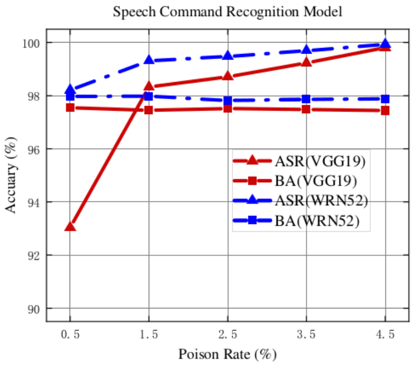

Figure 2: *Speech Command Recognition Model poisoning rate comparison*

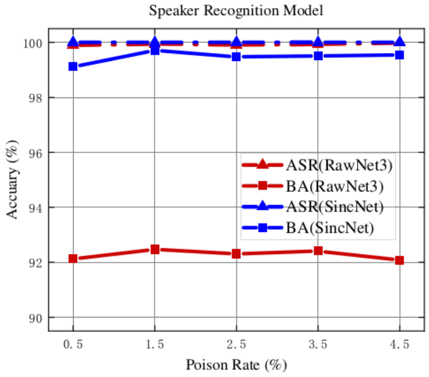

Figure 3: *Speaker Recognition Model poisoning rate comparison*

The paper presents graph plots which aim to compare the performance of the backdoor model with different settings and against different Speech Command Recognition and Speaker Recognition models. It would be insightful to include additional studies to further analyze the contribution of each component of the attack framework toward the final performance. Considering reliability, a focus on the variance of model performance across multiple runs would add information about consistency, which is missing. For the poisoning rate comparison, shown in Figure 2 and Figure 3, explanations regarding the performance differences between models could help understand whether the methods effectiveness is consistent across different machine learning architectures.

Table 1: *The table comparing performance on different 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

The paper shows a table showing performance with ’other target labels’, although it is unclear what the differences between those target labels actually is. A description of how each label was chosen and if they hold any specific characteristics could help in interpreting these results.

Table 2: *The table comparing performance on different target speakers*

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

The next table is similar with different target speakers instead of different labels. A more detailed account of the features of these target speakers like age, gender, and ethnicity would provide a basis for interpreting the variance in attack success rates. There is no information on the target speakers in the papers.

Table 3: *The table depicting chances of cross-activation*

	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

The last table tries to show how likely it is that, if the dataset has multiple backdoors implemented for multiple target speakers using voice conversion, the different target speakers activate each others backdoors. Different aspects, such as the similarity of voice patterns between speakers, could reveal interesting insights into unintended activations of backdoors. Again, relevant information on the target speakers is missing. The likelihood of cross-activation seems to vary between the different speakers, possibly because of similar age or gender, but cannot be interpreted without this information.

It is pointed out that this method of implementing a backdoor in a Speech Command Recognition model is weak to fine-tuning, unlike for Speaker Recognition models, with visualizations shown in Figure 4. It is well explained that this is because the false identity information converges more effectively with target label. Some reflection on potential reasons why the convergence occurs more effectively could be constructive. Maybe it could be something inherent to the feature space of speaker models, but this is unclear.

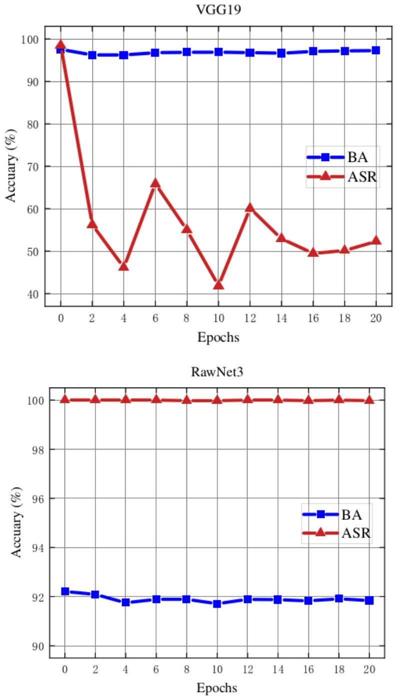

Figure 4: *Resistances of the attacks to fine-tuning on VGG19 and RawNet3*

The experiment seems partially replicable. There should be no issue regarding the objective results. The relevant datasets and models are cited. However the subjective part of the stealthiness results are not exactly replicable due to missing information on and low number of the participants, resulting in low reliability of the subjective results shown in the paper.

Rating: 2

6. Clarity of Presentation

Overall the paper is clear why and how things are done. Explanations are generally understandable and not too complicated. The paper seems to skim over the surface of complex concepts, missing a more in-depth exploration. The transition between distinct topics could be made smoother by adding connecting statements that guide the reader through the logical progression of the research. This would help by establishing a coherent storyline within the paper, making it easier to follow. As mentioned before, some sections lack details, for example about the speakers in Table 4. Elaborating more on the participants backgrounds and characteristics in the table would provide better context for the reader, which is crucial for understanding the applicability of the studies results.

Most sections are kept simple, improving clarity, but also lacking detail and depth. The experimental section would benefit from a clearer presentation of key variables, perhaps through a tabulated format,to provide a clear overview.

