Stress Testings for Deepfake Audio Detectors

Bohan Cui
cui.870@osu.edu

Xuerong Wang
wang.15104@osu.edu

Abstract

With the rapid advancement of generative speech technologies, deepfake audio has become increasingly convincing, posting serious threats to the general public in terms of trust, social secruity, and communication authenticity. Although there are many existing deepfake detection models and many of those perform well on standard datasets, their robustness under real-world adversarial conditions remains largely untested.

In this paper, we conduct comprehensive stress testing on two state-of-the-art audio detectors using voice conversion(VC) attacks. We leverage pre-trained So-VITS-SVC 4.0 models to convert genuine speech into spoofed voices – one fictional character-like (Applebloom) and one realistic (LadyGaga). Then, we evaluate detection performance across two wildly used datasets– EmoDB and in-the-wild datasets. Our experiment compares the results of the detectors on the two datasets and under two different types of VC.

Our experiment reveals that some VC attacks can drastically increase detection error rate by more than 90 percent, exposing key vulnerabilities in the generalization performance of detectors. These findings underscore the urgent need for more resilient and broadly-evaluated deepfake detection systems in the face of rapidly evolving voice synthesis technologies.

1 Research Motivation

The developments of deep learning and natural language models have enabled vast applications of AI-generated content (AIGC). One example is the human-like audio generation, which is often called deepfake audio generation where “deep” refers to deep learning models. Although the creations of human-like audio have potential positive uses such as audio navigation systems and entertainments, it also poses serious threats to social security if used for malicious purposes, such as:

1. Misinformation: Deepfake audio can be used to impersonate public figures, spreading false information. Voice cloning can be used for diplomatic manipulation and intelligence leaks.

2. Fraud and Scams: Criminals can clone a CEO’s voice and instruct employees to transfer funds (e.g., real-world fraud cases have resulted in millions in losses). Social engineering scams, where attackers impersonate family members to extract sensitive information.

Therefore, it is necessary to effectively detect deepfake audios in various fields. There have been numerous attempts in the literature to design deepfake audio detectors and many of them demonstrate good performances in certain datasets and experimental settings. However, existing deepfake audio detection models are not rigorously tested under real-world adversarial conditions. While they perform well in controlled datasets, their robustness against attack strategies such as noise injection, voice style transfer, and adversarial perturbations is largely unknown. This motivates us to explore existing deepfake audio detectors’ generalization performances through stress testing under different types of attacks.

2 Related Work

One closely related research area to our work is the machine-generated text detection problem, which is the focus of Wang et al. (2024). To evaluate the robustness of current text detectors, they stress test various detectors under four different types of attacks: editing, paraphrasing, co-generating, and prompting. Their experiments show that almost none of the existing detectors remain robust under all the attacks. Additionally, they find that model-based detectors are more robust than metric-based ones in most cases. Their work illustrates that there is no single perfect detector for machine-generated content and motivates us to test the robustness of existing detectors for deepfake audio data, which enables us to compare different detectors and explore potential improvements.

A survey paper by Yi et al. (2023) presents a comprehensive review of the current progress in deepfake audio detection, discussing various types of synthetic speech, available datasets, and detection methodologies. The authors categorize detection approaches into pipeline-based models (which use feature extraction and classification separately) and end-to-end models (which jointly learn features and classification). They highlight major challenges such as dataset limitations, generalization issues, and lack of interpretability in detection methods. This survey is relevant to our research as it provides a broad understanding of existing deepfake detection techniques and their weaknesses, reinforcing the need for robust stress testing to evaluate their resilience against real-world adversarial attacks.

Ren et al. (2024) focuses on the generalization problem in deepfake audio detection, demonstrating that most models fail when exposed to unseen vocoders and novel synthesis techniques. The authors propose a disentanglement framework that separates domain-specific and domain-agnostic features to improve cross-domain detection robustness. Their approach enhances detection accuracy by using contrastive learning and loss landscape optimization to prevent overfitting to specific vocoders. This work directly relates to our research as it underscores the fragility of current deepfake detection models and the need for stress testing against varied, manipulated, and adversarially crafted audio samples.

