Paper 8162024

SpoofCeleb: Speech Deepfake Detection and SASV in the Wild

Jee-weon Jung†, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti,
Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Joon Son Chung,
Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, Shinji Watanabe †: currently at Apple

[1] Carnegie Mellon University, USA [2] Renmin University of China, China [3] National Institute of Informatics, Japan
[4] Korea Advanced Institute of Science and Technology, South Korea [5] Meta, USA [6] University of Tokyo, Japan
[7] EURECOM, France [8] Shanghai Jiao Tong University, China [9] Yonsei University, South Korea [10] Keio University, Japan

Overview

❖Motivation

- Existing speech deepfake detection (SDD) datasets consist of clean bona fide speech and spoofed samples
- Spoofing-robust automatic speaker verification (SASV) lacks data

❖ Goal

- Create a dataset that: (1) Focuses more on the recognition side, (2) Involves large number of speakers (3) Consists of real-world noisy bona fide data, (4) Includes spoofed samples from diverse TTS models

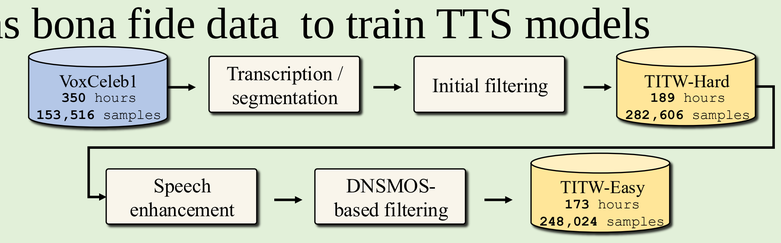
Backgrounds

❖Generation-recognition trade-off

- Generation side: text-to-speech (TTS) model training requires studio-quality data
- Recognition side: diverse noise, reverberation is crucial

❖ Source dataset: TITW (TTS in the wild)

- Derived by processing VoxCeleb through an automatic pipeline
- Used as bona fide data to train TTS models

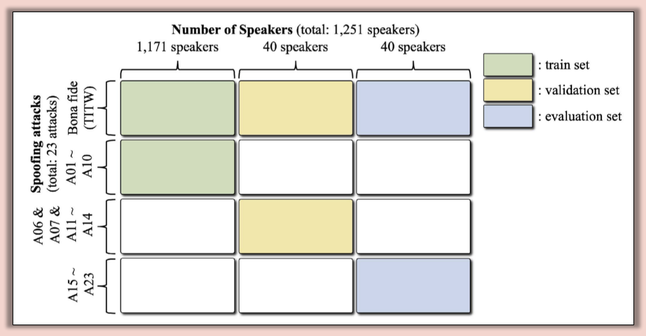


Overall Framework

❖Generating spoofing attacks

- SpoofCeleb includes spoofed speech samples synthesized via 23 TTS systems
  - Utilizes 4 acoustic models, 6 waveform models, and 5 end-to-end models including speechLMs
  - Paired acoustic and waveform models to augment the number of attacks

❖ Data partitioning



❖ Number of speech files and protocols

	# speech files	# trials in SASV protocols
Train	2,540,421	N/A
Validation	55,741	39,353
Evaluation	91,130	133,448

Number of trials equals that of speech files for SDD protocols.

❖ Spoofing attack algorithms

AttackID	Partition	Acoustic model	Waveform model
A01	trn	VITS	N/A
A02	trn	MQTTS‡	N/A
A03	trn	VALL-E	N/A
A04	trn	Delay†	N/A
A05	trn	GradTTS	DiffWave
A06	trn&dev	GradTTS	BigVGAN
A07	trn&dev	GradTTS	WaveGlow
A08	trn	MatchaTTS	DiffWave
A09	trn	MatchaTTS	BigVGAN
A10	trn	MatchaTTS	WaveGlow
A11	dev	MQTTS	N/A
A12	dev	Delay	N/A
A13	dev	GradTTS	NSF HiFiGAN
A14	dev	MatchaTTS	HiFiGAN
A15	eval	MQTTS‡,◇	N/A
A16	eval	VALL-E†	N/A
A17	eval	GradTTS	HiFiGAN
A18	eval	MatchaTTS	NSF HiFiGAN
A19	eval	Multi-scale Transformer	N/A
A20	eval	Multi-scale Transformer†	N/A
A21	eval	TransformerTTS	ParallelWaveGAN
A22	eval	BVAE-TTS	HiFiGAN
A23	eval	BVAE-TTS	NSF HiFiGAN

There are 23 different attacks stemming from 23 different TTS systems.†: pre-trained, ‡: decoder is pretrained, ◇: speaker embeddings from target utterances.

