

Speech Anti-Spoofing

Neil Zhang

ECE 277/477 - Computer Audition, Fall 2023

Announcement

HW6 was released on Tuesday.

https://drive.google.com/drive/folders/1Y4_N1aEtBSbEU9W3jEujVmVIKZ7RO40B?usp=sharing

Outline

Introduction to speech anti-spoofing

Generalization ability to unseen synthetic attacks

- One-class learning: OC-Softmax, SAMO

Robustness to channel variation:

- Channel-robust training strategies, phase perturbation

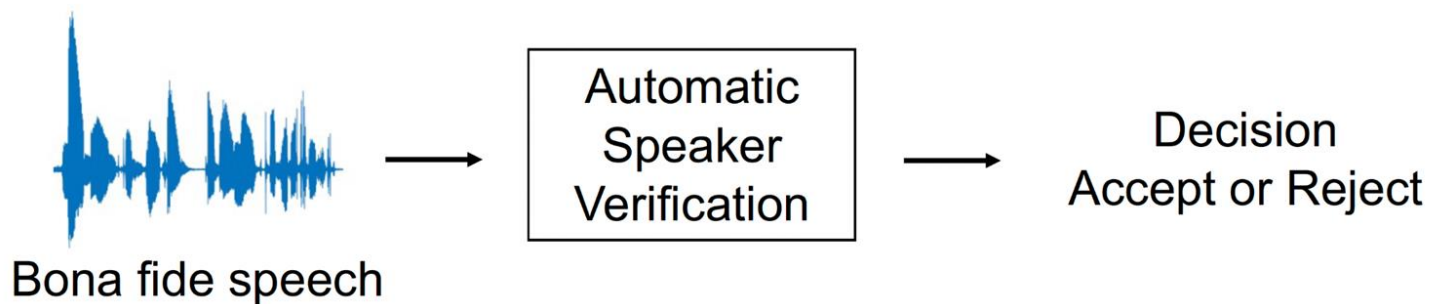
Joint optimization with speaker verification: Probabilistic fusion framework

Beyond speech anti-spoofing: **Singing voice** deepfake detection

Future directions

Voice Biometrics

Verify the identity of a speaker



We expect that the input (bona fide speech) is from a real person.

ARTIFICIAL INTELLIGENCE

Microsoft's New AI Can Simulate Anyone's Voice From a 3-Second Sample

By John P. Mello Jr. • January 11, 2023 8:06 AM PT • Email Article

EDITORS' PICK

Fraudsters Cloned Company Director's Voice In \$35 Million Bank Heist, Police Find

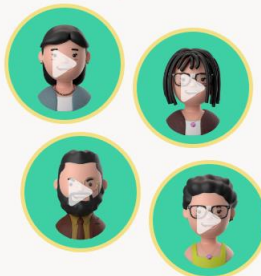
[top left](#)
[top right](#)
[bottom left](#)
[bottom right](#)

RESEMBLE.AI

PRODUCTS USE CASES PRICING SIGN IN

Your Complete Generative Voice AI Toolkit

community built voices



- Text-to-Speech
- Speech-to-Speech
- Neural Audio Editing
- Language Dubbing

Resemble's AI voice generator lets you create human-like voice overs in seconds.

Home > Music Industry News

AI Voice Tool Abused to Make Celebrity Deepfake Audio Clips

Ashley King February 1, 2023

Spoofing attacks

Impersonation

- twins and professional mimics

Replay

- reuse pre-recorded audio, most accessible

Text-to-speech (TTS)