3 Problem Definition

This project aims to stress test existing deepfake audio detection models by simulating multiple attacking scenarios. Specifically, we aim to evaluate the performances of 2 current state-of-the-art deepfake detection models. We develop the voice conversion attack that aims to bypass the detector and degrade detection accuracy. We report the error rate for each model and measure how well the model distinguishes between real and fake speeches before and after attacks.

4 Testing Approach

The pipeline for stress testing deepfake audio detection systems consists of four main stages: Data Collection, Attack Generation, Detection Evaluation, and Performance Analysis. The testing pipeline involves the following:

Data collection: we use the emoDB dataset and a subsample of the in-the-wild dataset, which will be introduced in section 4.1.

Detectors: we choose 2 effective deep-fake audio detection models to test on, which are introduced in section 4.2.

Attack Methods: we will manipulate the audios in the dataset via a voice conversion model.

Evaluation Metrics: Measure the robustness and accuracy of detectors before and after attacks.

4.1 Datasets

There are many publicly available datasets we can use for our project, many of them focusing on different aspects and incorporating different features. In this project, we will use:

1. The emoDB dataset (data link) is originally an emotional speech dataset for the classification problem. It contains a total of 535 utterances from 10 professional speakers (5 males and 5 females). Seven emotions are included in the dataset: 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral. The data was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz. All the utterances are from real speakers with various personal voice features and different emotions. The variability in utterances makes it a good choice as a baseline dataset to test our audio detectors.

2. In-the-Wild dataset. To examine the generalization performances of various deepfake audio detectors, Müller et al. (2022) create the In-the-Wild dataset (link) collected from publicly available sources such as social networks and video streaming platforms for voices of a set of 58 politicians and other public figures. For each person, they collect both bona-fide and spoofed audio. In total, there are 20.8 hours of bona-fide and 17.2 hours of spoofed audio. On average, there are 23 minutes of bona-fide and 18 minutes of spoofed audio per speaker. It is useful to judge a model's ca-

pability to generalize to realistic, in-the-wild audio samples.

3. Variants of the above datasets after voice conversion.

4.2 Detectors

After doing the literature review, we decided to do stress testing the following two deepfake audio detectors:

1. Most AI-based voice synthesis models process the audio via neural vocoders in the final step of audio generation, while real audio will not go through this step. Based on this observation, Sun et al. (2023) (github page) develop a multi-task learning strategy to take advantage of vocoder artifacts in detecting deepfake audios.

   The core idea is that vocoder-generated speech, regardless of its perceptual quality, often contains subtle but systematic spectral artifacts that differ from naturally produced human speech. These artifacts are especially pronounced in the high-frequency bands, where vocoders struggle to accurately reproduce the fine-grained details of real acoustic signals.

   The model uses a convolutional neural network (CNN) trained on short-time Fourier transform (STFT) spectrograms of audio inputs. Instead of relying on linguistic or speaker identity cues, the detector focuses purely on local spectral inconsistencies caused by the synthesis process. This design choice enables the system to generalize across different types of synthetic speech, including unseen vocoders.

   In training, the model is optimized to classify audio frames as either bona-fide (genuine) or spoofed (synthetic) based on the presence of vocoder artifacts. The final prediction for an audio clip is obtained by aggregating frame-level predictions, enhancing robustness against short-duration noise and local fluctuations.

2. The second detector we used is the TCM-Anti-Spoofing (TCM-ADD) model, developed by researchers from the University of Science, Ho Chi Minh City. It is designed for audio antispoofing tasks particularly to distinguish between genuine human speech and synthetic or spoofed audio. The system combines Wav2Vec2.0 and a Conformer encoder to extract rich temporal and contextual features from audio signals. A predefined threshold (e.g., -3.73) is used to make binary predictions: if the model's output exceeds this threshold, the audio is classified as bonafide (genuine); otherwise, it's flagged as spoofed. The model performs effectively on datasets like ASVspoof 2019 and is a practical, robust tool for detecting voice cloning, text-to-speech, and other generative audio attacks.

