An Initial Investigation of the attack-specific artifacts overfitting issue in speech anti-spoofing model

Yongyi Zang

Deepfake cause issues

Attackers use popular
Text-to-speech (TTS) and
Voice conversion (VC) toolboxes,

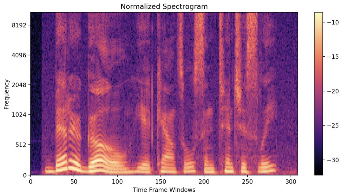
like **ESPnet** and **Coqui,**

which implements a lot of popular
TTS and VC algorithms.

Deepfake anti-spoofing systems

Visual

Image Deepfake Detection Systems
Detect **artifacts** in computer-generated image



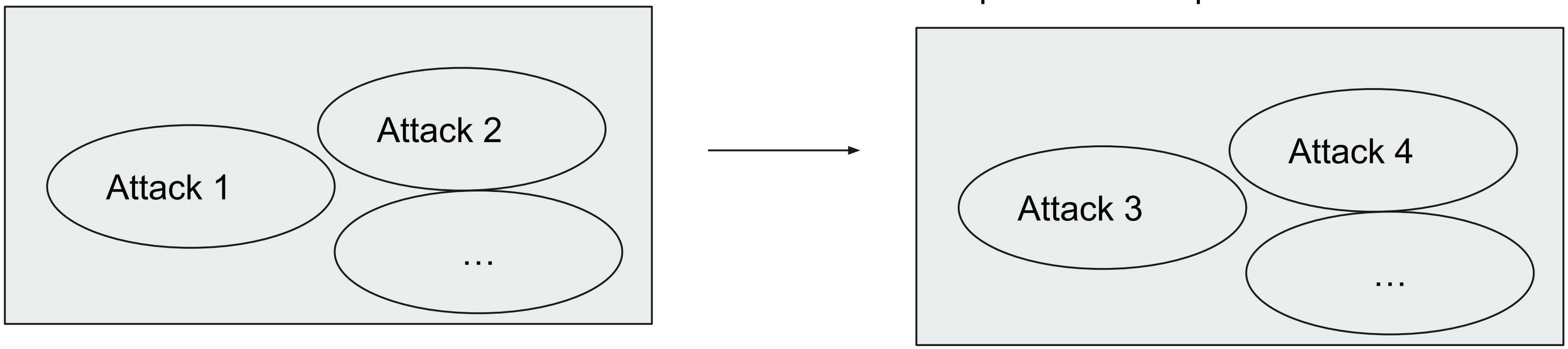
Audio

Speech Anti-spoofing Systems (or countermeasures, CM)
Detect **artifacts** in computer-generated speech