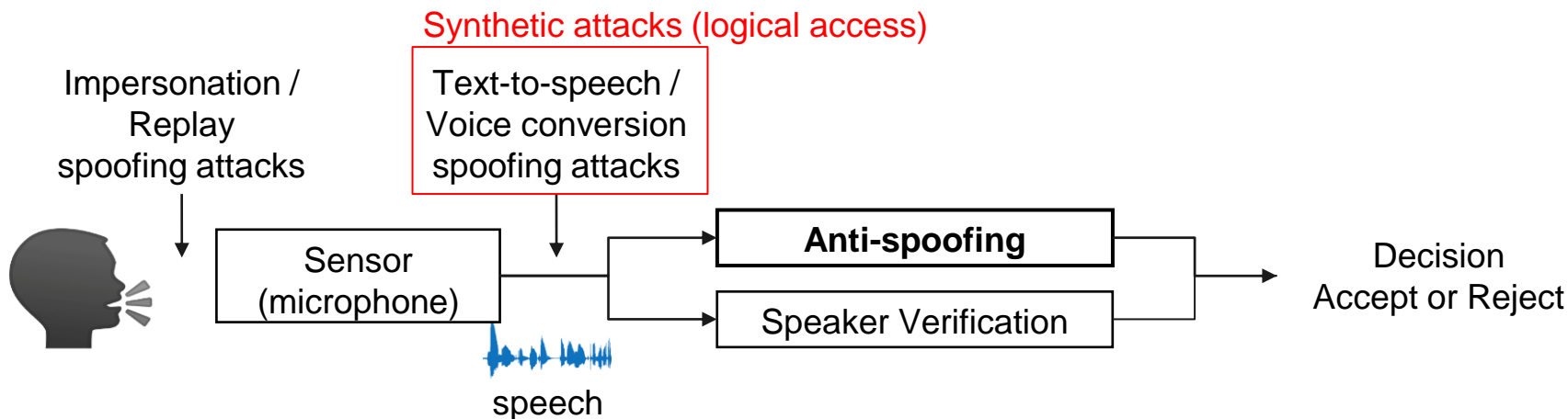
- convert written text into spoken words with speech synthesis

Voice conversion (VC)

- convert speech from source speaker to target speaker's voice

Speech Anti-Spoofing

A voice anti-spoofing system is desired to distinguish synthetic **speech** from **bona fide speech**.



ASVspoof challenge series

- LA: Robust to channel variability
- PA: Involve real replayed samples
- DF: a new speech deepfake task

2015

Replay spoofing
attacks detection

2019

Text-to-speech
(TTS) and voice
conversion (VC)
spoofing attacks
detection

2017

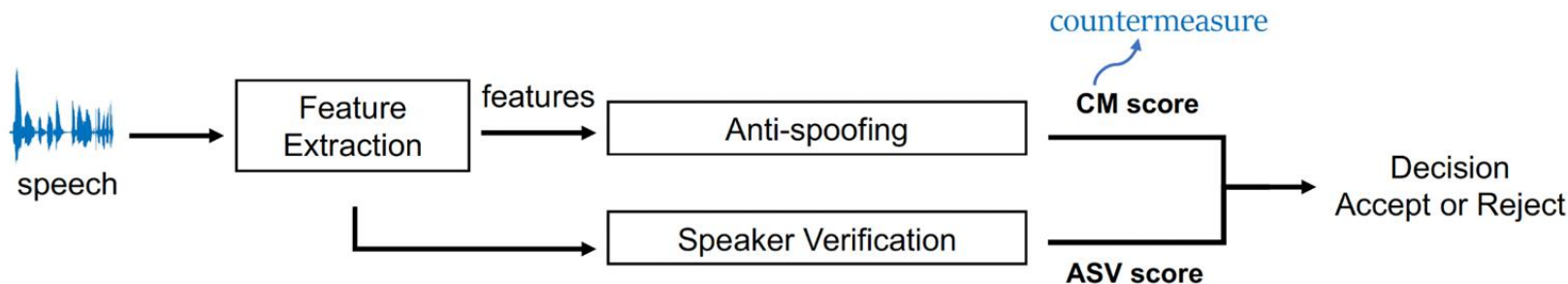
- LA: Advanced TTS and VC attacks
- PA: More controlled setup for replay attacks
- A new evaluation metric

2021

LA: algorithm-related artifacts

PA: device-related artifacts

Evaluation metric



- EER (Equal Error rate)

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

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

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

Dataset

ASVspoof 2019 Logical Access (TTS + VC)

- Bona fide speech (VCTK dataset)
- 6 known attacks (appear in training set)
- 11 unknown attacks (only appear in eval set)
- 2 attacks (use known algorithms but trained with more data)

	Bona fide	Spoofed	
	# utterance	# utterance	attacks
Training	2,580	22,800	A01 - A06
Development	2,548	22,296	A01 - A06
Evaluation	7,355	63,882	A07 - A19

Research question

Motivation:

- The fast development of speech synthesis are posing increasingly more threat.
- The **distribution mismatch** between the training set and test set for the spoofing attacks class.

