Spoofing Attack Augmentation: Can Differently-Trained Attack Models Improve Generalisation?

Wanying Ge [1], *Xin Wang* [2], *Junichi Yamagishi* [2], *Massimiliano Todisco* [1], *Nicholas Evans* [1]

[1]**EURECOM**, France [2]**National Institute of Informatics**, Japan

EU H2020 - GA-860813

Introduction

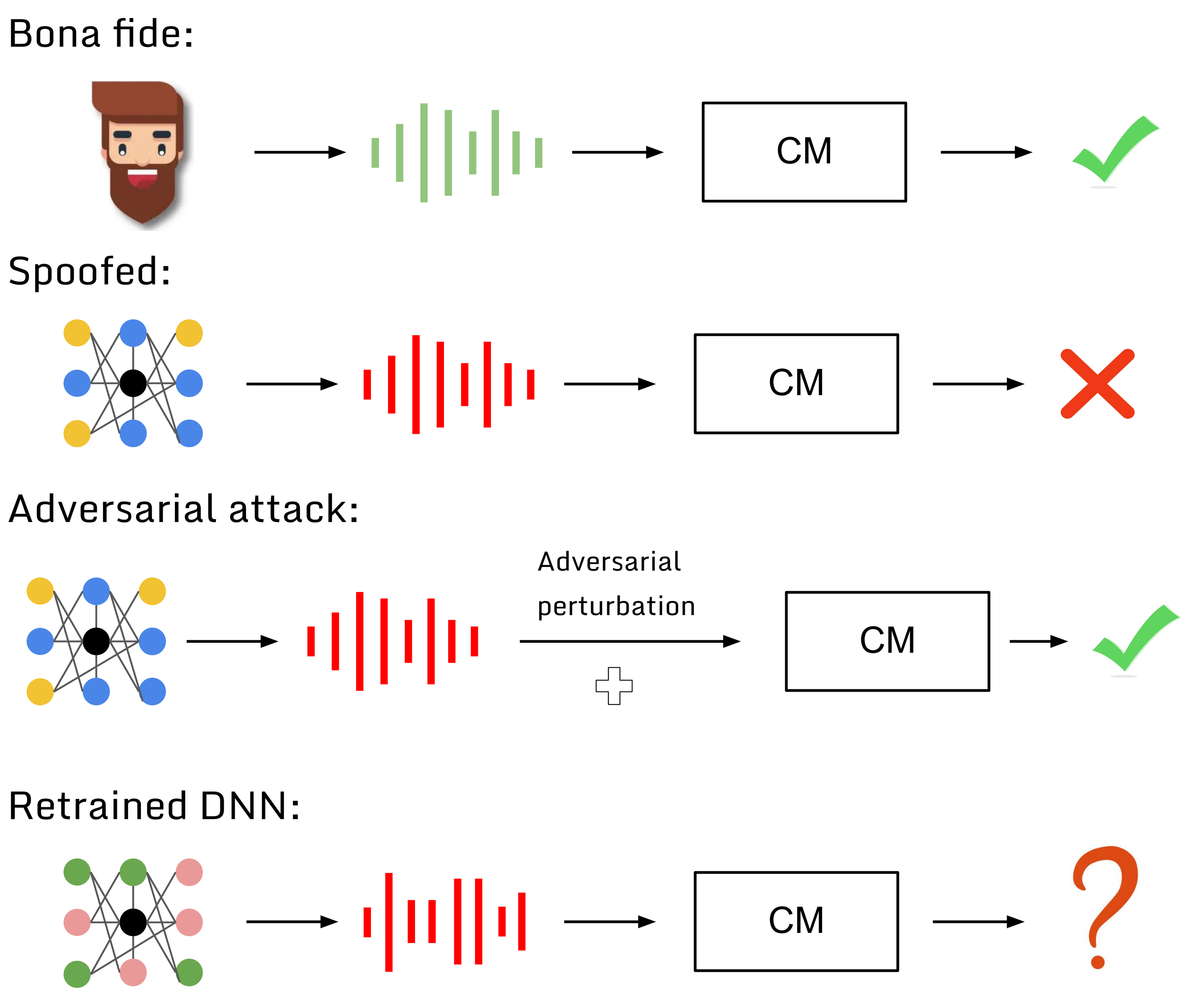
Motivation

Performance and behaviour of many deep-learning-based voice deepfake/spoofing ***detection*** models vary when retrained [1]. It is possible that the same could be true for deep-learning-based deepfake/spoofing ***generation*** models, potentially to an extent that own countermeasures (CMs) might fail to detect them.