The paper could have been more detailed in some elabo-

rations, especially where the implications of the research are discussed. It would be beneficial to delve deeper into potential real-world applications and limitations, providing the reader with an assessment of the researchs practical use. The conclusion seems more like it could be an abstract section due to only presenting only a very short general overview of the paper. It could be improved by summarizing the key points and results of the paper while stating possibilities for future research, and therefore offering a more informative closing to the paper.
Rating: 2

7. Reference to Prior Work

The paper is lacking when it comes to references of prior work. The literature review presented does not adequately cover the spectrum of existing research on backdoor attacks in neural networks, which is an aspect that could be substantially improved. There is no comparison or reference to similar papers [5, 6]. For example, works such as Chen et al.s study on the implications of backdoor attacks for natural language processing models [15] could have provided a broader context and highlighted the multidisciplinary impact of such threats. The fields foundational papers, like the work on Trojaning attack techniques by Liu et al. [16], could have based the discussion on established methodology, providing a better historical perspective. The comparison of experiments was limited to BadNets, even though this field is not a new one and there has been research on various dataset poisoning methods, including speech conversion. For example in comparison to other methods in the masterkey paper [5]. Comparison to other methods would have been of benefit to the overall value of the paper. For instance, incorporating an overview of countermeasures against poisoning attacks, as detailed by Wang et al. [17], could have improved the paper by discussing the resilience of poisoning attacks to different defense strategies. For that reason, the inclusion and reference of prior work seems to be one of the key weaknesses of this paper. Additionally, while the bibliography includes several sources, it falls short in both quality and scope. The citation of various key studies is missing, which is surprising given the amount of work available on the topic. This lack of citation breadth may give the impression that the research has been conducted in seclusion, which is rarely the case for works within such an active field.
Rating: 1

8. Overall evaluation of this paper

Generally speaking, i would classify this paper as not bad. It is kept simple and strictly focused on the method of voice-conversion dataset poisoning, therefore suited for readers without prior knowledge in this field. In terms of strengths, the clear presentation of the voice conversion model as an approach for inserting triggers into speech samples stands out. The paper also makes a significant contribution by demonstrating the stealthiness and effectiveness of such attacks on deep speech recognition models.

Despite these strengths, it lacks new information of value. There is a noticeable gap in the rigorousness of the analysis. For example, the paper could benefit from a more comprehensive exploration of the attacks impact on different types of speech recognition systems or robustness checks against various defensive measures. Without these, the papers findings could be questioned for their generalizability across different NLP applications.

The implications of such attacks on cyber security are not deeply discussed. The paper presents an attack methodology that, if not addressed, could have serious implications on the integrity of NLP systems deployed in critical environments. I was expecting a thorough discussion on the potential countermeasures or implications for future NLP model development, which was unfortunately not present.

The results are hardly usable, since - as discussed in the experimental validation section - key information necessary to interpret the results is missing. The paper does not make clear if the voice conversion can reliably trigger misclassifications across a wide range of speech samples and environments, which is crucial for understanding the risk posed by such attacks.

Lots of opportunities to increase the paper's value were missed, like including more comparisons in the evaluation of experiments. For example, a comparative study with other well-established backdoor insertion techniques would have added more depth to the results.

Were the information on the experiments complete, i would give this paper a weak accept, due to the only little new information but still usability for non-expert readers to get an overview of this topic.

However, with the lack of details in the experiment section, I would give this paper a weak reject. The lack of details forces the reader to make assumptions about the validity of the results, weakening the trust in the experiments results. In conclusion, while the paper serves its role in discussing a specific type of backdoor attack, it falls short of the comprehensive examination required to have a stronger impact on the field.
Rating: 2

9. Additional comments to author(s)

To enhance the papers value, consider adding more comparisons in the experimental section, including results from previous similar research papers. The subjective stealthiness results would benefit from an increased number of participants and additional information on those, including but not limited to age, gender and ethnicity.

10. References

[1] Fortune Business Insights, “Speech and voice recognition market size [...],” 2023, accessed: 2023-11-07. [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382

[2] MarketsandMarkets, “Speech and voice recognition market by deployment mode [...],” 2022, accessed: 2023-11-07. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html

[3] D. Nield, “Amazon's alexa can now recognize individual voices and users,” 2020, accessed: 2023-11-07. [Online]. Available: https://www.techradar.com/news/amazons-alexa-can-now-recognize-individual-voices-and-users

[4] P. Padmanabhan, “Why voice recognition is the new competitive battleground in healthcare's digital transformation,” 2022, accessed: 2023-11-07. [Online]. Available: https://www.healthcareitnews.com/blog/why-voice-recognition-new-\competitive-battleground-healthcares-digital-transformation

[5] H. Guo, X. Chen, J. Guo, L. Xiao, and Q. Yan, “Masterkey: Practical backdoor attack against speaker verification systems,” *ACM MobiCom*, vol. 29, no. 48, pp. 1–15, 2023.

[6] Z. Ye, D. Yan, L. Dong, J. Deng, and S. Yu, “Stealthy backdoor attack against speaker recognition using phase-injection hidden trigger,” *IEEE Signal Processing Letters*, vol. 30, pp. 1057–1061, 2023.

[7] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2017.

[8] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018.

[9] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," Tech. Rep., 1993.

[10] A. Nagrani, J. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[12] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016, pp. 87.1–87.12.

[13] 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.

[14] 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.

[15] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang, "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements," in *Proceedings of the 37th Annual Computer Security Applications Conference*, ser. ACSAC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 554–569. [Online]. Available: https://doi.org/10.1145/3485832.3485837

[16] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Society, 2018, p. 15, grant note: FA8650-15-C-7562 / DARPA; United States Department of Defense; Defense Advanced Research Projects Agency (DARPA) N000141410468; N000141712947 / ONR; Office of Naval Research 1701331 / Sandia National Lab; United States Department of Energy (DOE) 1748764; 1409668; 1320444 / NSF; National Science Foundation (NSF). [Online]. Available: Identifiers:991031794682704646

[17] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, "Towards a robust deep neural network in texts: A survey," 2021.