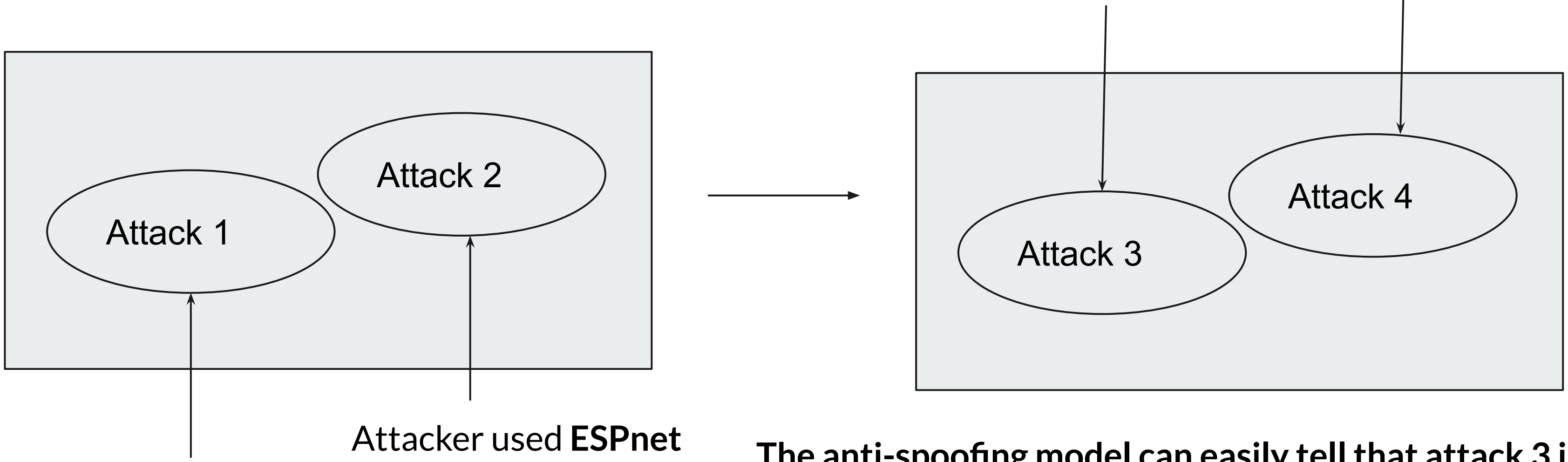
The generalization problem

We train on some attacks

Hope it can also spot out **other unseen attacks**



The generalization problem



Attack 1

Attack 2

Attacker used **ESPnet**

Attacker used **ESPnet**

Attack 3

Attack 4

Attacker used **ESPnet**

Attacker used **Coqui-TTS**

The anti-spoofing model can easily tell that attack 3 is fake,
But can't easily tell attack 4 is fake.

It overfitted on **ESPnet-specific artifacts.**

Training Setup

Used AASIST - the **SOTA** speech anti-spoofing model. (EER = 0.83% on ASVspoof2019LA)

Trained on...

ESPnet attack
FastSpeech2 TTS + Mel-GAN

Coqui attack
YourTTS

Validated on...

Fastpitch + Griffin-Lim

Training Setup

Then evaluate both	**on**
ESPnet-trained	**ESPnet-attack** VITS
Coqui-trained	**Coqui-attack** VITS

The problem does exist

ESPnet trained

Framework	EER
ESPnet-TTS	0.86%
Coqui-TTS	32.97%

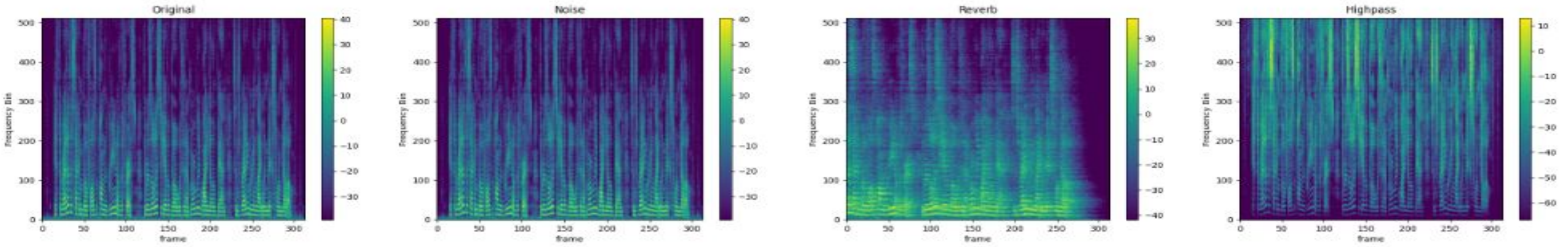
Performs better on ESPnet

Coqui trained

Framework	EER
ESPnet-TTS	14.14%
Coqui-TTS	2.87%

Performs better on Coqui

How do we mitigate it?



Inaudible **noise**
(0.1% amplitude)

Convolved with
reverb

Highpass Biquad
cutoff at 6 kHz
Q = 0.707

How do we mitigate it?

Noise

- Destroy amplitude slightly, destroy phase
- Spectra is preserved

Reverb

- Destroy amplitude and phase massively
- Spectra is not preserved

Filter

- Destroy amplitude, preserve some phase
- Spectra is somewhat preserved

Metrics

Performance: How well is the anti-spoofing model in telling fake speech apart from real ones?

- **Average EER (Avg.)**
- (ESPnet_attack_EER + Coqui_attack_EER) / 2

Overfitting: Does the anti-spoofing model still exhibit overfitting behavior?

