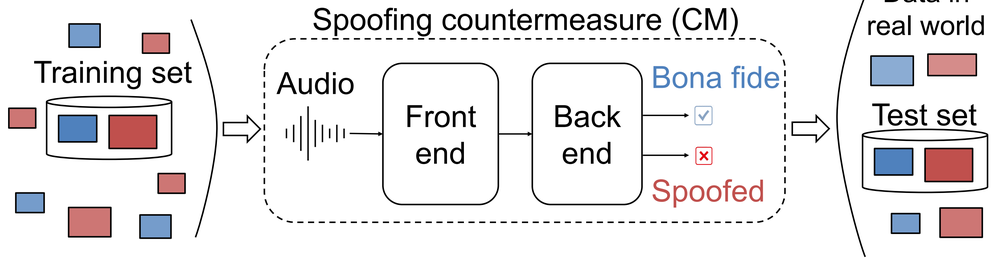
Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders

Xin Wang, Junichi Yamagishi
National Institute of Informatics

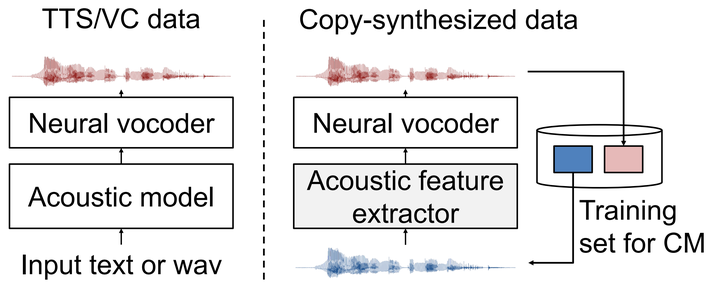
ICASSP2023
Paper ID: 1638

Introduction

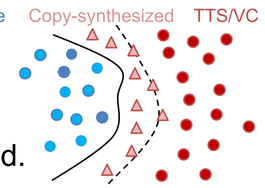


- A common practice: a training and a test set from a standard database
  - Can such a CM generalize? Not sufficiently good [1,2,3,4]
- More training data? Building diverse text-to-speech (TTS) and voice conversion (VC) is time-consuming: 6 months for ASVspoof 2019.
- *A more efficient way of creating spoofed training data?*

Copy-synthesis (CS)



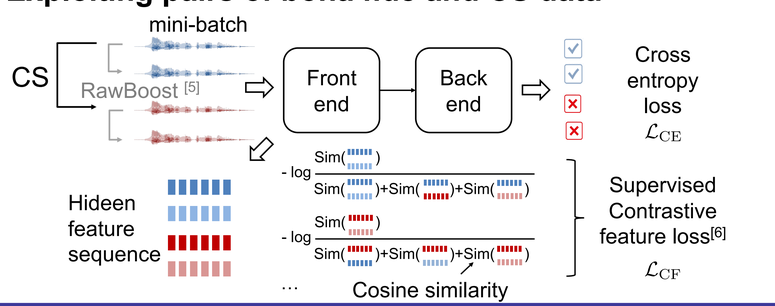
- CS is TTS/VC with a "perfect" acoustic model.
- CS data contains artifacts caused by vocoders.
- Vocoders are easier to train and fast to generate.



Assumptions:

- CS data is closer to the bona fide.
- A tight decision boundary is learned.

Exploiting pairs of bona fide and CS data



Experiments & findings

CM configuration following our previous work [3]

- Front end: wav2vec 2.0 (XLSR), fine-tuned with the back end
- Back end: linear layer, global average pooling, vanilla softmax

Test sets: ASVspoof test sets, WaveFake [8], In-the-Wild [9]

Training sets:

- ASVspoof 2019 logical access training set (`LA19trn`)
- WaveFake subset (`WFtrn`)
- Four training sets with `LA19trn` bona fide & CS spoofed data
  - No autoregressive (AR) or DSP vocoder is used

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Vocoder train/fine-tune data
`LA19trn`	20	2,580	22,800	full-fledged TTS/VC	-
`WFtrn`	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	LJSpeech / -
`Voc.v1`	20 same as `LA19trn`	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	LibriTTS / -
`Voc.v2`				HiFiGAN, Hn-NSF, NSF-HiFiGAN, WaveGlow	LibriTTS / -
`Voc.v3`				HiFiGAN, Hn-NSF, NSF-HiFiGAN, WaveGlow	`LA19trn` bona. / -
`Voc.v4`				HiFiGAN, Hn-NSF, NSF-HiFiGAN, WaveGlow	LibriTTS / `LA19trn` bona.

Questions: Is there any recommended way to train neural vocoders? Is the supervised contrastive feature loss useful?

