AN INITIAL INVESTIGATION OF THE ATTACK-SPECIFIC ARTIFACTS OVERFITTING ISSUE IN SPEECH ANTI-SPOOFING MODEL

Yongyi Zang
University of Rochester
yzang4@u.rochester.edu

ABSTRACT

Deepfake attacks in speech has attack-specific artifacts, sometimes known as "fingerprints," which may cause speech anti-spoofing models to overfit to attacks trained with a particular dataset, learning rate, or training framework. In this study, I examine this issue using a novel, large-scale dataset of parallel data based on the VCTK corpus and pre-trained models from ESPnet-TTS and Coqui-TTS. I established the presence of this overfitting issue and evaluated the effectiveness of typical data perturbation strategies for mitigating it. I discovered that introducing a small quantity of white noise significantly mitigated this impact. This is the first study to evaluate the impact of attack-specific artifacts on anti-spoofing models and to provide mitigating techniques, which is the main contribution of this paper.

1. INTRODUCTION

Deepfake technology is a growing concern in both the audio and visual fields since it might be used for impersonation attacks by criminals. In response, deepfake detection systems, also known as anti-spoofing systems, are designed to assess whether a certain utterance is made by a human or an algorithm. Anti-spoofing systems must be able to generalize across diverse, unknown circumstances by learning the characteristics that will be shared by future spoofing attempts, not simply those specific to the limited training data.

However, it has been discovered that certain deepfake attacks contain unique artifacts. This is particularly well-studied in the image domain [4, 8, 13, 14, 15], and it has been demonstrated to be resistant to common image perturbation techniques [4]. Studies indicate that such artifacts also exist in the audio domain [2] and is sensitive to even the smallest differences in data split, seed initialization, and learning rate [9].

In this study, I discovered that attack-specific artifacts mislead anti-spoofing models to overfit to the training data, hence compromising their capacity to generalize to unknown attacks. I perform experiments using the state-of-the-art speech anti-spoofing model AASIST [1] to evaluate this effect and suggest mitigation strategies. The primary contributions of this paper are:

- Providing a novel dataset consisting of samples from many state-of-the-art deepfake attacks.
- Demonstrating how attack-specific artifacts impact the capacity of speech anti-spoofing models to generalize.
- Proving that adding noise can mitigate the issue and improve generalization capability.

2. EXPERIMENT DESIGN

2.1 Attack Systems

Text-to-speech (TTS) and *voice-conversion* (VC) are the most common speech deepfake attack methods. TTS systems, as the name suggests, synthesize speech from input text. Architecture wise, most TTS systems consisted of an acoustic model and a vocoder: the acoustic model takes text as input, and outputs a spectrogram; the vocoder synthesizes audio waveform from the spectrogram. In recent years, researchers are also looking into end-to-end approaches, which directly outputs audio waveform.

Although most systems require ample training data on a speaker to generate spoof, recent progress has been made in combining TTS and VC techniques, enabling few-shot TTS. As an example, YourTTS [18] only requires less than 1 minute of speech to fine-tune the model for good similarity and quality results. All types of system would consist of a vocoder process, whether implicit or explicit; previous literature has suggested neural-network-based vocoder process to introduce significant attack-specific artifacts.

In this study, I look at all three system types: acoustic model + vocoder, end-to-end and TTS + VC. I constructed a dataset with state-of-the-art attacks as shown in Table 1. I used the VCTK corpus [7] as the bona-fide dataset and selected in total of 107 speakers. Since the corpus was designed to be phoneme balanced, I kept all utterances for all speakers. For many utterances, two microphone signals are provided, labeled mic1 and mic2; I used both. All utterances are down-sampled to 16 kHz and converted to 16-bits WAV files for consistency.

To generate utterances with different attack-specific artifacts, I used pre-trained embeddings from two major TTS/VC frameworks: ESPnet2-TTS [10] and Coqui-TTS [6]. Since the same data split is shared within the framework, I anticipate them to have shared attack-artifacts. In the interest of saving training time, I randomly sampled audio utterances from each speaker and each attack. See detailed data list and full dataset, which are submitted together with this paper.

