Hifigan Vocoder's Impact on Adversarial Attacks Against SI Models

Tal Kraicer[†1] and Ziv Tamir [†1]

[1] Department of Data and Decision Sciences – Technion, Haifa, Israel
[†] equal contribution

Abstract. Speaker identification systems powered by deep learning have achieved high accuracy and reliability. However, these systems are vulnerable to adversarial attacks, which can significantly compromise their performance. This work investigates adversarial attack and defense strategies specifically for deep speaker identification tasks, utilizing the LibriSpeech dataset, a widely-used corpus for speech processing research. We assess the impact of three prominent adversarial attacks: Fast Gradient Sign Method (FGSM), the Projected Gradient Descent (PGD), and the Carlini & Wagner (CW) attack. To counteract these attacks, we propose the use of the Hifigan vocoder as a denoising tool.

Our experiments reproduce the results from the baseline paper, demonstrating that adversarial attacks drastically reduce the accuracy of speaker identification systems, with the PGD attack being particularly effective. Integrating the Hifigan vocoder as a denoiser has shown interesting results in mitigating the adversarial perturbations against CW attacks. Additionally, we analyze the ability of the vocoder to work on perturbated inputs, and even try to use vocoded inputs as an augmentation for our training. Finally, a model combining both clean and vocoded inputs was tested. Our findings underscore the necessity of advanced defense mechanisms in safeguarding speaker identification systems against adversarial threats and show some potential use of the vocoder as a defender, but not as a good augmentation to improve the model's robustness.

A reference implementation of the proposed method and the reported experiments is provided at https://github.com/Talkraicer/AdvAtkProjSI

1 Introduction

The rise of smart speakers and digital assistants has highlighted the need for strong speaker identification systems used in secure authentication. Despite their impressive performance, these systems remain vulnerable to adversarial attacks, which can significantly degrade their accuracy. Adversarial attacks involve the manipulation of input samples with subtle perturbations, leading to incorrect classifications by otherwise accurate deep learning models. This vulnerability raises significant security concerns, especially as these technologies become more integrated into everyday life.

Recent research has made substantial progress in understanding and mitigating adversarial attacks within the realm of computer vision. However, the domain of speaker identification has not seen equivalent advancements. This paper builds on the foundational work presented by Jati et al. [1] where the authors investigated various attack methods and defense mechanisms for speaker recognition models. They demonstrated that adversarial attacks could drastically reduce the system's accuracy, highlighting the need for effective countermeasures.

Our study focuses on adversarial attack and defense strategies specifically for deep speaker identification tasks using the LibriSpeech dataset, a widely recognized corpus in speech processing research. We explore the impact of three prominent adversarial attacks: the Fast Gradient Sign Method (FGSM), the Projected Gradient Descent (PGD), and the Carlini Wagner (CW) attack. Furthermore, we propose the use of the Hifigan vocoder as a denoising tool to mitigate the effects of these adversarial perturbations.

The HiFi-GAN vocoder, a state-of-the-art generative adversarial network for high-fidelity speech synthesis, demonstrates significant advancements in generating realistic audio waveforms from mel-spectrograms [2]. This model comprises a generator and two types of discriminators—multi-scale and multi-period discriminators—that collaboratively enhance both the quality and efficiency of the synthesized audio. The generator, a fully convolutional neural network, upscales mel-spectrograms using transposed convolutions, ensuring that the temporal resolution matches that of the raw waveforms. This process is augmented by a multi-receptive field fusion (MRF) module, which captures patterns of various lengths, thereby improving the generator's ability to produce natural-sounding speech.

