SLT 2024 Challenge Session

Singing Voice Deepfake Detection

Overview and Results

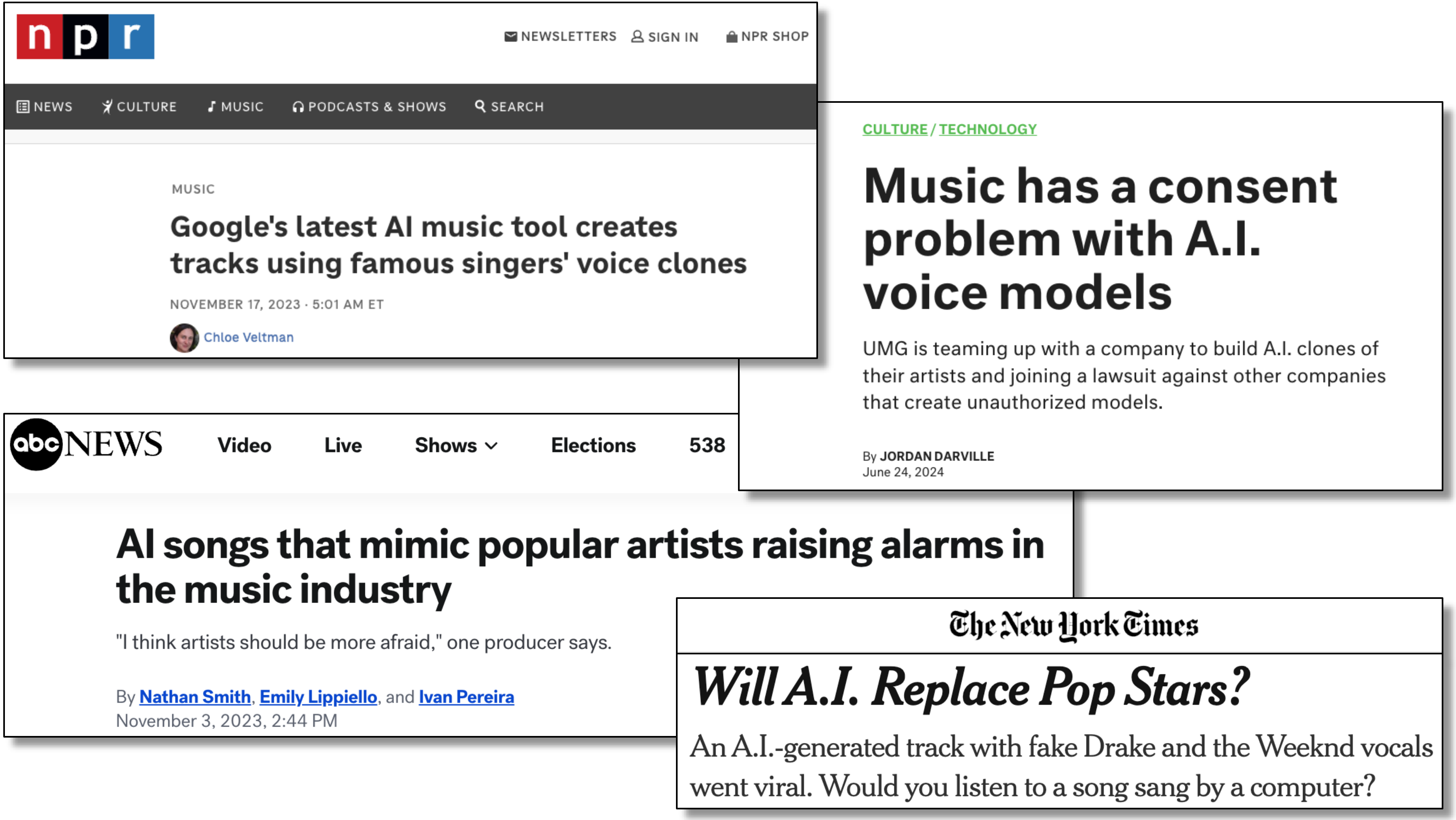
*You Zhang[1], Yongyi Zang[1], Jiatong Shi[2], Ryuichi Yamamoto[3], Tomoki Toda[3], and **Zhiyao Duan[1]***

[1] University of Rochester

[2] Carnegie Mellon University

[3] Nagoya University

Macau, China - Dec 2, 2024



MUSIC

Google's latest AI music tool creates tracks using famous singers' voice clones

NOVEMBER 17, 2023 · 5:01 AM ET

Chloe Veltman

CULTURE / TECHNOLOGY

Music has a consent problem with A.I. voice models

UMG is teaming up with a company to build A.I. clones of their artists and joining a lawsuit against other companies that create unauthorized models.

By JORDAN DARVILLE
June 24, 2024

AI songs that mimic popular artists raising alarms in the music industry

"I think artists should be more afraid," one producer says.