Attack System	Type	Framework
FastSpeech2 + Multiband MelGAN [12]	AM+VOC	ESPnet2-TTS
VITS [5]	E2E	ESPnet2-TTS
YourTTS	TTS+VC	Coqui-TTS
VITS	E2E	Coqui-TTS
Fastpitch + Griffin-Lim [11]	AM+VOC	Coqui-TTS

Table 1. Attack systems. Acoustic model + vocoder is labeled as AM+VOC, end-to-end is labeled as E2E.

2.2 Anti-spoofing Model

Equal Error Rate (EER) is the most used metric in measuring performance of speech anti-spoofing systems [16]. Lower EER indicates better performance. I used the state-of-the-art anti-spoofing system AASIST to conduct experiments, which reports 0.83% EER on the ASVspoof2019LA dataset [3], one of the largest and most used datasets in the speech anti-spoofing field [16].

To form a unified batch, I randomly select a 5-second consecutive audio snippet from the utterance if the audio is longer, pad the audio repeatedly until it reaches 5 seconds if the audio is shorter.

2.3 Training

The anti-spoofing system is trained on the FastSpeech2 + Multiband MelGAN attack from ESPnet2-TTS and the YourTTS attack from Coqui-TTS in two separate trial settings. FastPitch + Griffin-Lim is chosen as the validation set for both situations, since Griffin-Lim is not a neural-network-based vocoder and could therefore serve as a middle ground in terms of attack-specific artifacts.

To prevent overfitting the model by overtraining it, I trained for 50 epochs and evaluated using the checkpoint with the lowest EER on the validation set. With a batch size of 36, all training is performed on a single NVIDIA GeForce RTX 3090 GPU. Training script, along with all necessary configuration files, are submitted together with this paper.

2.4 Data Perturbation

To mitigate the overfitting problem, I examine several commonly found data perturbation techniques as listed below.

No perturbation. No alteration is done to the audio samples. All audio signals are normalized before sending into the anti-spoofing model.

Adding white noise. A 0.1% amplitude of white noise (0.0001 on a -1 to 1 scale) is applied to the audio signal. The audio signal is then normalized.

Convolving with reverb. All signals are convolved using the "Small Drum Room" impulse response from the Voxengo Reverb Impulse Responses collection [13]. After adding reverb, the audio signal is then normalized.

Filtering with a high-pass filter. Motivated by previous literature that found higher frequencies to have a higher concentration of artifacts [17], a high pass biquad filter is applied to all audio signals, with the cutoff frequency at 6000 Hz, and Q-value at 0.707. The audio is then normalized.

Examples of these perturbations can be found in the GitHub repository; to better visualize the results, spectrograms for all four scenarios are generated, as depicted in Figure 1. There's no clear difference between Noise and Original scenarios since the added noise is barely audible for human beings to hear and is not strong enough to form a visually distinctive difference on the spectrogram.

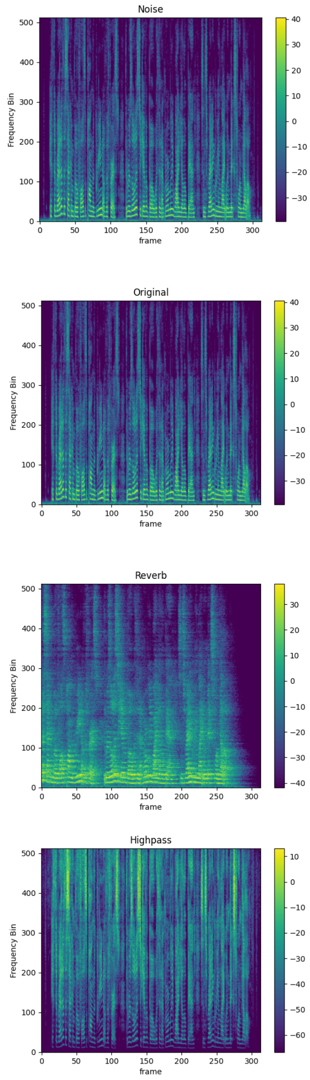

Figure 1. Spectrograms of original and modified utterances; using bona-fide utterance p333_023_mic1 as example. Spectrograms are extracted with 1024-point FFT.