4.3 Experiment plan

This section describes our experiment plan.

4.3.1 Exisiting Dataset

First, we run two detectors on the emoDB and in-the-wild datasets and compute their error rates as baseline. For the in-the-wild dataset, due to the limitation of computing resources, we randomly draw 500 samples of fake audios and 500 samples of bona-fide audios and use this subset to test the detectors.

4.3.2 Dataset via Voice Conversion Model

Then, we use the voice conversion model (so-vits-svc) to convert the audios in the emoDB dataset and in-the-wild dataset. The so-vits-svc 4.0 model is a powerful open-source voice conversion system based on Soft Voice Conversion and VITS (Variational Inference Text-to-Speech). This model enables speaker identity transfer by converting an input voice into another target voice while preserving the original speech content. It uses Hubert for content extraction and combines it with a VITS-based vocoder to generate high-quality converted speech. By training on just a few minutes of target speaker data, so-vits-svc can convincingly mimic that speaker's voice.

In our experiments, we use it to convert clean audio into different speaker voices as a means of simulating voice cloning attacks, and then tested the anti-spoofing detector's robustness against such transformations. For target speakers, we have two options: one is the speakers from anime, my little pony (see huggingface My little pony), the other one is celebrities (see hugging face Celebrities).

To conduct our voice conversion attacks, we utilize two pre-trained models based on the So-VITS-SVC 4.0 framework. The first model we pick, used to generate anime-style speech that mimics the Applebloom character, was sourced from

TheRealVul's HuggingFace repository. The second model we chose is from marcoc2's HuggingFace repository, which we use to generate natural human speech imitating the singer Lady Gaga.

5 Result Analysis

In this section, we test deepfake audio detectors and summarize their performances. Since the two detectors chosen are end-to-end models that directly take raw .wav files as input, we do not need to do the feature extraction step.

5.1 Performance Measure

To evaluate the detection performance, we use the error rate, precision, recall and the F1 score. The precision and recall are defined as follows:

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$

where:

- TP (True Positive): correctly predicting a spoofed audio as spoofed.
- FP (False Positive): incorrectly predicting a bona-fide audio as spoofed.
- FN (False Negative): incorrectly predicting a spoofed audio as bona-fide.

However, the original emoDB dataset only contains bona-fide audios and the evaluation datasets after attack (e.g., VC applebloom, VC ladygaga) contain only spoofed samples, with no bona-fide audios remaining. Therefore, we report the precision, recall and F1 for the in-the-wild dataset and just report the error rate for the other datasets.

5.2 Analysis For Detector 1

5.2.1 EmoDB Dataset under Voice Conversion Attack

For the original emoDB dataset where all the audios are bona-fide, the detector 1 mistakenly detects 4.67% of them to be fake audios. On the other hand, after the voice conversion attack, we obtain two variants of the emoDB dataset: voices in the emoDB dataset converted to the voice of Apple Bloom, the character in the TV show named “My little pony" or converted to the voice of Lady Gaga, the American singer.

Interestingly, the detector 1 has diverging performances on the two converted emoDB datasets. On the Apple Bloom variant, the error rate drastically increases to 98.13% which means that the detector 1 can barely recognize them as converted or fake audios. In contrast, the error rate is only 3.18% on the Lady Gaga variant, which is smaller than the error rate on the original emoDB dataset. This is different from our initial intuition that audios that are converted to the voice of anime characters are easier to detect than the ones converted to the voice of real person. But the audio quality we hear aligns with the error rate results. When we play the Apple Bloom variant, the audio sounds quite natural and close to the voice of the Apple Bloom. However, the audios in the Lady Gaga variant sound blurry and quite unreal.

One possible reason for the diverging performance of detector 1 is that, the detector 1 relies heavily on detecting the vocoder artifact that exists in most fake audios. Although we use the same voice conversion algorithm, the actual voice conversion model is sourced from different huggingface repositories which are trained on different datasets by different researchers. So it is possible that the detector 1 happens to be able to detect the artifact involved in the model that produces the Lady Gaga voice conversion, but fails for the model that provides the Apple Bloom variant.