By Nathan Smith, Emily Lippiello, and Ivan Pereira
November 3, 2023, 2:44 PM

Will A.I. Replace Pop Stars?

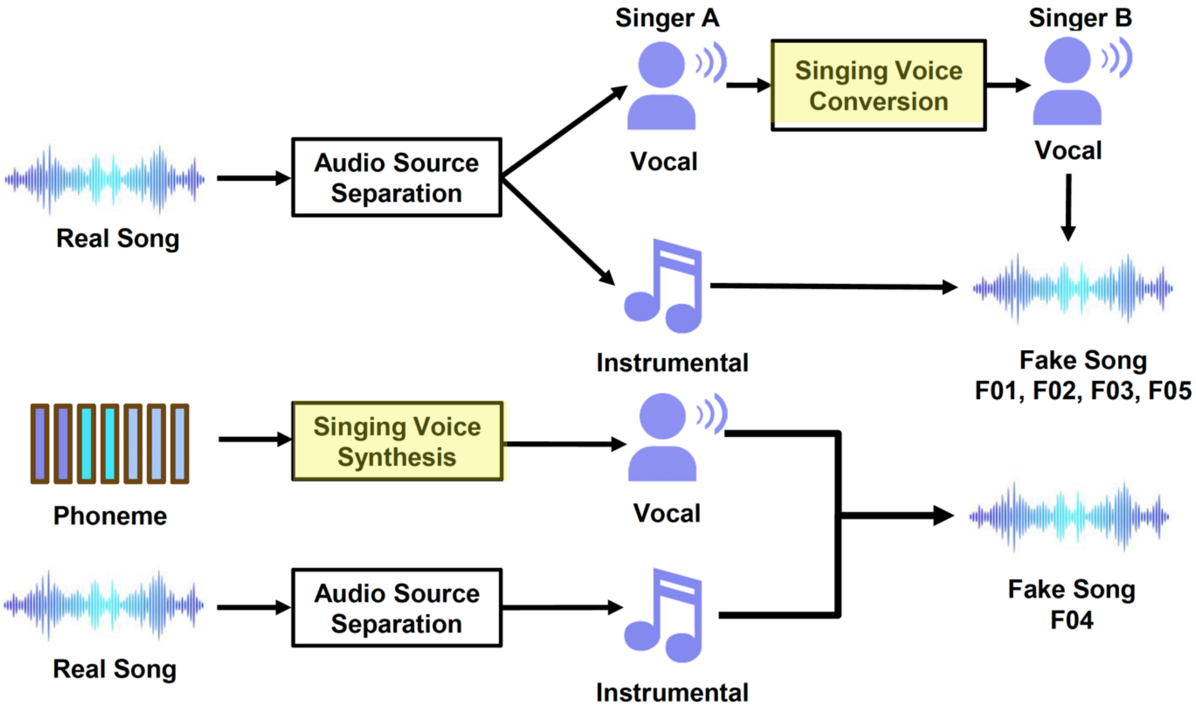
An A.I.-generated track with fake Drake and the Weeknd vocals went viral. Would you listen to a song sang by a computer?

Demo of singing voice deepfakes

https://www.youtube.com/watch?v=dHBOKfHZwL8

Life Is a Highway (Song by Rascal Flatts, Covered by AI Taylor Swift)

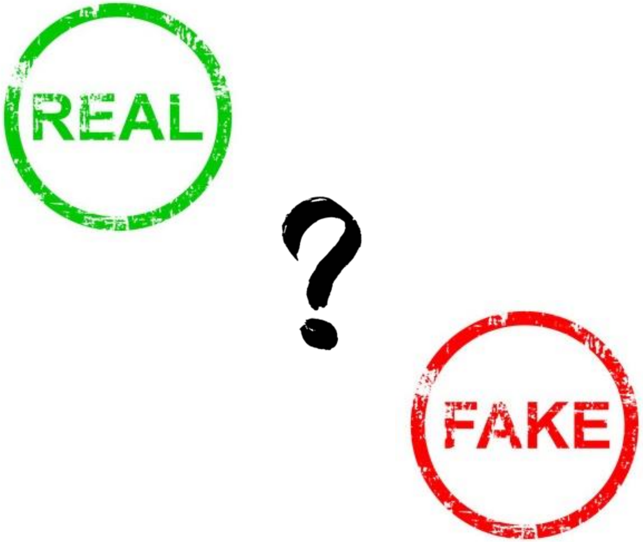
Singing voice deepfake generation



Xie et al. "FSD: An Initial Chinese Dataset for Fake Song Detection", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