3. EXPERIMENT SETUP AND RESULTS

For evaluation purposes, I computed the EER on both ESPnet2-TTS and Coqui-TTS produced VITS attacks. I calculate the average EER and the absolute EER difference between these two attacks. Less overfitting to attack-specific artifacts would be indicated by a lower difference in EER; a lower EER average would indicate better generalization ability. Since the anti-spoofing model randomly collects 5-second snippets, three evaluation results are averaged for each EER measurement. No data perturbation is present during evaluation. The outcomes are depicted in Table 2 and Table 3.

Perb.	**Framework**	**EER**	**Avg.**	**Diff.**
None	ESPnet-TTS	0.86%	16.92%	32.11%
	Coqui-TTS	32.97%		
Noise	ESPnet-TTS	1.76%	3.70%	3.87%
	Coqui-TTS	5.63%		
Reverb	ESPnet-TTS	6.72%	20.62%	27.80%
	Coqui-TTS	34.52%		
Filter	ESPnet-TTS	13.50%	15.90%	4.79%
	Coqui-TTS	18.29%		

Table 2. Evaluation results from model trained on ESPnet2-TTS attack.

Perb.	**Framework**	**EER**	**Avg.**	**Diff.**
None	ESPnet-TTS	14.14%	8.51%	11.27%
	Coqui-TTS	2.87%		
Noise	ESPnet-TTS	1.47%	3.64%	4.33%
	Coqui-TTS	5.80%		
Reverb	ESPnet-TTS	20.78%	11.69%	18.19%
	Coqui-TTS	4.10%		
Filter	ESPnet-TTS	13.17%	10.40%	5.54%
	Coqui-TTS	7.63%		

Table 3. Evaluation results from model trained on Coqui-TTS attack.

4. DISCUSSIONS

To better understand the revelation behind these perturbation scenarios, I selected the following pairs of perturbation scenarios for closer examination:

Non-perturbated scenarios. When only comparing the scenarios without any perturbation, it becomes clear that ESPnet-TTS trained model performs much better against ESPnet-TTS attack, whereas Coqui-TTS trained model performs significantly better against Coqui-TTS attack. This indicates that attack-specific artifacts can lead an anti-spoofing model to overfit and perform poorly on unknown data.

Noise-added versus non-perturbated. They reveal significant improvements in the model's ability to generalize, as indicated by the significantly lower average EER for both attacks on both models. The absolute difference in EER is also much smaller, indicating that the attack-specific artifacts are mitigated by the addition of white noise.

Reverb-added versus non-perturbated. Convolving with room impulse response decreases generalization ability, which could be explained by how reverb drastically impacted the overall spectra of audio. However, it seems that the attack-specific artifacts may not be mitigated in all scenarios.

Filter-added versus non-perturbated. A smaller difference in EER can be noticed in both scenarios, indicating that a high-frequency filter can help mitigate the addition of white noise. The average EER is roughly on the same level compared to the non-perturbated ones in both scenarios as well, showing no clear performance difference.

Noise-added versus Filter-added. These two sets of scenarios have relatively similar difference in EER, but the performance of noise-added scenarios is much better. This may be due to the noise-added samples preserving more information, while the filtered utterances are lacking in low-frequency information, which is rich in speech-related information.

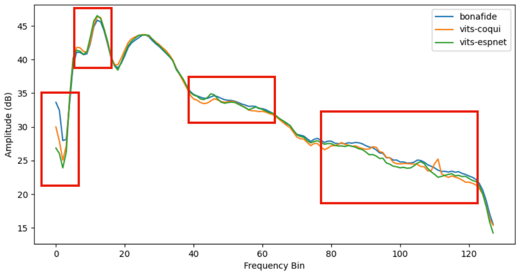

Figure 3. Average frequency energy curve from bonafide, VITS (Coqui-TTS generated, labeled as vits-coqui) and VITS (ESPnet-TTS generated, labeled as vits-espnet). Discarding non-speech frames using a naïve voice activity detection algorithm based on energy per frame. Red boxes denote areas where three frequency energy curves differ.

A further investigation with the average frequency energy curve for each attack shows the average energy different in different frequency domain. As we could see from Figure 2, although many differences do concentrate on higher frequency ranges, there are some differences in middle and low frequencies. This led to the hypothesis that anti-spoofing models rely on these artifacts to tell attacks apart, and by filtering out low frequencies, the model has less artifacts to identify, leading to worse overall performance. At the same time, since there's less model-specific artifacts as well, the anti-spoofing model is less likely to overfit.