How can the anti-spoofing system defend against **unseen** spoofing attacks?

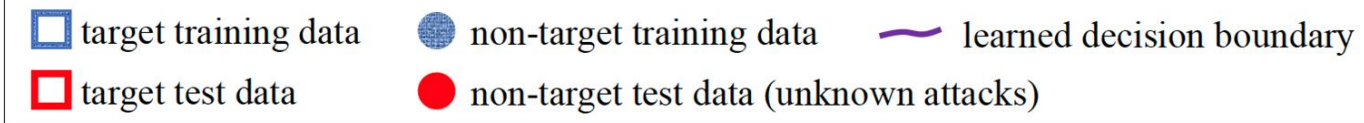
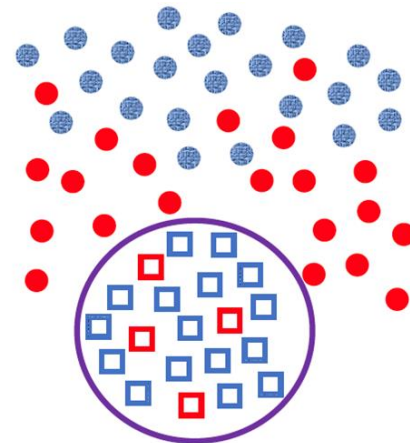
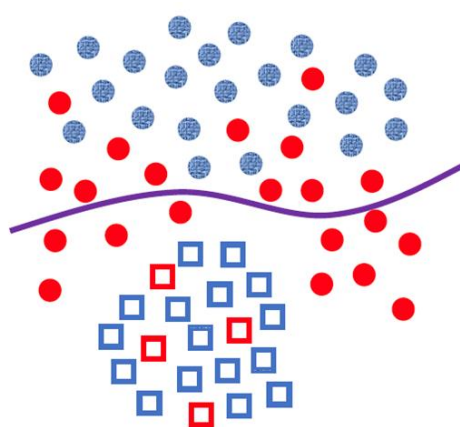
Generalization ability!

Definition of one-class classification

- “One of the classes (referred to as the positive class or **target** class)
 - is **well characterized** by instances in the training data.
- For the other class (**nontarget**),
 - it has either **no instances** at all,
 - **very few** of them,
 - or they do **not form a statistically-representative** sample of the negative concept.”

Khan, S. S., & Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3), 345-374.

Illustration of comparison



(a) Binary classification

(b) One-class classification

You Zhang, Fei Jiang, Ge Zhu, Xinhui Chen, and Zhiyao Duan. "Generalizing Voice Presentation Attack Detection to Unseen Synthetic Attacks and Channel Variation", *Handbook of Biometric Anti-spoofing (3rd edition)*, Springer, 2023.

One-class learning

- Compact the bona fide speech representation
- Isolate the spoofing attacks

Training: OC-Softmax loss (Proposed)

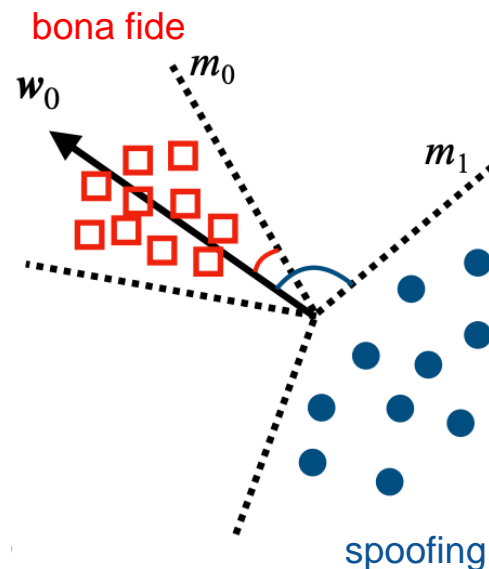
$$\mathcal{L}_{OCS} = \frac{1}{N} \sum_{i=1}^N \log \left(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \hat{x}_i)(-1)^{y_i}} \right).$$

Diagram illustrating the OC-Softmax loss formula with annotations:

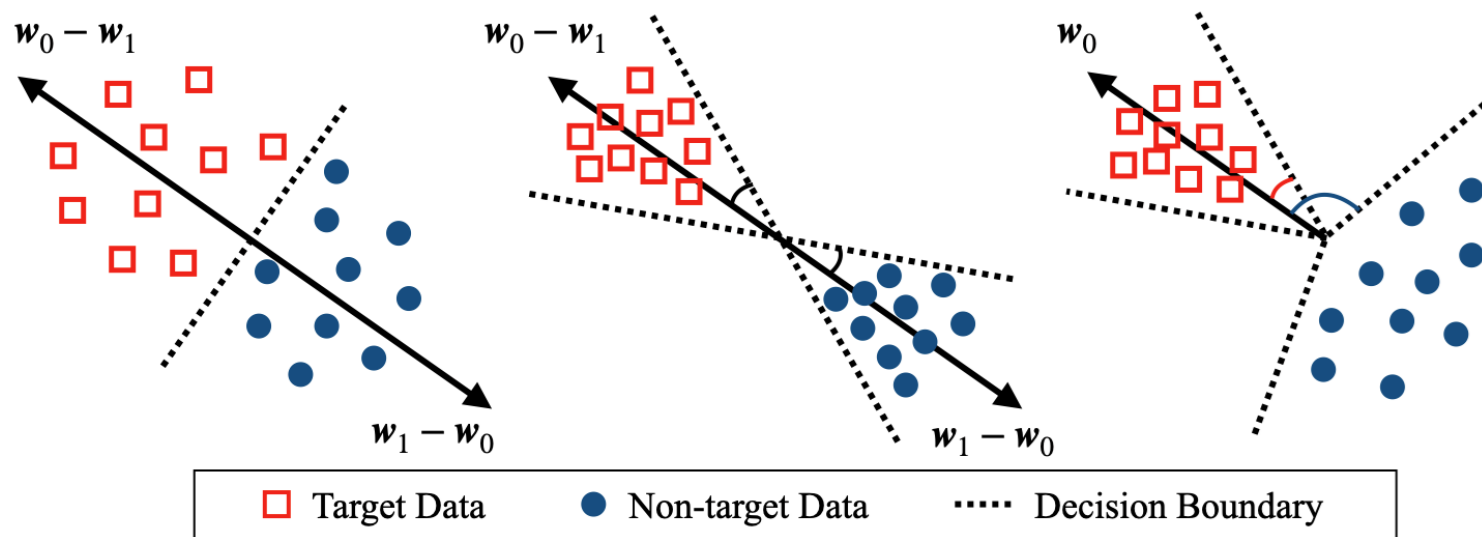
- $\frac{1}{N}$: # samples
- α : scale factor
- m_{y_i} : margin
- \hat{w}_0 : center vector
- \hat{x}_i : embedding
- $(-1)^{y_i}$: label

Inference: cosine similarity

$$S_{OCS} = \hat{w}_0 \hat{x}_i.$$



Comparing OC-Softmax with binary classification



(a) Original Softmax

(b) AM-Softmax

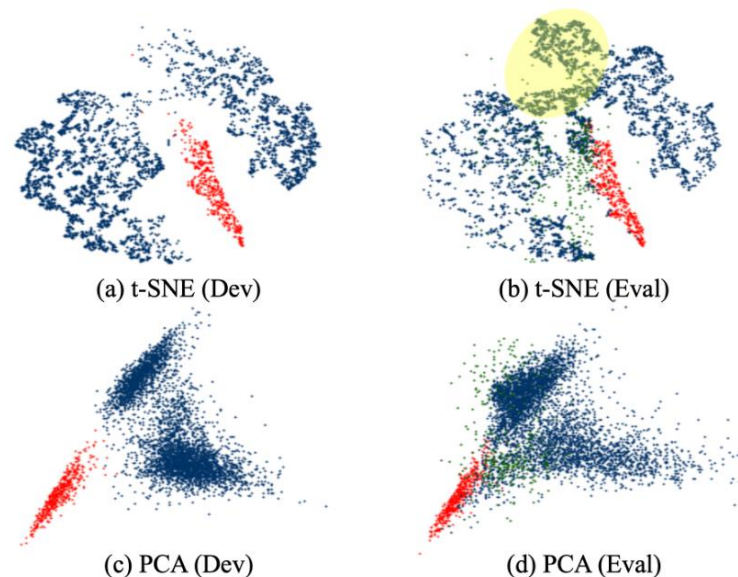
(c) OC-Softmax (Proposed)