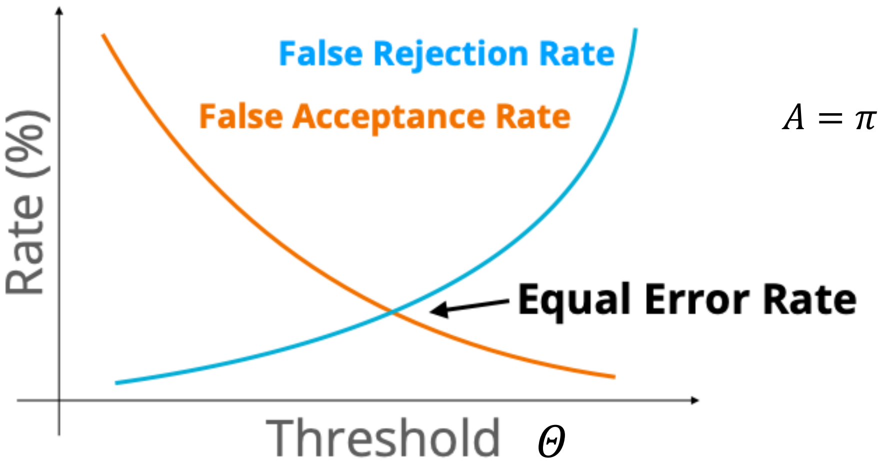
Singing Voice Deepfake Detection (SVDD)

- Aims to **detect AI-generated singing voices**



Evaluation metric

- Equal Error Rate (EER)



more fake more real

$A = \pi r^2$

$$P_{fa}(\theta) = \frac{\#\{\text{spoof trials with score} > \theta\}}{\#\{\text{total spoof trials}\}},$$

$$P_{miss}(\theta) = \frac{\#\{\text{human trials with score} \leq \theta\}}{\#\{\text{total human trials}\}}$$

$$P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER})$$

Preliminary work: SingFake

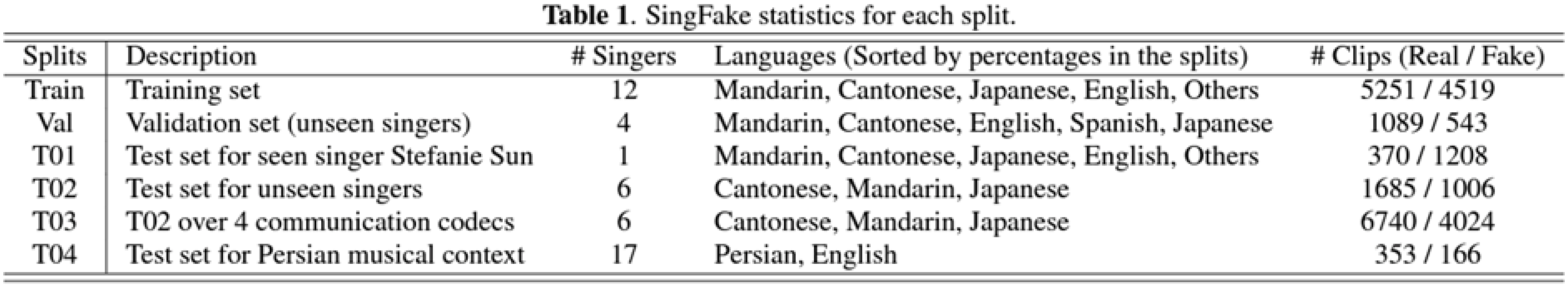


Table 1. SingFake statistics for each split.

Splits	Description	# Singers	Languages (Sorted by percentages in the splits)	# Clips (Real / Fake)
Train	Training set	12	Mandarin, Cantonese, Japanese, English, Others	5251 / 4519
Val	Validation set (unseen singers)	4	Mandarin, Cantonese, English, Spanish, Japanese	1089 / 543
T01	Test set for seen singer Stefanie Sun	1	Mandarin, Cantonese, Japanese, English, Others	370 / 1208
T02	Test set for unseen singers	6	Cantonese, Mandarin, Japanese	1685 / 1006
T03	T02 over 4 communication codecs	6	Cantonese, Mandarin, Japanese	6740 / 4024
T04	Test set for Persian musical context	17	Persian, English	353 / 166

- Speech anti-spoofing models heavily degrade on SVDD task!