❖Metrics

- For evaluation of the data quality: SPF-EER, MCD, UTMOS, DNSMOS, WER
  - WER measured with OpenAI Whisper
- For evaluation of baselines: EER, minDCF, min a-DCF, SV-EER, SPF-EER

❖ Source dataset: TITW (TTS in the wild)

- Derived by processing VoxCeleb through an automatic pipeline
- Used as bona fide data to train TTS models

Trial type \ metric	a-DCF [66]	SV-EER	SPF-EER [13]
Target	+	+	+
Bona fide non-target	-	-	
Spoof non-target	-		-

a-DCF measures the overall performance.SV-EER measures the ability to reject non-target speakers and SPF-EER measures spoofrobustness. "+": a system should accept, "-": a system should reject.

EER: equal error rate
SPF: spoof
MCD: mean cepstral distortion
MOS: mean opinion score
WER: word error rate
DCF: detection cost function
SV: speaker verification

Spoofing attack quality

❖**Main goal**

- High SPF-EER
- Small discrepancy of quality metrics including WER and DNSMOS compared to bona fide

❖**Observations**

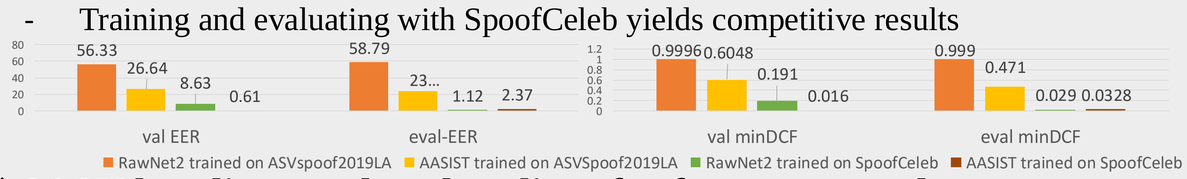
- 18 attacks have SPF-EER over 20%
- 9 attacks have higher DNSMOS than bona fide

Attack ID	Partition	SPF-EER (%)↓	MCD↓	UTMOS↑	DNSMOS↑	WER (%)↓
A00 (bona fide)	trn&val&eval	N/A	N/A	3.32	2.78	9.10
A01	trn	29.22	8.61	2.77	2.74	53.00
A02	trn	49.47	7.09	3.08	2.83	23.60
A03	trn	12.51	10.85	3.28	2.93	28.50
A04	trn	20.86	10.42	3.59	2.83	4.80
A05	trn	23.63	6.76	2.18	2.39	11.90
A06	trn&val	29.42	9.23	2.08	2.16	11.30
A07	trn&val	24.61	5.61	1.30	1.51	11.90
A08	trn	32.00	5.36	2.47	2.59	15.80
A09	trn	31.07	9.10	2.38	2.48	15.90
A10	trn	26.20	5.66	1.32	1.79	15.70
A11	val	47.78	6.99	3.08	2.83	23.30
A12	val	14.52	10.91	3.26	2.84	32.50
A13	val	27.13	5.52	1.97	2.13	12.90
A14	val	29.36	5.11	2.52	2.48	14.50
A15	eval	65.21	6.79	3.14	2.83	21.20
A16	eval	21.63	10.43	3.87	2.93	3.40
A17	eval	30.20	5.44	2.62	2.43	11.20
A18	eval	25.84	5.19	2.04	2.24	16.00
A19	eval	17.20	10.69	3.29	2.88	11.80
A20	eval	22.36	10.73	3.53	2.92	5.50
A21	eval	22.32	11.68	2.06	2.50	24.90
A22	eval	5.65	5.74	1.37	1.62	21.50
A23	eval	6.75	5.65	1.30	1.50	25.90

Baseline Results

❖**SDD baseline results**

- RawNet2 and AASIST used as baselines
- Huge domain gap exists with ASVspoof2019 dataset
- Training and evaluating with SpoofCeleb yields competitive results



❖**SASV baseline results – baselines for future research**

- Model: SKA-TDNN
- Conventional ASV has high SPF-EER but competitive SV-EER
- Again, huge domain gap exists with ASVspoof2019 dataset