SVDD 2024

The Inaugural Singing Voice Deepfake Detection Challenge

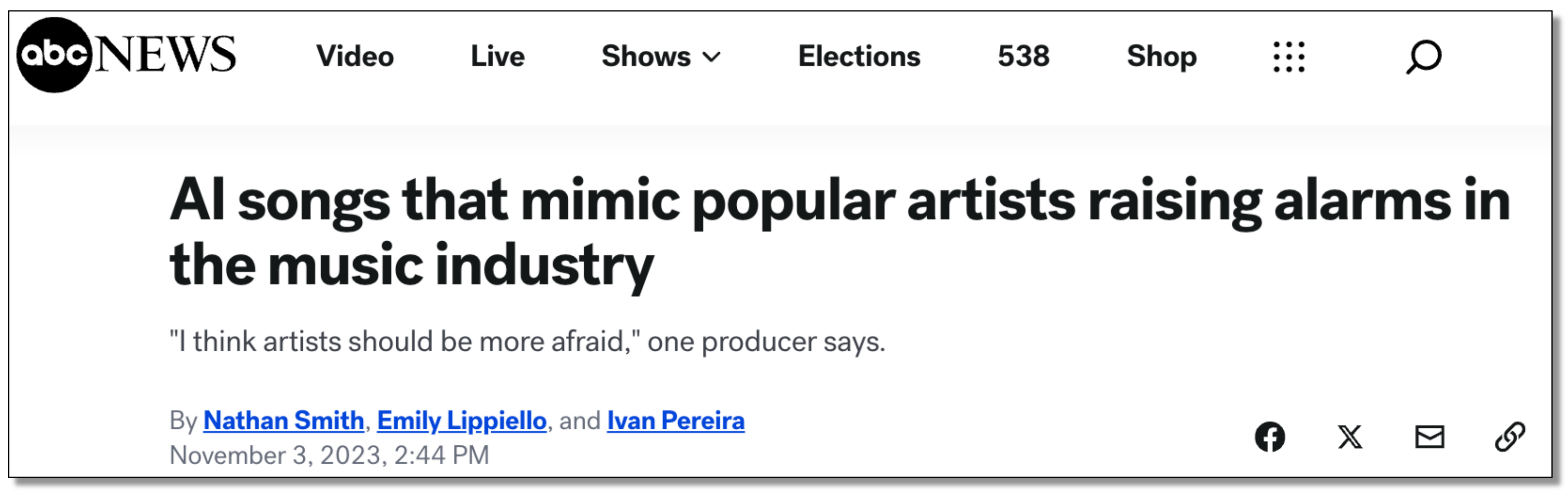
You Zhang [1], Yongyi Zang [1], Jiatong Shi [2], Ryuichi Yamamoto [3], Tomoki Toda [3], Zhiyao Duan [1]

1 University of Rochester
2 Carnegie Mellon University
3 Nagoya University

SLT 2024

Background

Singing Voice Deepfakes are raising public and industry concerns.