Evaluation of OC-Softmax

Results on the development and evaluation sets of ASVspoof 2019 LA using different losses

Loss	Dev Set		Eval Set	
	EER (%)	min t-DCF	EER (%)	min t-DCF
Softmax	0.35	0.010	4.69	0.125
AM-Softmax	0.43	0.013	3.26	0.082
OC-Softmax	0.20	0.006	2.19	0.059

- OC-Softmax performs the best on unseen attacks.
- Achieved the state-of-the-art single-system performance.

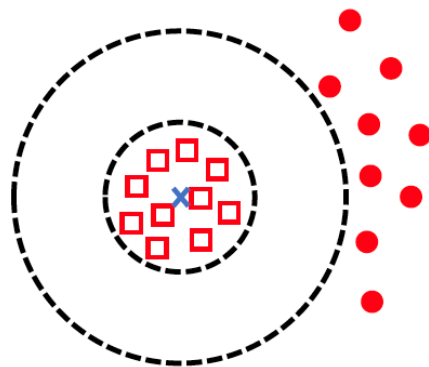


Feature Embedding Visualization
(red: bona fide, green: A17 attack, blue: spoofing attacks)

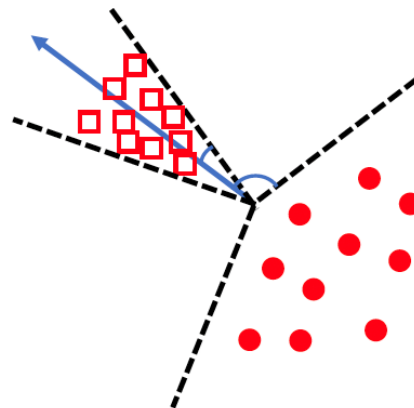
Other one-class loss functions

Euclidean distance-based one-class loss (isolate loss, single-center loss)

Cosine similarity-based one-class loss (OC-Softmax, angular isolate loss)



(a) Euclidean distance-based



(b) Cosine similarity-based

Research question

Motivation:

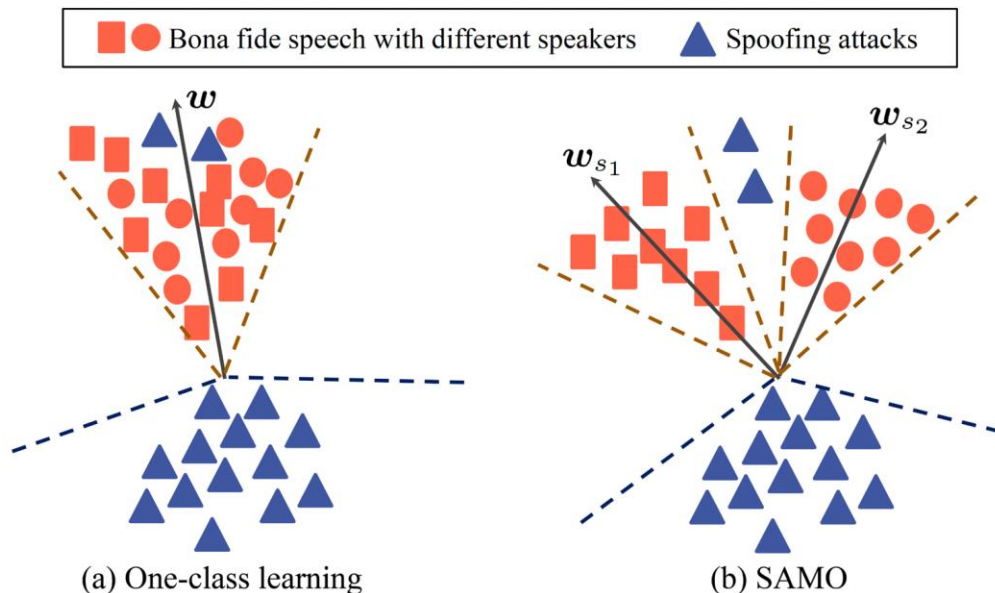
- In our previous work, we compact the embedding space of the bona fide speech into one cluster.
- However, due to **the variety of timbre and speaking traits of different speakers**, the bona fide speech of different speakers naturally forms multiple clusters in the embedding space.