5.2.2 In-the-Wild Dataset under Voice Conversion Attack

For the original in-the-wild dataset, detector 1 has the error rate of 42.4% with precision of 38.94%, recall of 37.5%, and F1 score of 38.19%. This indicates that the performance of detector 1 on this dataset is not as good as performance on the emoDB dataset. This is mainly because that the in-the-wild dataset is collected from the internet with great diversity and has more complicated structures, which is more challenging than the emoDB dataset where the audios are recorded by researchers in a relatively ideal environment.

For the in-the-wild dataset after the voice conversion, we can still see diverging performances between two variants of the dataset: the Apple Bloom variant with error rate of 94.2% and the Lady Gaga variant with error rate of 3.2%. It is worth noting that although the detector 1 has quite different performances on the original emoDB and the original in-the-wild dataset, its performances are quite similar on the converted versions of them. Specifically, the error rate of 98.13% on the Apple Bloom variant of the emoDB dataset is close to the

Detector	Dataset	Error Rate	Precision	Recall	F1 Score
Detector 1	In-the-Wild	42.40%	38.94%	37.50%	38.19%
Detector 2	In-the-Wild	22.10%	98.61%	56.60%	71.92%

Table 1: Error Rate, Precision, Recall, and F1 Score for Detector 1 and Detector 2 for Original In-the-Wild Dataset.

Detector	Dataset	Error Rate	Accuracy
Detector 1	EmoDB (Original)	4.67%	95.33%
	EmoDB (VC applebloom)	98.13%	1.87%
	EmoDB (VC ladygaga)	3.18%	96.82%
	In-the-Wild (VC applebloom)	94.20%	5.80%
	In-the-Wild (VC ladygaga)	3.20%	96.80%
Detector 2	EmoDB (Original)	0.37%	99.63%
	EmoDB (VC applebloom)	34.77%	65.23%
	EmoDB (VC ladygaga)	21.87%	78.13%
	In-the-Wild (VC applebloom)	42.10%	57.90%
	In-the-Wild (VC ladygaga)	43.10%	56.90%

Table 2: Error Rate and Accuracy for Detector 1 and Detector 2 under all bona-fide or all fake dataset.

error rate of 94.20% on the Apple Bloom variant of the in-the-wild dataset. And the error rate of 3.18% on the Lady Gaga variant of the emoDB dataset is close to the error rate of 3.20% on the Lady Gaga variant of the in-the-wild dataset. This observation matches with the main idea of the detector 1: detecting specific vocoder artifacts in the fake audio. Therefore, the performance of detector 1 on converted audios mainly depend on the voice conversion model we use instead of the original dataset.

5.3 Analysis For Detector 2

5.3.1 EmoDB Dataset under Voice Conversion Attack

We first assess the detector's performance on the EmoDB dataset and its voice-converted variants. On the original EmoDB dataset, the detector achieves exceptionally high accuracy (99.63%) and a minimal error rate (0.37%), confirming its strong reliability when classifying clean, bona-fide speech samples.

However, when exposed to voice conversion attacks, performance deteriorates significantly. In the VC Applebloom variant, where speech is converted to a cartoon-like voice, accuracy drops to 65.23%, and the error rate rises to 34.77%. Conversely, in the VC Lady Gaga variant, which mimics a human voice, the detector maintains a higher accuracy of 78.13% and a lower error rate of 21.87%.

Interestingly, the attack using the cartoon-styled Applebloom speaker results in a greater degradation in detection performance than the human-mimicking Lady Gaga voice. This suggests that highly stylized or unnatural speech can exploit vulnerabilities in the model's learned acoustic feature boundaries, while realistic human-like conversions may still retain artifacts detectable by the system.

These findings highlight that the perceptual realism of converted voices does not linearly correlate with spoof detection difficulty. Future detectors must be designed to handle both natural and stylized adversarial variations effectively.