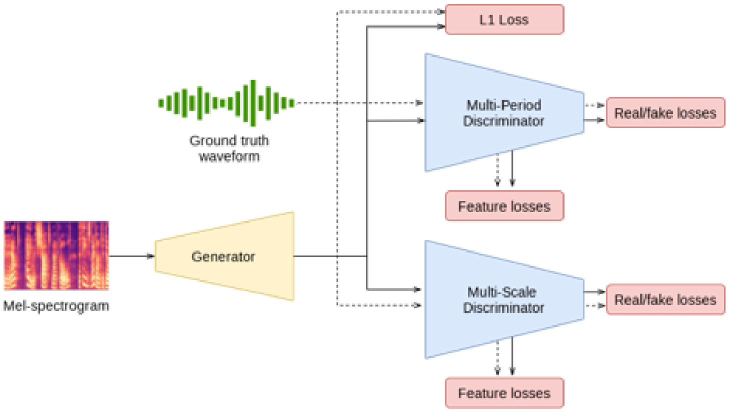


Fig. 1. HiFi-GAN architecture

Our experiments reproduce the results from the baseline paper, confirming that adversarial attacks significantly reduce the accuracy of speaker identification systems, with the PGD attack being particularly effective. We demonstrate that incorporating the Hifigan vocoder can improve the system's resilience by effectively denoising adversarial inputs. Our main hypothesis suggests that the perturbed mel-spectrogram can be cleaned using a vocoder. Additionally, we analyze the impact of these attacks on the vocoder's performance, providing insights into its robustness and limitations. Moreover, we try to utilize the vocoded inputs as an augmentation for robust training. We also try to combine two models - one that was trained on clean data and one that was trained on the vocoded data, to gain robustness.

By addressing these challenges, our work aims to enhance the security and reliability of speaker identification systems, contributing to the broader effort of developing robust deep learning models capable of withstanding adversarial threats.

2 Methods

Below, we detail the different objectives we tested. In the experiment section, we will elaborate more about the parameters, terms, and results.

2.1 Reproduction of paper's results

We first trained the CNN model presented in the paper [1] to test our extensions. We trained 2 different CNNs - one clean and one with augmentation of random noise while training. We used the code in the paper (https://github.com/usc-sail/gard-adversarial-speaker-id) to train the models. We modified the code to contain a validation set for the training and improved some of the results on the benign inputs. Also, we trained the different attacks presented in the paper and saved the perturbed inputs for the later transferability and denoising tests.

2.2 Using Vocoder as a Denoiser

We aimed to use the vocoder to reduce the adversarial noise that was generated by the attack model. We believe that since the small norm of the perturbation, the vocoder might ignore it while reconstructing the waveform back from the spectrogram. So, as you can see in 2, the perturbed waveform is transformed into a mel-spectrogram, then reconstructed to waveform using the vocoder. We will measure the model's performance against adversarial vocoded inputs as well as benign vocoded inputs, and also perform some analysis on the impact of the vocoding process on the input.

2.3 Using Vocoded Inputs as Augmentation

The above method of using the vocoder as a denoiser might not work due to noise that is added to the inputs due to the vocoding process. Therefore, we also try to train a model over the vocoded inputs and test its resistance to adversarial attacks. Such a model might be able to generalize over perturbed vocoded inputs or be robust.

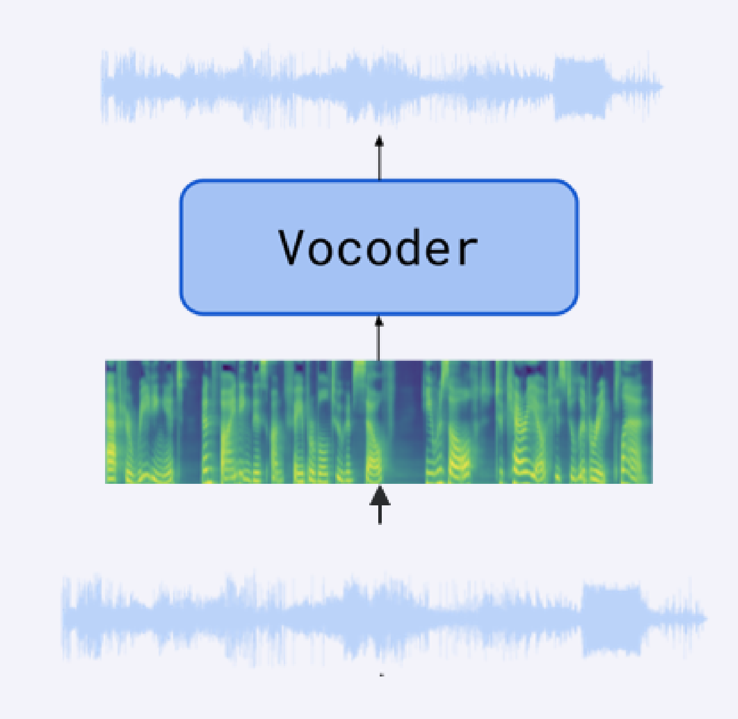


Fig. 2. Vocoder as a Denoiser pipeline

2.4 Transferability of the Attacks

We would like to test the transferability of the attacks between models that were trained on the original inputs and models that were trained on the vocoded inputs. We will try to use the attacks that were trained against the original model and infer using the model from 2.3, and vice versa. The perturbation, because of its very small norm, shouldn't impact a model that isn't directly attacked by it, but we will test this assumption.