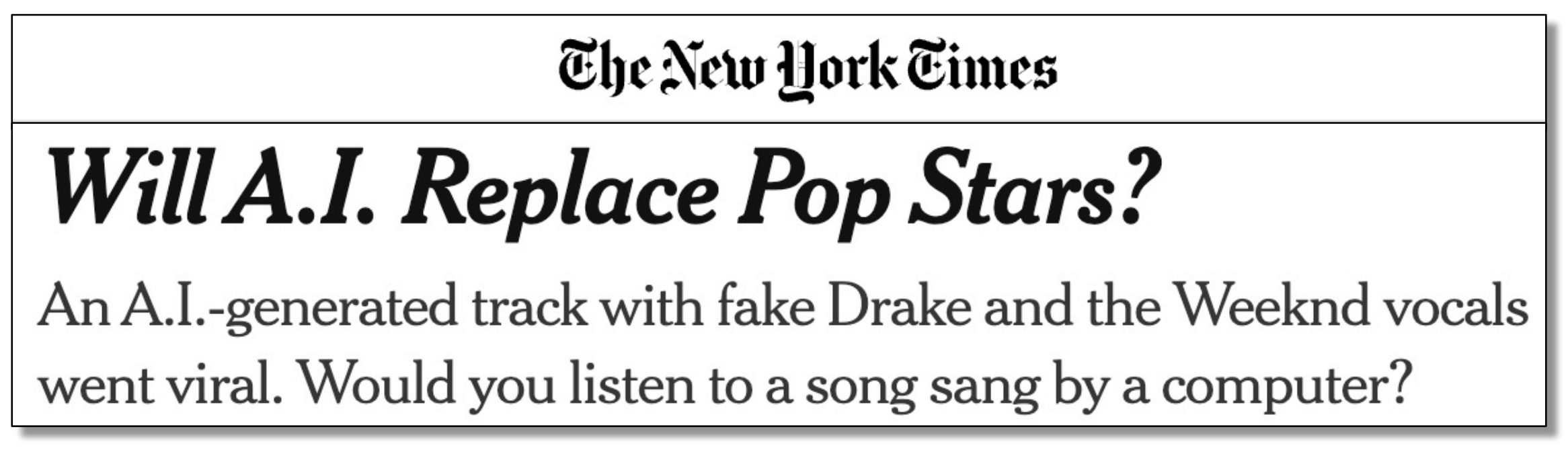


abc NEWS Video Live Shows Elections 538 Shop

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



The New York Times

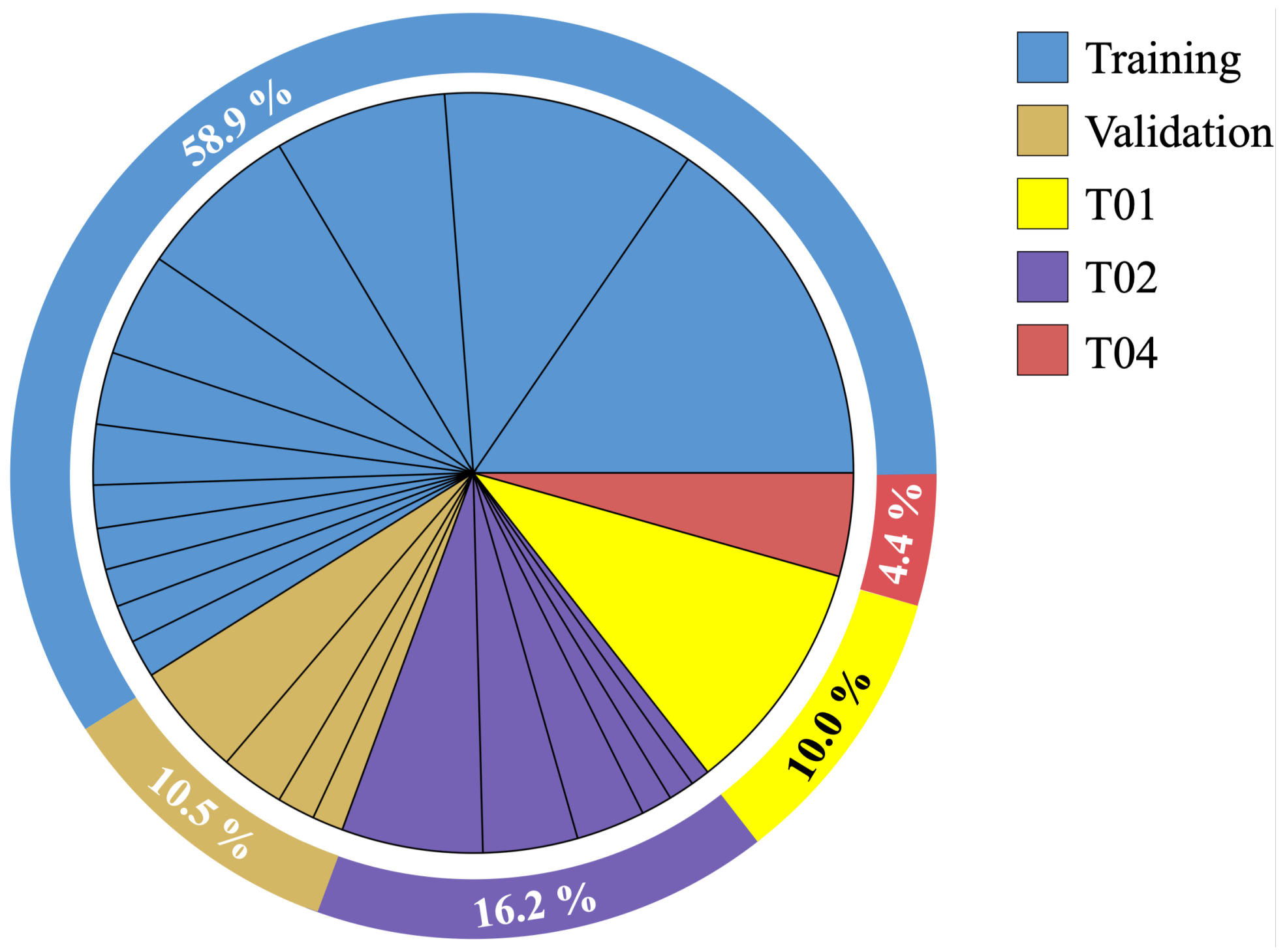
Will A.I. Replace Pop Stars?