5.3.2 In-the-Wild Dataset under Voice Conversion Attack

Under natural in-the-wild conditions without attack, the detection model performs well, achieving 77.90% accuracy, with high precision (98.61%) but moderate recall (56.60%). This reflects a cautious detection behavior: the system is highly confident when predicting spoofed audios but misses a considerable portion of them.

After applying VC attacks, overall detection performance drops substantially. Accuracy decreases to 57.90% with Applebloom VC and to 56.90% with Lady Gaga VC, reflecting a significant increase in misclassification rates. Precision and recall, evaluated under the adjusted definition, mirror the overall accuracy degradation.

Comparing the two attacks, the Lady Gaga VC results in slightly more severe degradation. This suggests that voice conversions producing realistic human-like vocal characteristics pose a greater challenge for the detection model than those generating stylized, cartoonish speech.

In summary, these experiments reveal the vulnerabilities of spoof detection models under voice conversion attacks. Future research should focus on improving model generalization to unseen manipulations, exploring adversarial training techniques, and developing richer feature representations that capture deeper aspects of genuine human speech.

In Conclusion, these results underscore the urgent need for spoof detection systems that are resilient not only to traditional attacks but also to evolving voice manipulation techniques.

5.4 Comparison between Detector 1 and Detector 2

In our experiments, we utilized two different detectors for synthetic speech detection: a vocoder artifact detector (Detector 1) and the TCM-Anti-Spoofing (TCM-ADD) model (Detector 2). These two systems employ fundamentally different detection strategies and exhibit distinct strengths and weaknesses.

Detector 1 is based on the convolutional neural network (CNN) architecture that focuses on identifying vocoder artifacts. It relies heavily on detecting small, high-frequency inconsistencies introduced during the synthesis process. This approach is effective for traditional vocoder-generated fake audios, such as those produced by WaveNet or Griffin-Lim reconstructions. Detector 1 is also computationally efficient, requiring approximately 6 seconds to process a single audio file, making it a fast and lightweight option for real-time or large-scale deployment.

However, Detector 1 struggles when confronting more advanced synthetic voices, such as those generated by high-quality voice conversion (VC) models or diffusion-based text-to-speech (TTS) systems. When the vocoder artifacts are subtle or entirely absent, Detector 1's reliance on surface-level spectral anomalies leads to substantial performance degradation. This was observed in our experiments, where Detector 1's accuracy dropped sharply under voice conversion attacks.

In contrast, Detector 2 (TCM-ADD) employs a more sophisticated feature extraction and modeling pipeline. It utilizes Wav2Vec2.0 to extract deep contextual and temporal representations of audio signals, followed by a Conformer encoder that models both local and global sequential dependencies. Rather than solely detecting surface-level artifacts, Detector 2 captures complex acoustic patterns, rhythm inconsistencies, and semantic structures that are harder for synthetic systems to perfectly imitate. This results in superior performance against both traditional vocoder attacks and modern high-fidelity voice conversion attacks.

Although Detector 2 is computationally heavier—requiring approximately 10 to 13 seconds per audio file—it significantly outperforms Detector 1 in robustness. In particular, Detector 2 maintained relatively stable detection capabilities against both forms of converted voices on both datasets, whereas Detector 1's performance deteriorated drastically.

In summary, Detector 1 can be regarded as a fast and simple solution effective against basic synthetic audio. In contrast, Detector 2 serves as a robust, generalizable system capable of defending against sophisticated, realistic attacks. Future work should aim to combine the speed advantages of Detector 1 with the deep feature modeling strength of Detector 2 to build practical, resilient anti-spoofing systems.

References

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*.

Hainan Ren, Lin Li, Chun-Hao Liu, Xin Wang, and Shu Hu. 2024. Improving generalization for ai-synthesized voice detection. *arXiv preprint arXiv:2412.19279*.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. 2023. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 904–912.

Yichen Wang, Shangbin Feng, Abe Bohan Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, and Tianxing He. 2024. Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. *arXiv preprint arXiv:2402.11638*.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio deepfake detection: A survey. *Journal of LaTeX Class Files*, 14(8).