5. CONCLUSIONS

This paper illustrates the influence of attack-specific artifacts on speech anti-spoofing systems through the creation of a novel, large-scale dataset. In addition, I demonstrate that by introducing a little amount of white noise, both the overfitting to model problem and the anti-spoofing model's capacity to generalize can be significantly mitigated and enhanced, respectively. Future work will involve the employment of new techniques to anti-spoofing models, such as representation learning, that are more independent towards specific model artifacts.

6. ACKNOWLEDGEMENTS

This study would not have been feasible without the discussions with You (Neil) Zhang and Ge Zhu of the University of Rochester's AIR Lab. Specifically, thank Neil for his assistance in adding to the paper's background section, and providing the general training codebase for AASIST. I'd also like to recognize Meiying (Melissa) Chen's assistance in providing insights about vocoder artifacts and AIR Lab's assistance in providing GPU resources to develop and generate attack samples.

7. REFERENCES

[1] J. Jung et al., “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks.” arXiv, Oct. 04, 2021. doi: 10.48550/arXiv.2110.01200.

[2] X. Yan et al., “An Initial Investigation for Detecting Vocoder Fingerprints of Fake Audio,” in Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Oct. 2022, pp. 61–68. doi: 10.1145/3552466.3556525.

[3] X. Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech & Language, vol. 64, p. 101114, Nov. 2020, doi: 10.1016/j.csl.2020.101114.

[4] M. Joslin and S. Hao, “Attributing and Detecting Fake Images Generated by Known GANs,” in 2020 IEEE Security and Privacy Workshops (SPW), May 2020, pp. 8–14. doi: 10.1109/SPW50608.2020.00019.

[5] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.” arXiv, Jun. 10, 2021. Accessed: Oct. 26, 2022. [Online]. Available: http://arxiv.org/abs/2106.06103

[6] G. Eren and The Coqui TTS Team, “Coqui TTS.” Jan. 2021. doi: 10.5281/zenodo.6334862.

[7] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/~idea/readings/rainbow.htm)., Nov. 2019, doi: 10.7488/ds/2645.

[8] Y. Ding, Y. Ding, and N. Thakur, “Does a GAN leave distinct model-specific fingerprints?,” p. 13.

[9] Anonymous, “How Distinguishable Are Vocoder Models? Analyzing Vocoder Fingerprints for Fake Audio,” in Submitted to The Eleventh International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=cCjxF2QB-AT

[10] T. Hayashi et al., “ESPnet2-TTS: Extending the Edge of TTS Research.” arXiv, Oct. 14, 2021. Accessed: Jul. 01, 2022. [Online]. Available: http://arxiv.org/abs/2110.07840

[11] A. Łancucki, “FASTPITCH: PARALLEL TEXT-TO-SPEECH WITH PITCH PREDICTION,” p. 5.

[12] Y. Ren et al., “FASTSPEECH 2: FAST AND HIGH-QUALITY END-TO- END TEXT TO SPEECH,” p. 15.

[13] “Free Reverb Impulse Responses,” Voxengo. https://www.voxengo.com/impulses/ (accessed Dec. 05, 2022).

[14] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, “Making DeepFakes more spurious: evading deep face forgery detection via trace removal attack.” arXiv, Mar. 21, 2022. Accessed: Nov. 17, 2022. [Online]. Available: http://arxiv.org/abs/2203.11433

[15] V. Wesselkamp, K. Rieck, D. Arp, and E. Quiring, “Misleading Deep-Fake Detection with GAN Fingerprints.” arXiv, May 25, 2022. Accessed: Nov. 15, 2022. [Online]. Available: http://arxiv.org/abs/2205.12543

[16] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, “Voice Spoofing Countermeasures: Taxonomy, State-of-the-art, experimental analysis of generalizability, open challenges, and the way forward.” arXiv, Oct. 01, 2022. Accessed: Oct. 09, 2022. [Online]. Available: http://arxiv.org/abs/2210.00417

[17] J. Frank and L. Schönherr, “WaveFake: A Data Set to Facilitate Audio Deepfake Detection,” arXiv:2111.02813 [cs, eess], Nov. 2021, Accessed: May 04, 2022. [Online]. Available: http://arxiv.org/abs/2111.02813

[18] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone.” arXiv, Feb. 16, 2022. doi: 10.48550/arXiv.2112.02418.