How to improve the **generalization** ability while **maintaining the variation** of bona fide speech?

Speaker attractor multi-center one-class learning

Model speaker diversity while maintaining the generalization ability brought by one-class learning

- Discriminate bona fide vs. spoofing attacks
- Cluster bona fide speech according to speakers



Siwen Ding, You Zhang, and Zhiyao Duan. "SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Speech Anti-Spoofing, ECE 277/477 - Computer Audition, Fall 2023

Loss function for multi-center one-class learning

- Compact the bona fide speech representation belonging to the same speaker
- Push away the spoofing attacks from all speaker attractors

$$\mathcal{L}_{SAMO} = \frac{1}{N} \sum_{i=1}^N \log \left(1 + e^{\alpha(m_{y_i} - d_i)(-1)^{y_i}} \right),$$

Annotations for the equation:

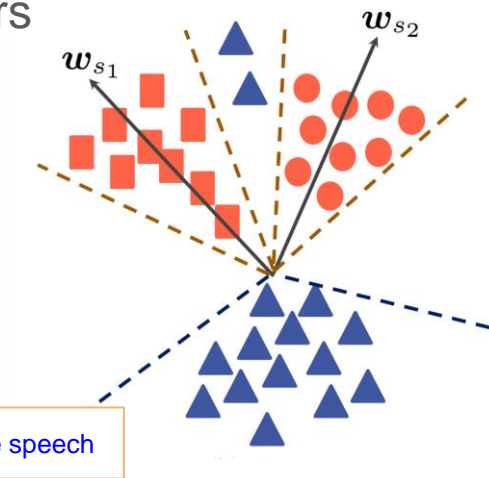
- # samples (points to N)
- scale factor (points to α)
- margin (points to m_{y_i})
- label (points to y_i)

where d_i is calculated by

$$d_i = \begin{cases} \hat{w}_{s_i} \hat{x}_i & \text{if } y_i = 0 \\ \max_s (\hat{w}_s \hat{x}_i), s \in \mathcal{S}_{train} & \text{if } y_i = 1 \end{cases}$$

Annotations for the definition:

- bona fide speech (points to $y_i = 0$)
- spoofing attacks (points to $y_i = 1$)



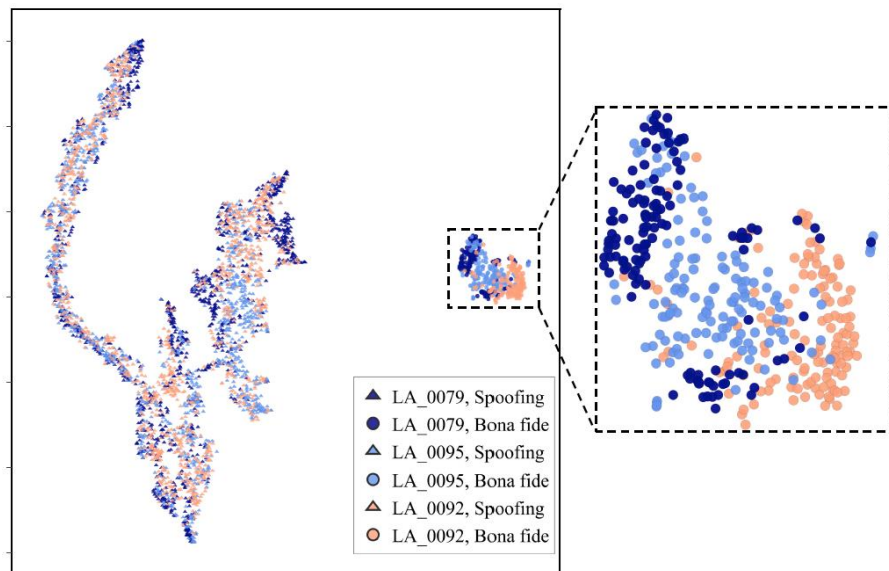
Siwen Ding, You Zhang, and Zhiyao Duan. "SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Speech Anti-Spoofing, ECE 277/477 - Computer Audition, Fall 2023