- **Absolute Difference in EER (Diff.)**
- abs(ESPnet_attack_EER - Coqui_attack_EER)

Noise works

ESPnet trained

Coqui trained

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%		
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 doesn't work

ESPnet trained

Coqui trained

Perb.	Framework	EER	Avg.	Diff.
None	ESPnet-TTS	0.86%	16.92%	32.11%
	Coqui-TTS	32.97%		
Reverb	ESPnet-TTS	6.72%	**20.62%**	**27.80%**
	Coqui-TTS	34.52%		
None	ESPnet-TTS	14.14%	8.51%	11.27%
	Coqui-TTS	2.87%		
Reverb	ESPnet-TTS	20.78%	**11.69%**	**18.19%**
	Coqui-TTS	4.10%		

Highpass works

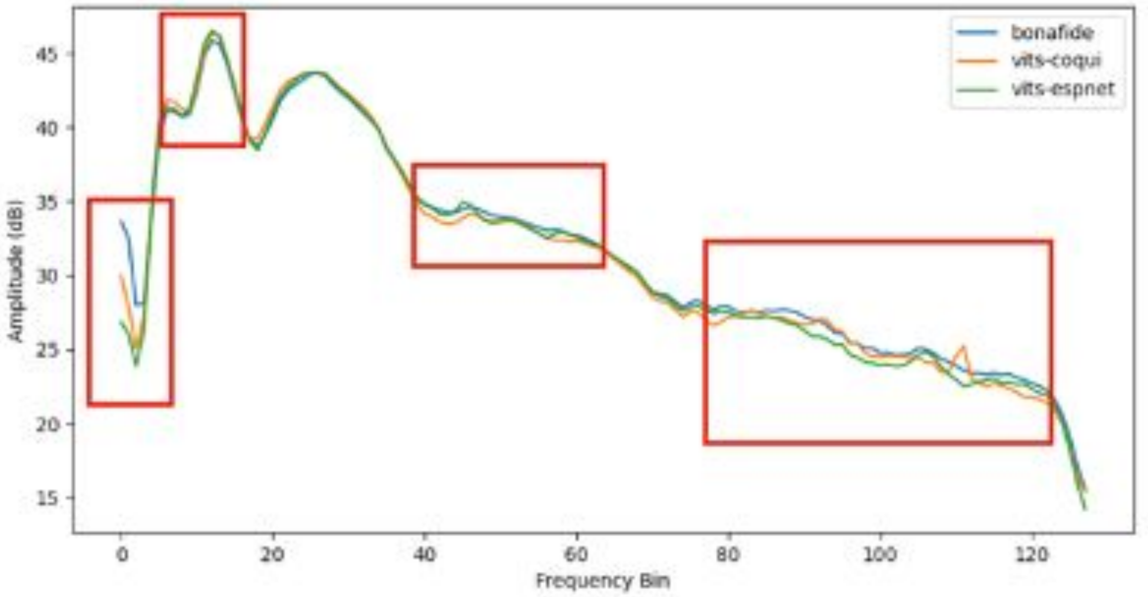
	Perb.	Framework	EER	Avg.	Diff.
ESPnet trained	None	ESPnet-TTS	0.86%	16.92%	32.11%
		Coqui-TTS	32.97%		
	Filter	ESPnet-TTS	13.50%	**15.90%**	**4.79%**
		Coqui-TTS	18.29%		
Coqui trained	None	ESPnet-TTS	14.14%	8.51%	11.27%
		Coqui-TTS	2.87%		
	Filter	ESPnet-TTS	13.17%	**10.40%**	**5.54%**
		Coqui-TTS	7.63%		

Noise works, Reverb doesn’t work, Filter works. Why?

It’s possible that…

- **Spectra** should be **preserved**
- **Frequency** with artifacts should be **distorted**
- **Phase** should be **destroyed**

Which frequencies are rich with artifacts?



Future work

- Further investigation of the **frequency artifacts and phase artifacts**
    - Bandpass to see which frequency band is most rich with artifacts
- **Representation learning** to make the speech anti-spoofing model **immune to model-specific artifacts**

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.