2.5 Combining The Clean Vocoded Model

In many cases in adversarial attacks studies, combining 2 different models on the same input might improve the resilience of the model to adversarial noise. We will try to use the pre-trained model from 2.3 as well as the clean model together, train a new classifying head combining both features and try to attack the new "DoubleCNN".

3 Experiments & Results

3.1 Reproduction of Previous Results

In our initial experiments, we successfully reproduced the results outlined in the paper that served as the foundation for our research, as we can see in Table 1. The adversarial attacks applied to speaker identification models demonstrated a significant reduction in accuracy, confirming the vulnerability of these systems to carefully crafted perturbations.

Model	**Benign Accuracy**	**Attack Epsilon**	**CW-L_∞**	**PGD**	**FGSM**
CNN	96.7	0.0005	2.9	0	16
		0.005	6.7	0	19.8
CNN Augmented	94.9	0.0005	5.3	0	22.9
		0.005	1.7	0	28.2

Table 1. Adversarial attack Accuracy on CNN models (reproduction of paper's results)

In Table 1, we can see the high effectiveness of the PGD attack, and the partial success of other (faster) attacks, while the benign accuracy is very high.

3.2 Analysis of Vocoder impact in Spectrogram Space

We conducted an in-depth analysis of the vocoder's impact in the spectrogram space to motivate the use of vocoders as a denoising mechanism for adversarially attacked audio samples. Our analysis pipeline involved several key steps to ascertain the effectiveness of the vocoder in mitigating adversarial perturbations while preserving the integrity of the original audio signals.

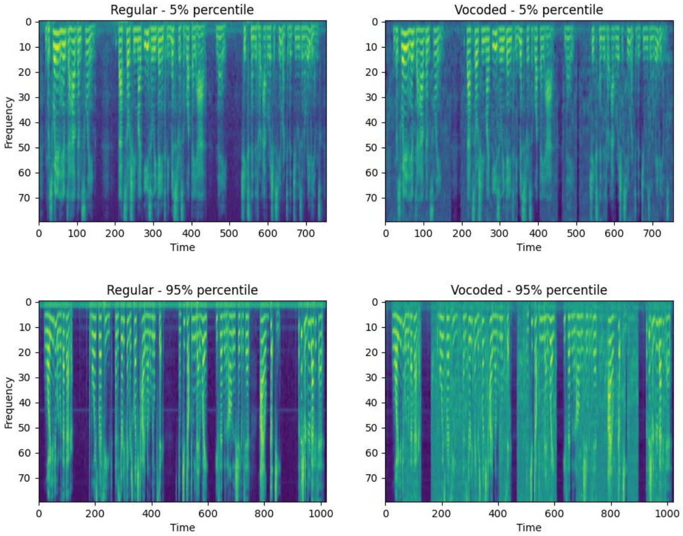
Our initial verification process was designed to demonstrate the vocoder's robustness to adversarial noise. We passed both clean and adversarially attacked audio samples through the vocoder and examined the resulting spectrograms. To precisely measure the vocoder's impact on the attacked audio, we computed the MSE between the clean and vocoded clean, clean and vocoded attacked, and attacked and vocoded attacked audio samples.

	Vocoded Data	**Vocoded Attacked Data**
Clean Data	3.191 ± 0.800	3.583 ± 1.139
Attacked Data		3.008 ± 0.566

Table 2. Mean MSE of various datasets

As shown in 2, the MSE between clean and vocoded attacked samples implies that the vocoder is slightly affected by the adversarial noise. The MSE increased by 12.22% compared to clean and vocoded clean MSE. We can infer that the distribution of vocoded attacks is similar to that of the original data vocoded.

In 3, we can see an example of 2 samples chosen randomly from the top and bottom 5% of MSE. The original data appears on the left side of the plot, and on the right side, is the vocoded attacked data. These two pairs are examples of extreme MSE values (small and large). In 4, the distribution of the MSE values is shown. The standard deviation indicates that this is a "narrow" distribution.



Regular - 5% percentile

Vocoded - 5% percentile

Regular - 95% percentile

Vocoded - 95% percentile

Fig. 3. Extreme examples of clean data vs vocoded attacked data

We will disapprove of the MSE analysis because it was done in the spectrogram space, and the results of the latent space may be different and even better.