[1] Paul, et al., A Case Study with ASVspoof 2015 and BTAS 2016 Corpora. In Proc. ICASSP, 2047–2051, 2017
[2] Das, et al., Assessing the Scope of Generalized Countermeasures for Anti-Spoofing. In Proc. ICASSP, 6589–6593. 2020.
[3] Wang, X., et al., Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In Odyssey, 2022
[4] Liu, X., et.al., ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. ArXiv 2022.
[5] Tak, H., et. al., RawBoost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In Proc. ICASSP, 6382–6386. 2022.
[6] Khosla. P, et.al., Supervised Contrastive Learning. In Proc. NIPS, 18661–18673. 2020.
[7] Frank. J., et al.. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In Proc. NeurIPS Datasets and Benchmarks 2021.
[8] Müller. N.,,et. al.,. Does Audio Deepfake Detection Generalize? In Proc. Interspeech, 2783–2787. 2022.
[9] Jung. J., et. al.,. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In Proc. ICASSP, 6367–6371. 2022.

Results Averaged over three training & evaluation runs

Equal error rates (EER, %)

Test sets		`LA19 trn`	`WF trn`	`Voc. v1`	`Voc. v2`	`Voc. v3`	`Voc. v4`
ASVspoof 2019 test set →	`LA19eval`	2.98	44.48	5.78	5.32	8.74	4.36
2021 LA eval track →	`LA21eval`	7.53	41.57	26.30	17.98	19.29	24.39
2021 DF eval track →	`DF21eval`	6.67	24.26	11.95	11.54	9.71	13.31
LA19 silence trimmed →	`LA19etrim`	15.56	31.62	23.29	16.16	14.99	9.52
2021 LA hidden track →	`LA21hid`	28.80	27.60	28.30	19.49	17.62	21.43
2021 DF hidden track →	`DF21hid`	23.62	26.18	22.01	13.92	13.50	16.99
WaveFake (all) →	`WaveFake`	15.76	-	39.27	34.05	17.10	10.89
In-the-wild →	`InWild`	26.65	19.98	41.06	36.46	22.26	19.45
	`Pooled`	14.24	-	36.57	39.95	19.39	16.35

Training criterion	$\mathcal{L}_{CE}$				$\mathcal{L}_{CE} + \mathcal{L}_{CF}$		
Data augmentation	×		RawBoost		RawBoost		
Training set	`LA19 trn`	`Voc. v4`	`LA19 trn`	`Voc. v4`	`LA19 trn`	`Voc. v4`	`Voc. v4`
Bona-spoof paired	×	×	×	×	×	×	✓
ID	①	②	③	④	⑤	⑥	⑦
`LA19eval`	2.98	4.36	0.22	3.46	0.21	2.63	2.21
`LA21eval`	7.53	24.39	3.63	16.55	3.30	16.67	17.90
`DF21eval`	6.67	13.31	3.65	9.60	4.12	6.92	5.04
`LA19etrim`	15.56	9.52	9.16	6.09	9.00	4.48	3.79
`LA21hid`	28.80	21.43	21.18	19.37	26.98	15.05	14.57
`DF21hid`	23.62	16.99	13.64	14.29	16.85	8.17	7.78
`WaveFake`	15.76	10.89	26.37	6.87	24.62	4.03	2.50
`InWild`	26.65	19.45	16.17	12.08	17.07	9.37	7.55
`Pooled`	14.24	16.35	13.12	13.13	13.68	13.15	11.27

Answers to question 1:

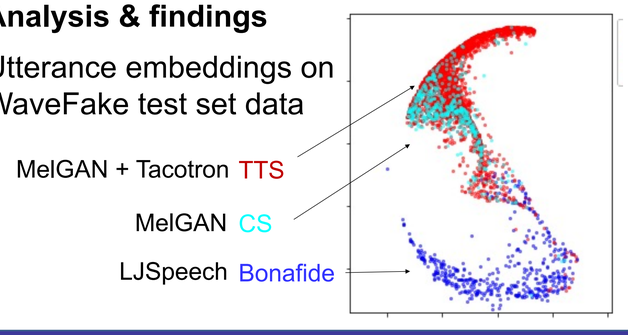
- ✓ Training neural vocoders on the bona fide data (v3 & v4)
- ✓ Pre-training + fine-tuning is better (v4)
- x Not useful to CMs LFCC-LCNN and AASIST[9] ☞ arxiv

Highlighted answers to question 2:

- ✓⑦ VS ⑥ and ④: contrastive feature loss performed best on *paired* bona fide and CS data
- ✓⑦ is competitive on most test sets ☞ paper Tab.4

Analysis & findings

Utterance embeddings on WaveFake test set data



Conclusions & future work

- Spoofed CM training data can be created by neural vocoders
- The trained CM well detected actual spoofed data from TTS/VC using unseen DSP vocoders, but not those from *unseen AR vocoders* ☞ paper Fig.1
- Larger vocoded database? VoxCeleb2!

Vocoders
CMs
Databases

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Contact: wangxin@nii.ac.jp, Code: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts project/09-asvspoof-vocoded-trn
Partially supported by JST CREAST grant JPMJCR18A6, JPMJCR20D3, JSPS KAKEN 21K17775, 21H04906, and Google AI for Japan program.