Ensemble Models for Spoofing Detection in Automatic Speaker Verification

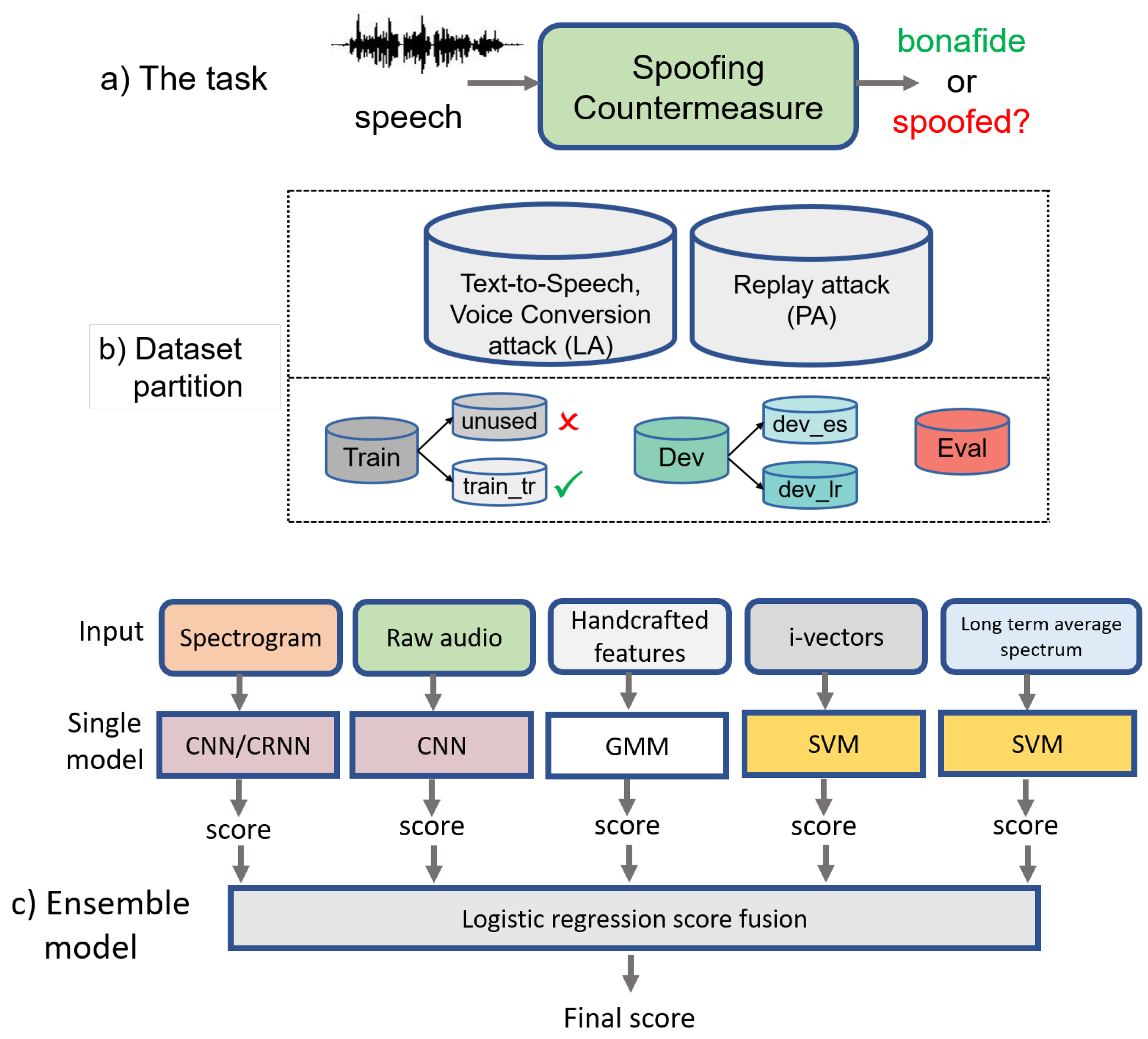
Bhusan Chettri, Daniel Stoller, Veronica Morfi, Marco A. Martínez Ramírez, Emmanouil Benetos, Bob L. Sturm

b.chettri@qmul.ac.uk

1 Introduction

- We explore ensemble models for spoofing detection on the ASVspoof 2019 logical access (LA) and physical access (PA) datasets [1].
- We find models appear to have improved generalisation when we partition those datasets to ensure disjoint attack conditions [2].
- We examine why some models work so well and find they are using specific irrelevant cues in the recordings.

2 Tasks and Model description



3 Experimental results

- Metric: tandem-DCF (t-DCF) [3] and equal error rate (EER)
- LFCC GMM (B1) and CQCC-GMM (B2) are official baselines

Model	Set	LA attack		PA attack	
		t-DCF	EER%	t-DCF	EER%
B1	Dev	0.0663	2.71	0.2554	11.96
B2		0.0123	0.43	0.1953	9.87
ensemble		**0.0**	**0.0**	**0.0354**	**1.33**
B1	Eval	0.2116	8.09	0.3017	13.54
B2		0.2366	9.57	0.2454	11.04
ensemble		**0.0755**	**2.64**	**0.1492**	**6.11**

4 What is the CNN exploiting in the PA dataset?

- We find that a CNN performs much better when trained on the last 4 seconds of every recording than on the first 4 seconds.
- We find this comes from silent segments in the spoof recordings.

Intervention I: remove silence from the end at test time.

Model	t-DCF	EER %
B1	$0.2036 \to 0.2741$	$9.18 \to 13.27$
B2	$0.1971 \to 0.2959$	$10.06 \to 15.59$
CNN	$0.1672 \to 0.5018$	$5.98 \to 19.8$

Intervention II: train the models removing silence from the end.

Model	t-DCF	EER %
B1	$0.2036 \to 0.9528$	$9.18 \to 54.76$
B2	$0.1971 \to 0.9463$	$10.06 \to 57.98$
CNN	$0.1672 \to 0.2626$	$5.98 \to 11.20$

Intervention III: remove silence during both training and testing.

Model	t-DCF	EER %
B1	$0.2036 \to 0.8614$	$9.18 \to 41.09$
B2	$0.1971 \to 0.9448$	$10.06 \to 58.71$
CNN	$0.1672 \to 0.3129$	$5.98 \to 12.85$

How about the evaluation set?

- Models show similar behaviour under above interventions.

5 Conclusion

- We find ensemble models are better than the baselines in detecting unseen spoofing attacks, yielding 3^{rd} rank in the LA task.
- We find their performance on the PA task is inflated due to a cue (existence of silence) in the recordings of the dataset [4].
- We propose removing this cue in the PA dataset [5] for more reliable estimate of performance.

[1] Massimiliano et. al. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Proc. Interspeech*, 2019.

[2] Partition details: https://github.com/BhusanChettri/ASVspoof2019/.

[3] Kinnunen et. al. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. In *Proc. Speaker Odyssey*, 2018.

[4] B. L. Sturm. A Simple Method to Determine if a Music Information Retrieval System is a "Horse". In *IEEE Transactions on Multimedia*, 2014.

[5] B. Chettri and B. L. Sturm. A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems. In *IEEE ICASSP*, 2018.