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

WildSVDD: Singing Voice Deepfake Detection in the Wild

Datasets collected from media platforms

Previous work: **SingFake** [1] proposed the novel task of SVDD, presented the SingFake dataset, and identified several challenges.



WildSVDD: An expanded SingFake with newly collected data

Participants can freely split the development set from the training set.
Test A: Unseen singers, similar to T02 in SingFake
Test B: Unseen musical context, same as T04 in SingFake

WildSVDD Baselines

AASIST [3] with various front-ends

Front-end	WildSVDD Test A		WildSVDD Test B	
	Mixtures	Vocals	Mixtures	Vocals
Raw Waveform	10.50	8.48	16.85	14.91
Spectrogram	27.93	20.55	30.97	24.41
Mel-Spectrogram	29.27	27.35	32.18	30.78
MFCC	17.78	19.14	22.92	23.31
LFCC	22.60	23.25	26.82	26.94
Wav2vec2 XLS-R	9.57	6.09	21.45	24.09

Acknowledgments

Resources

Website

Paper

REFERENCES

[1] Zang, Y., Zhang, Y., Heydari, M., & Duan, Z. (2024). Singfake: Singing voice deepfake detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12156-12160).

[2] Zang, Y., Shi, J., Zhang, Y., Yamamoto, R., Han, J., Tang, Y., ... & Duan, Z. (2024). CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. *Proc. Interspeech* (pp. 4783-4787).

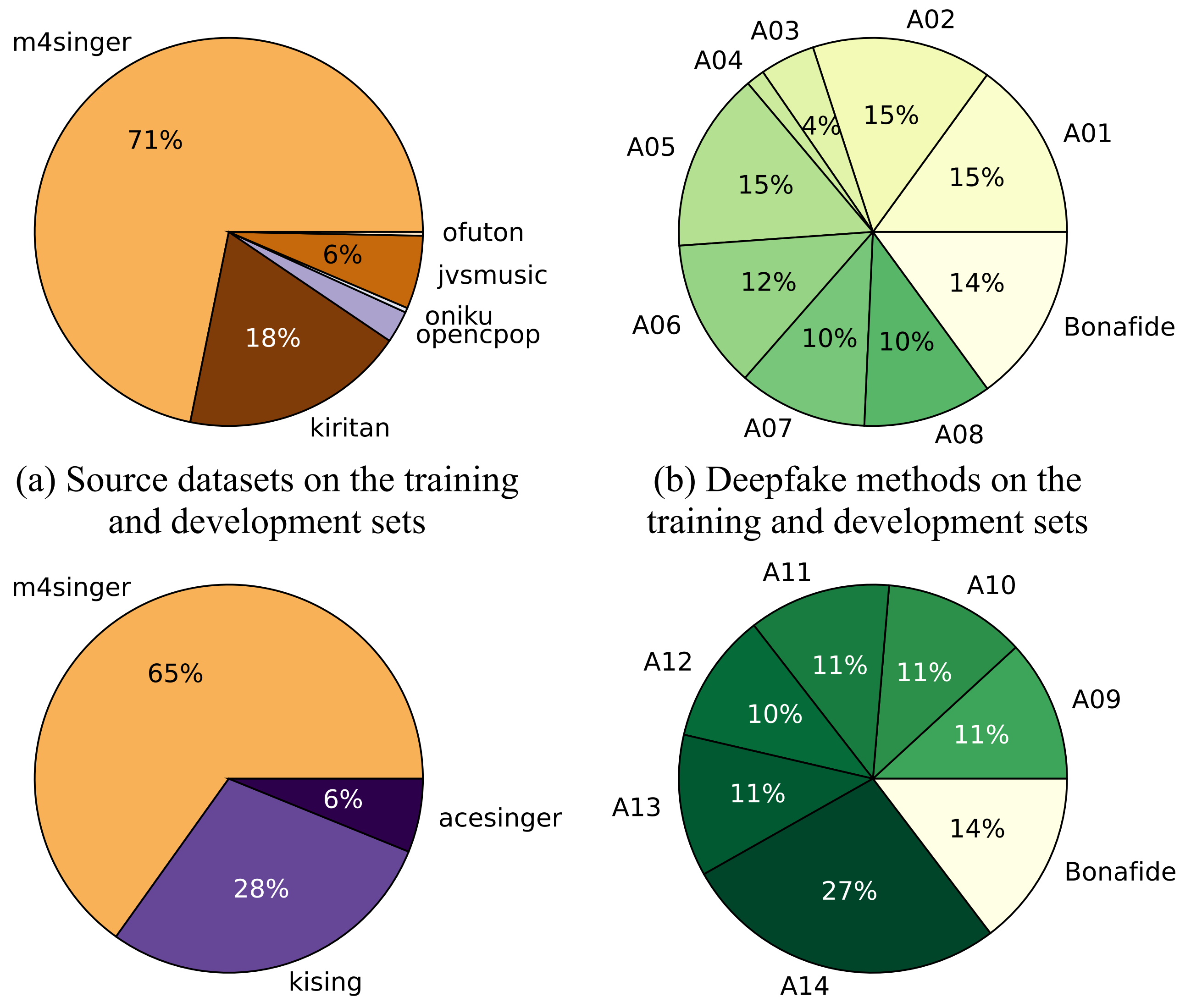
[3] Jung, J. W., Heo, H. S., Tak, H., Shim, H. J., Chung, J. S., Lee, B. J., ... & Evans, N. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6367-6371).