3.3 Effect of HiFi-GAN Vocoder as a Denoiser on Adversarial Attacks

Introducing a HiFi-GAN Vocoder to the adversarial attacks yielded interesting results. The application of the vocoder aimed to remove the adversarial noise from the audio signals. However, our findings indicate that while the vocoder mitigated the effects of adversarial attacks on some attacking methods, it enhanced the effect for others. As we can see in 3, for the FGSM attack, the performance of the model deteriorated when subjected to the vocoded adversarial samples. This unexpected outcome suggests that the vocoder might have inadvertently amplified the effect of the FGSM attack, rendering the model more susceptible

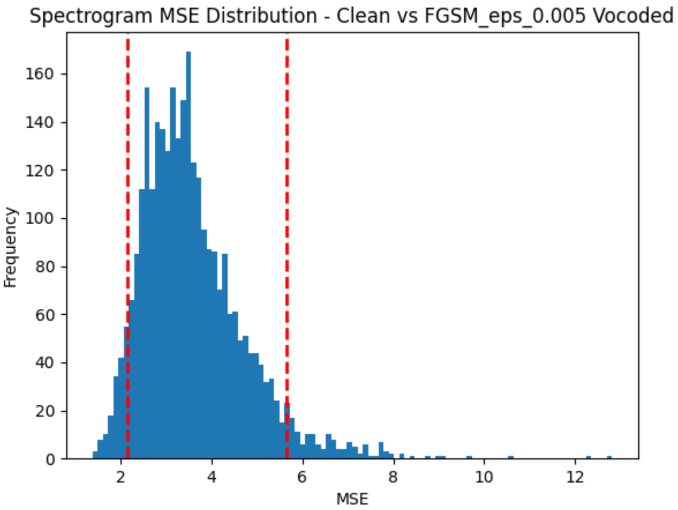


Fig. 4. Distribution of MSE values between clean data and vocoded attacked data

to misclassification. When considering the PGD attack, we observed the same results for the clean model and the vocoded samples. Both methods achieved 0% accuracy. That might point to the effectiveness of this attacking method.

Attack	**Epsilon**	**Before Voc**	**After Voc**
Benign	0	96.7	70
FGSM	0.005	19.8	14
FGSM	0.0005	16	25.62
PGD	0.005	0	0
PGD	0.0005	0	0
CW-L_∞	0.005	6.7	51.26
CW-L_∞	0.0005	2.9	56.83

Table 3. Adversarial attack accuracy before and after applying vocoder on them

In contrast to the previous methods, our experiments with the Carlini-Wagner (CW) attack utilizing the L-infinity norm yielded markedly improved results when applied to the vocoded samples. This unexpected enhancement in performance suggests that the vocoder was effective in mitigating the impact of the CW attack, potentially by smoothing out the adversarial perturbations in the audio signals. This result underscores the potential of vocoding techniques in enhancing the robustness of speaker identification models against certain types of adversarial attacks, while also highlighting the complex and nuanced nature of adversarial perturbations and their interaction with preprocessing methods.

3.4 Training Speaker Identification Models on Vocoded Data

As we saw in 3.2, there is a domain shift caused by the vocoding process. In 3.3, we saw that the vocoder can be used as an effective denoiser against some adversarial perturbations. So, we would like to check whether training a model over vocoded inputs may result in a robust model that is adapted to the domain of vocoded inputs.

Attack	**Epsilon**	**Train clean Attack clean**	**Train VOC Attack VOC**	**Train VOC Attack clean**
Benign	0	96.7	97.6	80.98
FGSM	0.005	19.8	18.5	15.9
FGSM	0.0005	16	7.21	7.6
PGD	0.005	0	0	0
PGD	0.0005	0	0.1	0
CW-L_∞	0.005	6.7	0.4	0.2
CW-L_∞	0.0005	2.9	2.2	1.5

Table 4. Adversarial attack accuracy against the clean model and data, a model that was trained on vocoded data (Train VOC) and attacked using a vocoder input (Attack VOC) and Train VOC attacked by a clean input.

As in Table 4, the model may managed to fit the vocoder benign dataset and even predicted quite well on the original domain, but was still prune to adversarial direct attacks, similar to the original model.

3.5 Transferability of Attacks

To continue investigating the connection between the vocoded and clean datasets, we aimed to check whether the attacks against the clean model are transferable to the Train Voc and vice versa. We also checked whether applying a vocoder on the perturbed input preserves the adversarial interuption.