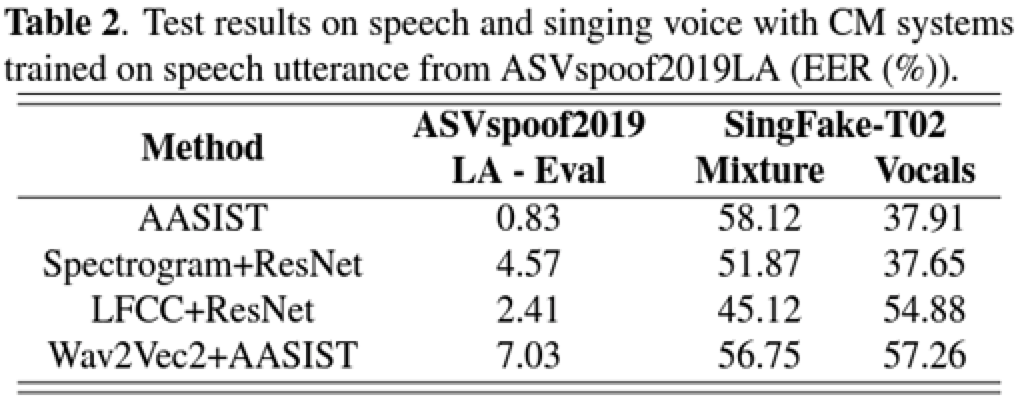


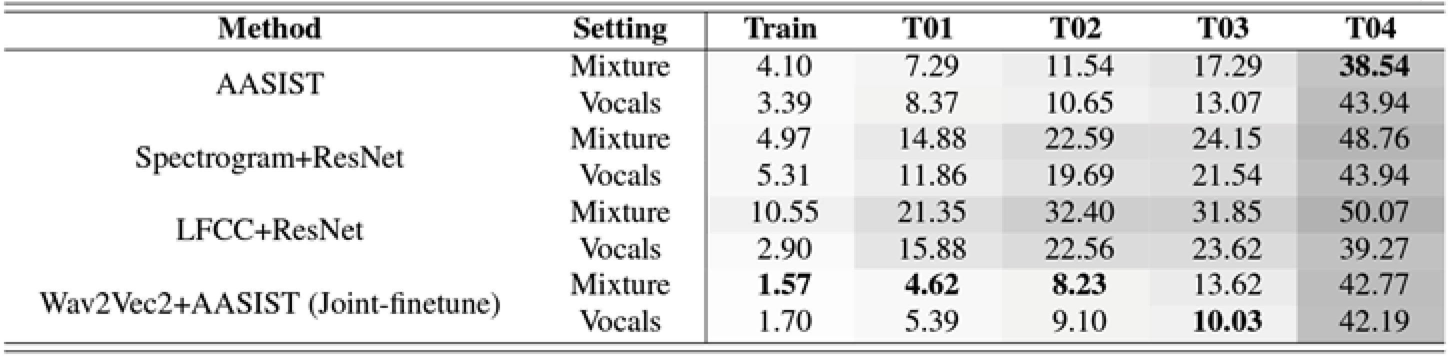
Table 2. Test results on speech and singing voice with CM systems trained on speech utterance from ASVspoof2019LA (EER (%)).

Method	ASVspoof2019	SingFake-T02	
	LA - Eval	Mixture	Vocals
AASIST	0.83	58.12	37.91
Spectrogram+ResNet	4.57	51.87	37.65
LFCC+ResNet	2.41	45.12	54.88
Wav2Vec2+AASIST	7.03	56.75	57.26

Yongyi Zang*, You Zhang*, Mojtaba Heydari, and Zhiyao Duan. “SingFake: Singing Voice Deepfake Detection”, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. (* equal contribution)

Preliminary work: SingFake

Results of training on SingFake data



	Song	Seen	Unseen	Unseen	Unseen	Unseen	
	Singer	Seen	Seen	Unseen	Unseen	Unseen	
	Codec	Seen	Seen	Seen	Unseen	Unseen	Trained on speech
	Context	Seen	Seen	Seen	Seen	Unseen	
Method	Setting	Train	T01	T02	T03	T04	T02
AASIST	Mixture	4.10	7.29	11.54	17.29	**38.54**	58.12
	Vocals	3.39	8.37	10.65	13.07	43.94	37.91
Spectrogram+ResNet	Mixture	4.97	14.88	22.59	24.15	48.76	51.87
	Vocals	5.31	11.86	19.69	21.54	43.94	37.65
LFCC+ResNet	Mixture	10.55	21.35	32.40	31.85	50.07	45.12
	Vocals	2.90	15.88	22.56	23.62	39.27	54.88
Wav2Vec2+AASIST (Joint-finetune)	Mixture	**1.57**	**4.62**	**8.23**	13.62	42.77	56.75
	Vocals	1.70	5.39	9.10	**10.03**	42.19	57.26

- Training on singing voices improves SVDD performance
- SVDD systems show limited robustness to unseen scenarios

Yongyi Zang*, You Zhang*, Mojtaba Heydari, and Zhiyao Duan. “SingFake: Singing Voice Deepfake Detection”, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. (* equal contribution)