CtrSVDD: Controlled Singing Voice Deepfake Detection

CtrSVDD Dataset [2]

47.64 hours of bonafide vocals from open-source singing datasets
260.34 hours of deepfake vocals using 14 synthesis methods

Overview of source datasets and deepfake methods distribution [2]



(a) Source datasets on the training and development sets

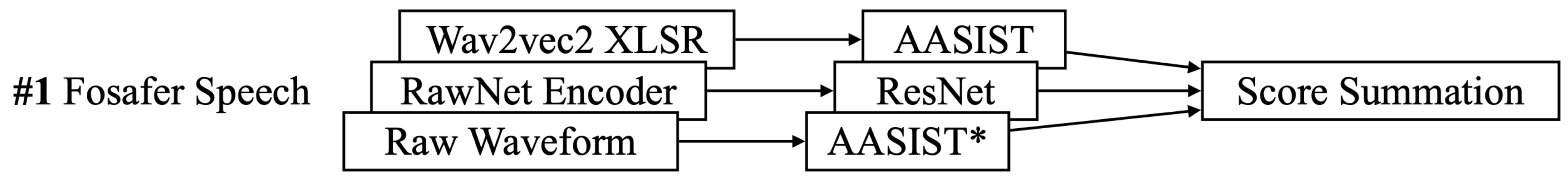
(b) Deepfake methods on the training and development sets

(c) Source datasets on the evaluation set

(d) Deepfake methods on the evaluation set

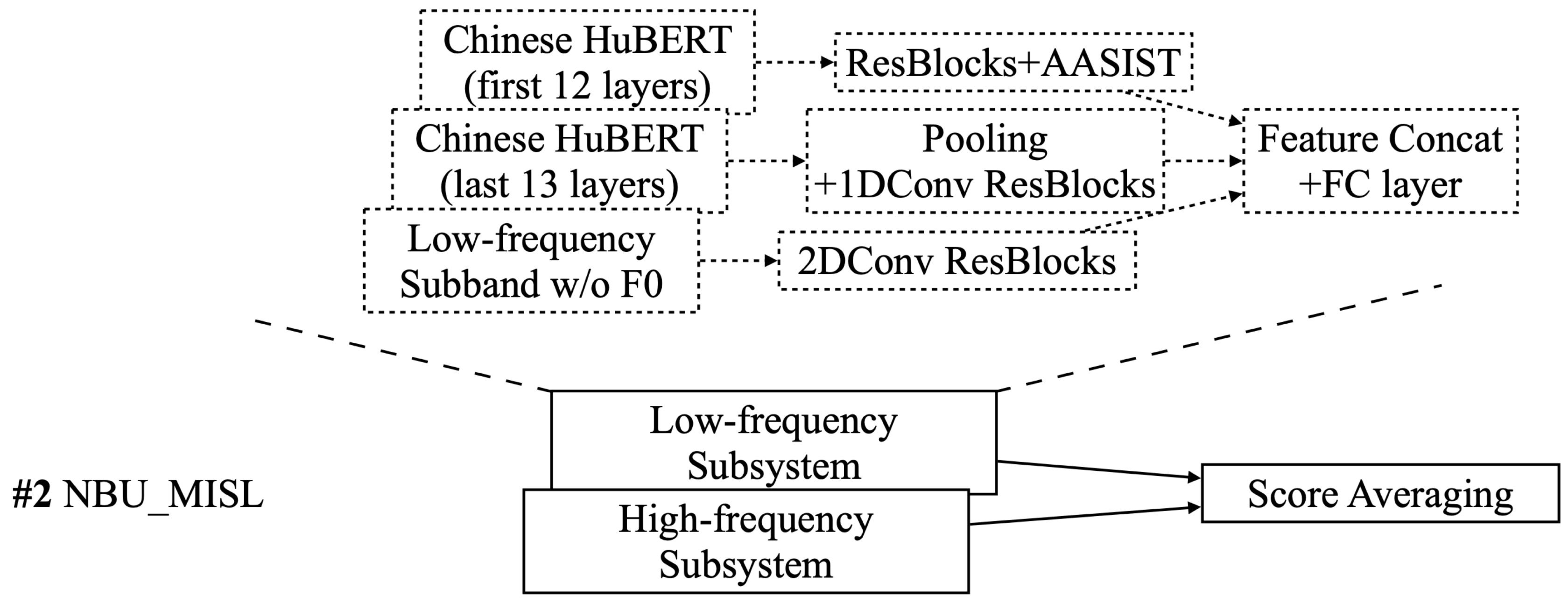
Winning solutions:

Illustration of the top-4 ranked system submissions for the CtrSVDD track



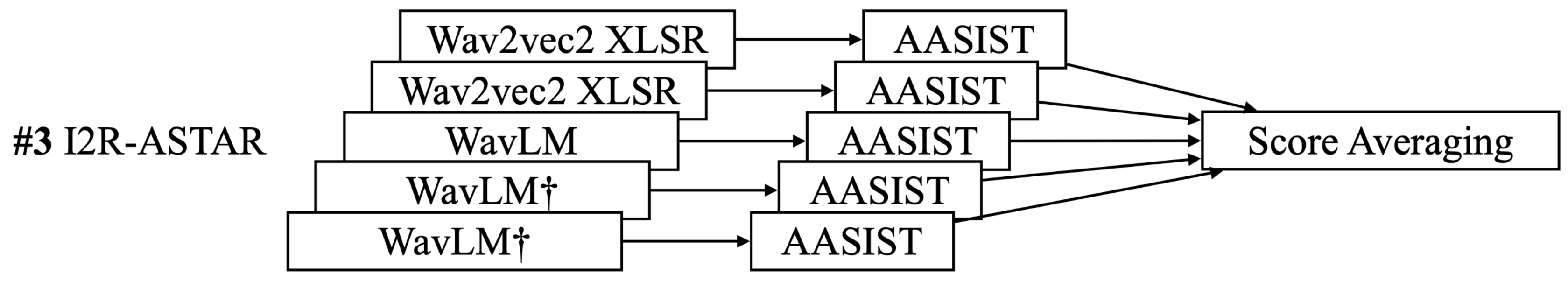
#1 Fosafer Speech
Wav2vec2 XLSR
RawNet Encoder
Raw Waveform
AASIST
ResNet
AASIST*
Score Summation

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



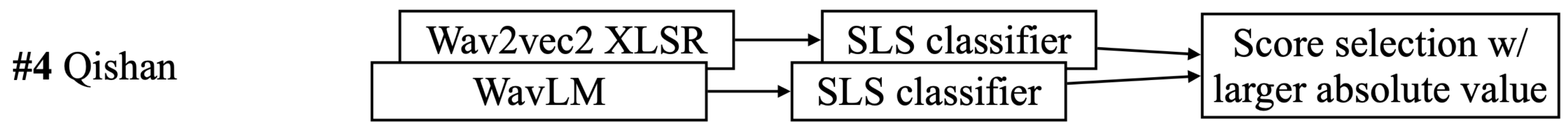
Chinese HuBERT (first 12 layers)
Chinese HuBERT (last 13 layers)
Low-frequency Subband w/o F0
ResBlocks+AASIST
Pooling +1DConv ResBlocks
2DConv ResBlocks
Feature Concat +FC layer
#2 NBU_MISL
Low-frequency Subsystem
High-frequency Subsystem
Score Averaging

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



#3 I2R-ASTAR
Wav2vec2 XLSR
Wav2vec2 XLSR
WavLM
WavLM†
WavLM†
AASIST
AASIST
AASIST
AASIST
AASIST
Score Averaging

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



#4 Qishan
Wav2vec2 XLSR
WavLM
SLS classifier
SLS classifier
Score selection w/ larger absolute value

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

Challenge results: Overview of the top-8 ranked submission results

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