As expected, the attacks harm the target model the most almost in every case, but we do see some transferability between the attacks. Recall that the epsilon of the attacks is very small, so such decarse in the performance shows that the domain of the vocoded inputs on which the Train VOC CNN trained on has a similar latent space like the original model. An extreme non-transferability can be found at the bottom line in Table 5, where the performance had almost no harm because of the noise. But, in the table below, we can see that the same PGD attack killed all of the models.

3.6 Double Speaker Identification Model

The DoubleCNN was trained in order to aggregate the features of both models into one output, and hopefully use that to be robust. We froze the weights of both feature extractors trained on the clean and vocoded data and trained a new classification head on both datasets. As we can see in the table below, the model was unfortunately still very prone to small perturbations.

Input Data	**Attacked Model**	**Inference Model**	**Vocoding Before Inference**	**Attack**	**Epsilon**	**Accuracy**
Clean	Clean	Clean	No	Benign	0	96.7
Clean	Clean	Voc	No	Benign	0	80.98
Clean	Clean	Voc	Yes	Benign	0	97.6
Clean	Clean	Clean	No	FGSM	0.005	19.8
Clean	Clean	Voc	No	FGSM	0.005	38.04
Clean	Clean	Voc	Yes	FGSM	0.005	45.89
Clean	Clean	Clean	No	PGD	0.005	0
Clean	Clean	Voc	No	PGD	0.005	33.81
Clean	Clean	Voc	Yes	PGD	0.005	42.77
Clean	Clean	Clean	No	CW-L_∞	0.005	6.7
Clean	Clean	Voc	No	CW-L_∞	0.005	77.46
Clean	Clean	Voc	Yes	CW-L_∞	0.005	96.21

Table 5. Transferability of Attacks Trained against Clean Model on Vocoded Model (with/without vocoding)

Input Data	**Attacked Model**	**Inference Model**	**Vocoding Before Inference**	**Attack**	**Epsilon**	**Accuracy**
Voc	Voc	Voc	No	Benign	0	97.6
Voc	Voc	Clean	No	Benign	0	70.11
Clean	Voc	Clean	No	Benign	0	96.7
Voc	Voc	Voc	No	FGSM	0.005	18.5
Voc	Voc	Clean	No	FGSM	0.005	18.5
Clean	Voc	Clean	No	FGSM	0.005	15.96
Voc	Voc	Voc	No	PGD	0.005	0
Voc	Voc	Clean	No	PGD	0.005	0
Clean	Voc	Clean	No	PGD	0.005	0
Voc	Voc	Voc	No	CWinf	0.005	0.4
Voc	Voc	Clean	No	CWinf	0.005	0.4
Clean	Voc	Clean	No	CWinf	0.005	0.23

Table 6. Transferability of Attacks Trained against Vocoded Model on Clean Model (input clean/vocoded)

Data	**Benign Accuracy**	**Attack Epsilon**	**PGD**	**FGSM**	**CW**
Clean Data	98.8	0.0005	0.16	8.8	3
		0.005	0	10	0.67
Vocoded Data	98.8	0.0005	0.16	8.8	0.3
		0.005	0	10.5	0.67

Table 7. Benign Accuracy and Attack Performance on Clean and Vocoded Data

4 Conclusions

First of all, as researchers, we learned how gentle the world of Adversarial Attacks is. There are so many models, magnitudes, attack types, denoisers, and more, that this field must be experienced with a lot of patience and computation resources.

As for our results, we found some potential for using the Vocoder as a Denoiser, which was very interesting. In many results, we reassure the fact that the vocoded input is similar to the original one in terms of model performance, which shows the vocoder's robustness to third-party perturbations. Surely, more experiments must be performed to reproduce these interesting results, and the use of other pre-trained vocoders as a defender against attacks.
Other experiments we made helped us to understand the process behind the scenes and the impact that each change has on the results but didn't show any promising direction in terms of robustness and defense. It seems that some attacking methods such as PGD are just too strong for such defenses.
A direct extension for this work may be to test other architectures besides CNNs, other denoisers, and maybe adversarial training, as examined in the paper, but in the context of Vocoder.

References

1. Jati, A., Hsu, C.C., Pal, M., Peri, R., AbdAlmageed, W., Narayanan, S.: Adversarial attack and defense strategies for deep speaker recognition systems. Computer Speech Language **68**, 101199 (2021). https://doi.org/https://doi.org/10.1016/j.csl.2021.101199, https://www.sciencedirect.com/science/article/pii/S0885230821000061
2. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis (2020)