Singing Voice Deepfake Detection (SVDD) challenge

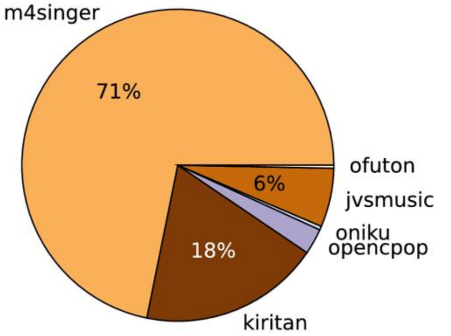
- CtrSVDD (Controlled setting)
  - Clean vocals generated by state-of-the-art singing voice synthesis (SVS) and singing voice conversion (SVC) systems based on open-source pop song datasets
- WildSVDD (In-the-wild setting)
  - Expanded SingFake dataset with newly collected data

https://svddchallenge.org

Zhang, Y., Zang, Y., Shi, J., Yamamoto, R., Toda, T., & Duan, Z. (2024). SVDD 2024: The inaugural singing voice deepfake detection challenge. *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

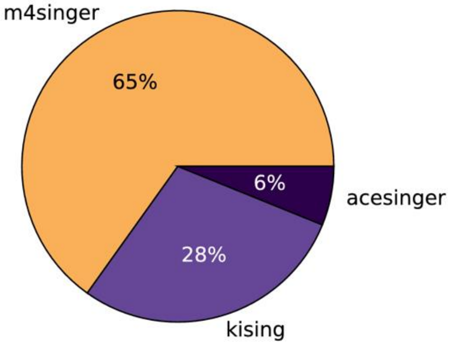
CtrSVDD Dataset

- 307.98 hours total (220,798 mono vocal clips)
  - 47.64 hours of bonafide vocals from **open-source singing datasets**
  - 260.34 hours of deepfake vocals using **14 synthesis methods**
- **164 singer identities**
- Average clip length: 5.02 seconds, 16 kHz sample rate
- **Fully accessible** under CC BY-NC-ND 4.0 license



(a) Source datasets on the training and development sets

Train + Dev

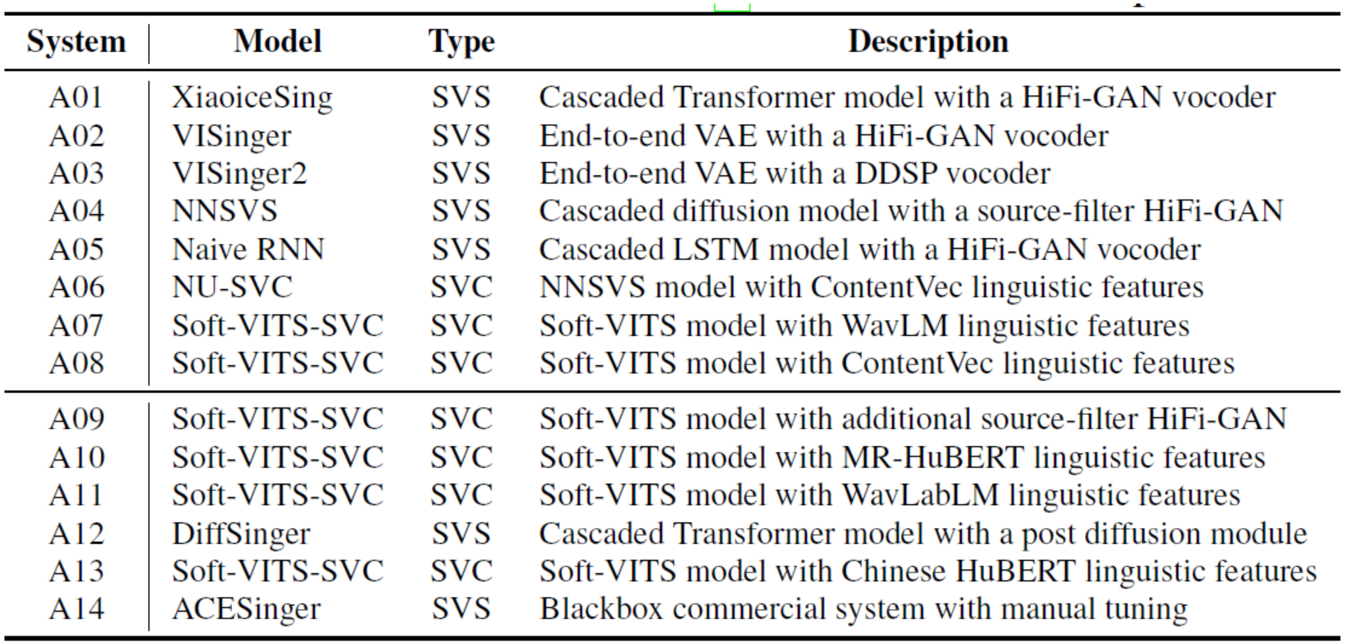


(c) Source datasets on the evaluation set

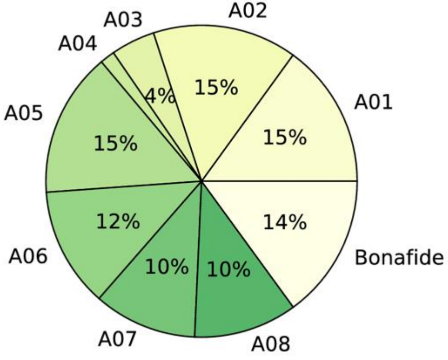
Test (with labels)