What we do:

- train CM with spoofed data generated from a text-to-speech (TTS) attack;
- retrain the TTS attack as adversarial attacks to a fixed CM;
- propose spoofing attack augmentation to improve CM generaliability.

Adversarial attack to spoofing detection



Spoofing attack models

Attack algorithm : Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS). [2]

CM training data : 3000 bonafide utterances from VCTK [3] and 3000 spoofed utterances generated by VITS models listed below:

Table 1: VITS training and generation settings across different sets ('-' indicates identical settings to V1).

	Training			Noise std. in generation	
Set ID	Train set	#. Mel chan.	Seed	For acoustic feat.	For duration
V1	set-1	80	seed-1	0.667	0.8
V2	-	40	-	-	-
V3	set-2	-	-	-	-
V4	-	-	seed-2	-	-
V1.2	same VITS model as V1			-	-
V1.3	same VITS model as V1			0.1	-
V1.4	same VITS model as V1			-	0.1
V1.5	same VITS model as V1			0.1	0.1

- V1 : Basic model trained with default conditions.
- V2 - V4 : Models trained each with one condition different to V1.
- V1.2 - V1.5 : Same V1 model, different generation conditions.

Experiments and results

Table 2: CM performance in terms of the EER (%) in different training and testing conditions.

	Trained on V1			Trained on V2			Trained on V3			Trained on V4		
Tested on	AASIST	RawNet2	SSL-AASIST	AASIST	RawNet2	SSL-AASIST	AASIST	RawNet2	SSL-AASIST	AASIST	RawNet2	SSL-AASIST
V1	0	0	0	0	13.27	0.04	0.03	6.17	1.37	0.27	12.60	0.57
V2	0.50	6.27	0.07	0	0.03	0	0.67	8.70	0.47	0.67	11.23	0.13
V3	2.43	8.50	0.03	2.20	18.00	0.10	0	0	0	1.73	10.60	0.07
V4	1.20	7.93	0	0.57	15.93	0.07	0.13	5.87	0.13	0	0.13	0
V1.2	0	0.67	0	0	13.03	0.30	0	5.47	1.40	0.23	12.47	0.60
V1.3	0	0.03	0	0	7.20	0.57	0	2.00	2.03	0.07	6.2	1.03
V1.4	0	0.93	0	0.03	12.63	0.33	0.03	7.27	1.80	0.33	15.03	1.07
V1.5	0	0.10	0	0	6.63	0.83	0.03	2.10	2.80	0.10	7.80	1.33
Pooled	0.77	3.73	0.01	0.50	11.49	0.37	0.16	5.03	1.50	0.57	10.11	0.63

- Matched training and testing conditions result in zero or near-zero equal error rate (EER) for all CM systems.
- EER increases under mismatched conditions; however, AASIST and SSL-AASIST systems are relatively more robust across different synthetic data.
- RawNet2 shows substantially higher EERs under mismatched conditions.

Table 3: Performance in terms of the EER (%) for CMs trained on combined sets V2-V4 and tested against unseen V1 and V1.2-V1.5 attacks.

	Trained on V2-4		
Tested on	AASIST	RawNet2	SSL-AASIST
V1	0	2.2	0
V2	0	2.93	0
V3	0	0.47	0
V4	0	1.37	0
V1.2	0	1.9	0
V1.3	0	0.77	0.03
V1.4	0	2.83	0
V1.5	0	0.87	0.03
Pooled	0	1.79	0.01

- Training CMs with spoofed data from multiple, differently configured attack algorithms improves generalisation to spoofing attacks.
- RawNet2 shows higher variability in EER across different attack configurations

References

[1] X.Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in Proc. INTERSPEECH 2021.

[2] J. Kim. J. Kong et al., "Conditional variational autoencoder with adversarial learning for end-to-end textto-speech," in International Conference on Machine Learning, 2021.

[3] J. Yamagishi, C. Veaux et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research, 2019.

Conclusions

- Spoofing countermeasures trained on data generated with one attack configuration are vulnerable to variations of the same algorithm.
- Training CMs with spoofed utterances from multiple, differently configured attack algorithms significantly improves generalisation.
- Future research should extend the evaluation of current CMs to other attack algorithms and explore the benefits of **spoofing attack augmentation** in improving generalisation to entirely different attacks.