Zang, Y., Shi, J., Zhang, Y., Yamamoto, R., Han, J., Tang, Y., Xu, S., Zhao, W., Guo, J., Toda, T., Duan, Z. (2024) CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. *Proc. Interspeech 2024*, 4783-4787, doi: 10.21437/Interspeech.2024-2242

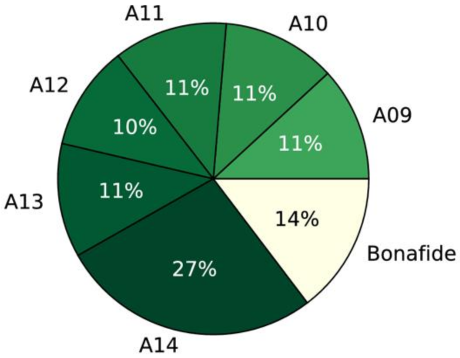
CtrSVDD Dataset



System	Model	Type	Description
A01	XiaoiceSing	SVS	Cascaded Transformer model with a HiFi-GAN vocoder
A02	VISinger	SVS	End-to-end VAE with a HiFi-GAN vocoder
A03	VISinger2	SVS	End-to-end VAE with a DDSP vocoder
A04	NNSVS	SVS	Cascaded diffusion model with a source-filter HiFi-GAN
A05	Naive RNN	SVS	Cascaded LSTM model with a HiFi-GAN vocoder
A06	NU-SVC	SVC	NNSVS model with ContentVec linguistic features
A07	Soft-VITS-SVC	SVC	Soft-VITS model with WavLM linguistic features
A08	Soft-VITS-SVC	SVC	Soft-VITS model with ContentVec linguistic features
A09	Soft-VITS-SVC	SVC	Soft-VITS model with additional source-filter HiFi-GAN
A10	Soft-VITS-SVC	SVC	Soft-VITS model with MR-HuBERT linguistic features
A11	Soft-VITS-SVC	SVC	Soft-VITS model with WavLabLM linguistic features
A12	DiffSinger	SVS	Cascaded Transformer model with a post diffusion module
A13	Soft-VITS-SVC	SVC	Soft-VITS model with Chinese HuBERT linguistic features
A14	ACESinger	SVS	Blackbox commercial system with manual tuning



(b) Deepfake methods on the training and development sets



(d) Deepfake methods on the evaluation set

Zang, Y., Shi, J., Zhang, Y., Yamamoto, R., Han, J., Tang, Y., Xu, S., Zhao, W., Guo, J., Toda, T., Duan, Z. (2024) CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. *Proc. Interspeech 2024*, 4783-4787, doi: 10.21437/Interspeech.2024-2242

WildSVDD Dataset

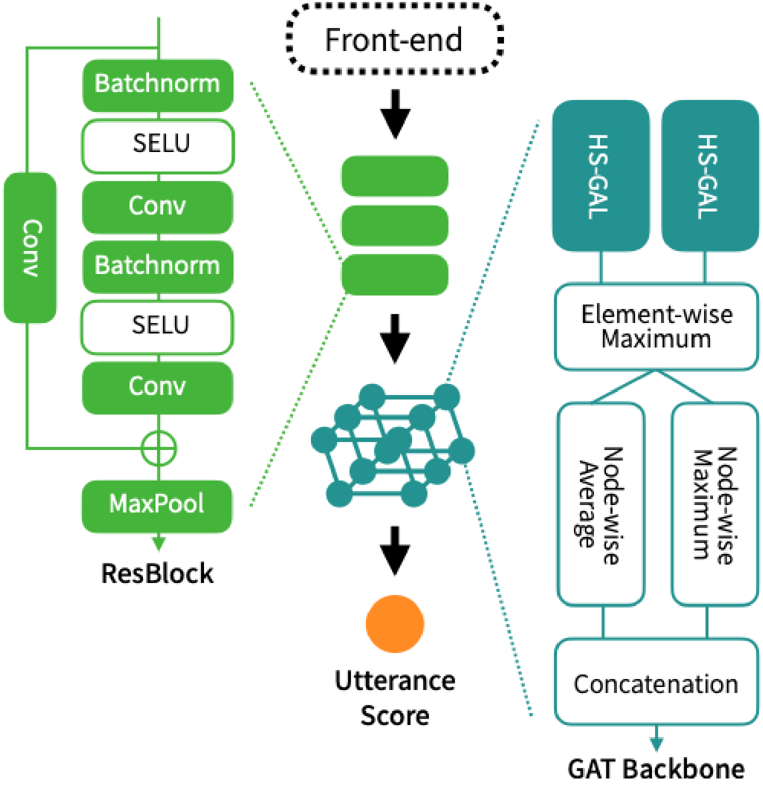
- Expanded SingFake with newly collected data
    - Nearly 2x SingFake
    - Multi-lingual: Mandarin, Cantonese, Korean, English, Japanese, Others…
    - Removed expired videos
- Freely split development set from the training set
- Test sets:
    - Test A: Unseen singers, similar to T02 in SingFake
    - Test B: Unseen musical context, same as T04 in SingFake

Zenodo link

Baseline system

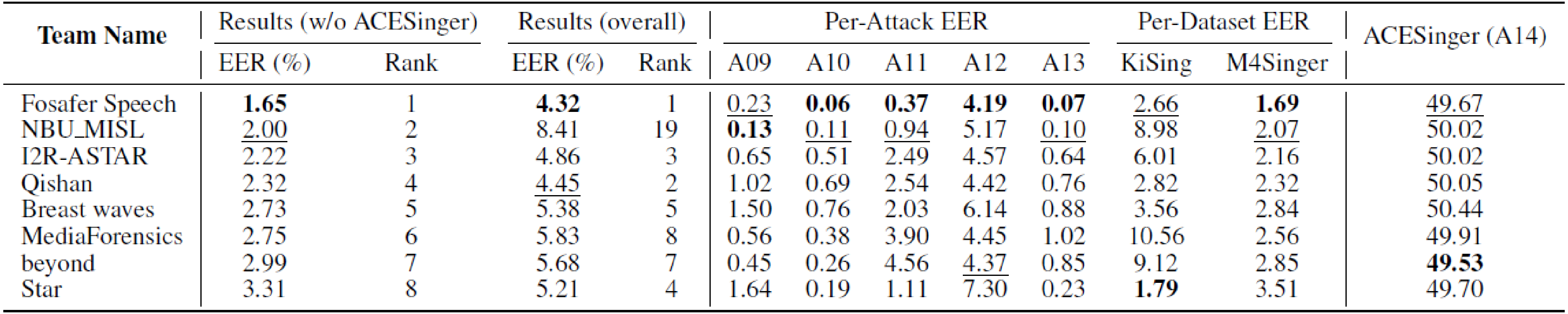
- AASIST: a graph-neural-network based backbone, well-recognized in speech anti-spoofing task

- Can be integrated with different front-ends:
  - Spectrogram
  - Mel-spectrogram
  - MFCC
  - **LFCC**
  - **Raw waveform**
  - **Self-Supervised Learning (SSL) feature (wav2vec)**



CtrSVDD challenge results

- 47 submissions, 37 out of which surpassed baselines
- Best performance: 1.65% EER

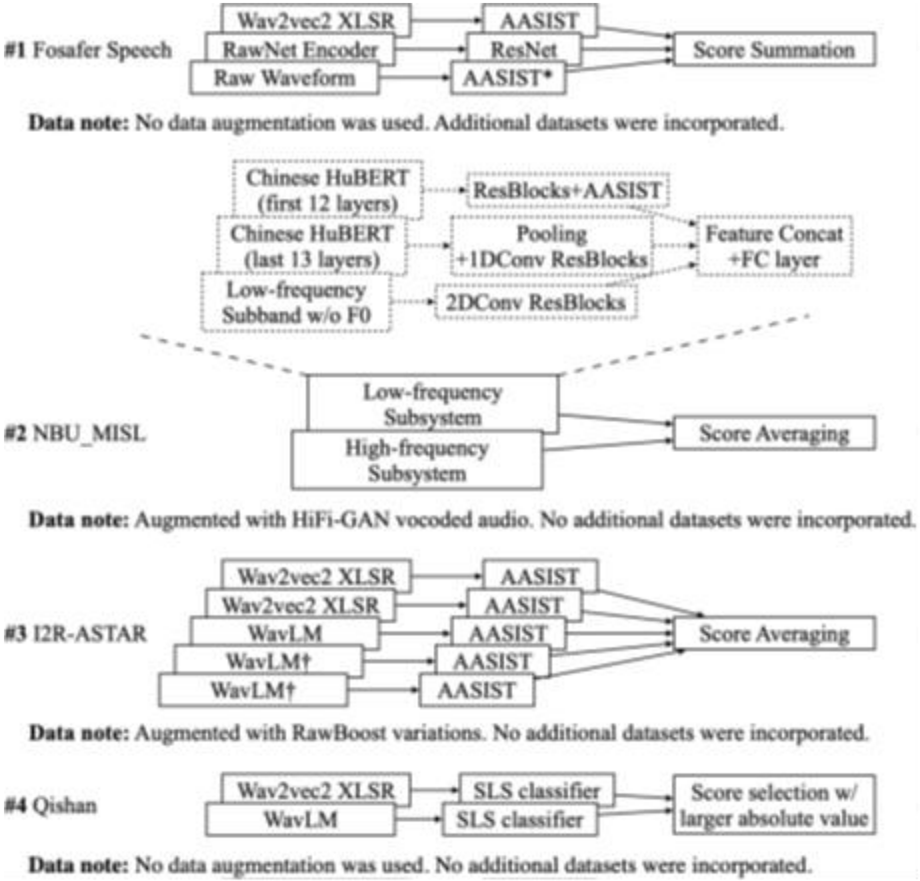


Team Name	Results (w/o ACESinger)		Results (overall)		Per-Attack EER					Per-Dataset EER		ACESinger (A14)
	EER (%)	Rank	EER (%)	Rank	A09	A10	A11	A12	A13	KiSing	M4Singer	
Fosafer Speech	**1.65**	1	**4.32**	1	0.23	**0.06**	**0.37**	**4.19**	**0.07**	2.66	**1.69**	49.67
NBU_MISL	2.00	2	8.41	19	**0.13**	0.11	0.94	5.17	0.10	8.98	2.07	50.02
I2R-ASTAR	2.22	3	4.86	3	0.65	0.51	2.49	4.57	0.64	6.01	2.16	50.02
Qishan	2.32	4	4.45	2	1.02	0.69	2.54	4.42	0.76	2.82	2.32	50.05
Breast waves	2.73	5	5.38	5	1.50	0.76	2.03	6.14	0.88	3.56	2.84	50.44
MediaForensics	2.75	6	5.83	8	0.56	0.38	3.90	4.45	1.02	10.56	2.56	49.91
beyond	2.99	7	5.68	7	0.45	0.26	4.56	4.37	0.85	9.12	2.85	**49.53**
Star	3.31	8	5.21	4	1.64	0.19	1.11	7.30	0.23	**1.79**	3.51	49.70

- A12 (diffusion-based) is a bit challenging
- A14 (out-of-domain) data is quite challenging

CtrSVDD winning solutions

- SSL features and ensemble learning are common winning strategies
- Most methods adapt from speech deepfake detection methods
- Lack specific design for singing voice



Data note: No data augmentation was used. Additional datasets were incorporated.

Data note: Augmented with HiFi-GAN vocoded audio. No additional datasets were incorporated.

Data note: Augmented with RawBoost variations. No additional datasets were incorporated.

Data note: No data augmentation was used. No additional datasets were incorporated.

WildSVDD challenge results

- 4 teams participated, all surpassed baselines

Team	Methods Used	EER on test_A	EER on test_B
UNIBS1	Log-spectrogram+ResNet - Vocals	2.38	9.81
UNIBS2	Log-spectrogram+ResNet - Mixtures	2.70	12.19
IMS-SCU1	Ensemble - Vocals	2.70	12.95
IMS-SCU2	WavLM - Vocals	3.54	15.32
IMS-SCU3	Ensemble - Mixtures	3.61	11.00
NTU	SingGraph - Mixtures	4.31	31.82
IMS-SCU4	WavLM - Mixtures	4.94	16.72
PDL	Ensemble - Vocals	5.80	22.01
Baseline1	Wav2vec - Vocals	6.09	24.09
Baseline2	Raw - Vocals	8.84	26.11
Baseline3	Wav2vec - Mixtures	9.57	21.45
Baseline4	Raw - Mixtures	10.88	17.69

Takeaways

- SVDD offers a new and challenging test ground for audio deepfake detection
  - Training on speech data does not work well, but model designs work
  - May need specific model designs to account for special characteristics of singing voices
- Some previous findings may or may not hold, serving as a retrospect / rethinking on deepfake detection research
  - Winner of WildSVDD@MIREX challenge is a LogSpec+ResNet method pre-trained on ImageNet.

Acknowledgments

- Organizers

You Zhang
University of Rochester
you.zhang@rochester.edu

Yongyi Zang
University of Rochester
yongyi.zang@rochester.edu

Jiatong Shi
Carnegie Mellon University
jiatongs@andrew.cmu.edu

Ryuichi Yamamoto
Nagoya University
zryuichi@gmail.com

Tomoki Toda
Nagoya University
tomoki@icts.nagoya-u.ac.jp

Zhiyao Duan
University of Rochester
zhiyao.duan@rochester.edu

- Funding agencies
- Challenge participants

Schedule for the SVDD special session

- First hour: SVDD Challenge
  - (15 min) Challenge overview presentation
  - (3 x 10 min) Lightning talks from CtrSVDD winners
  - (3 x 5 min) Lightning talks from WildSVDD winners
- Second hour: Discussions on SVDD research
  - (20 min) Invited Talk: Xueyao Zhang
  - (20 min) Invited Talk: Chang Zeng
  - (20 min) Panel Discussion

Special session
webpage