Embedding visualization

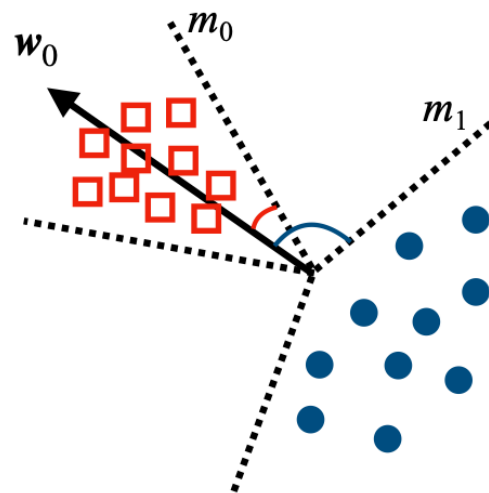
2D t-SNE visualization of SAMO feature embeddings of bona fide and spoofed speech of three speakers

- Bona fide utterances are grouped in a small region.
- Utterances of the three speakers are generally clustered according to speaker identity.



Takeaways

- One-class learning aims to **compact the target** class representation in the embedding space, and **push away non-target**.
- The proposed OC-Softmax and SAMO could improve the **generalization ability** of anti-spoofing system against **unseen spoofing attacks**.



Channel effects

-- Audio effects imposed onto the speech signal throughout the entire recording and transmission process

- Reverberation of recording **environments**
- Frequency responses of recording **devices**
- Compression algorithms in **telecommunication**

Research question:

How to improve the robustness to **channel variation**?

Cross-dataset performance

Table 1: *EER performance across different evaluation datasets (ASVspoof2019LA-eval, ASVspoof2015-eval, VCC2020). All of the three CM systems are trained on the training set of ASVspoof2019LA and validated on its development set.*

EER (%) Evaluation Datasets	CM Systems		
	LCNN [9]	ResNet [10]	ResNet-OC [15]
2019LA-eval	3.25	5.23	2.29
2015-eval	24.55	37.11	26.30
VCC2020	33.78	36.09	41.66

EER degradation across datasets for all three CM systems

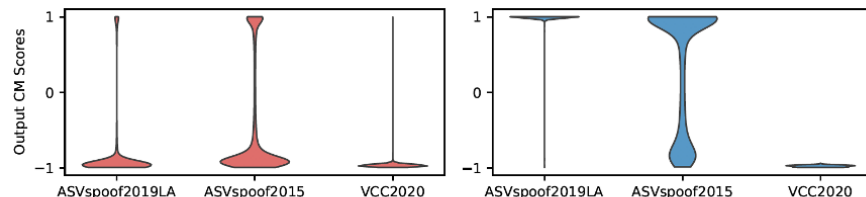


Figure 1: *Score distributions of ResNet-OC method on spoofing attacks (left) and bona fide (right) of cross-dataset evaluation.*

The main cause is some differences in **bona fide speech**, among which, **channel** variation is worth checking.

Channel mismatch

The **average magnitude spectrum** across all **bona fide** utterances of each dataset is different.

We hypothesize that channel mismatch is an important reason for the EER degradation.

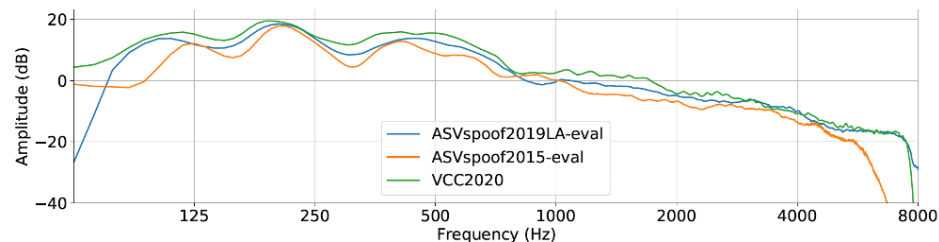


Figure 2: *Average magnitude spectra of bona fide utterances across different datasets.*

Channel-robust training strategies

Augmentation (AUG):

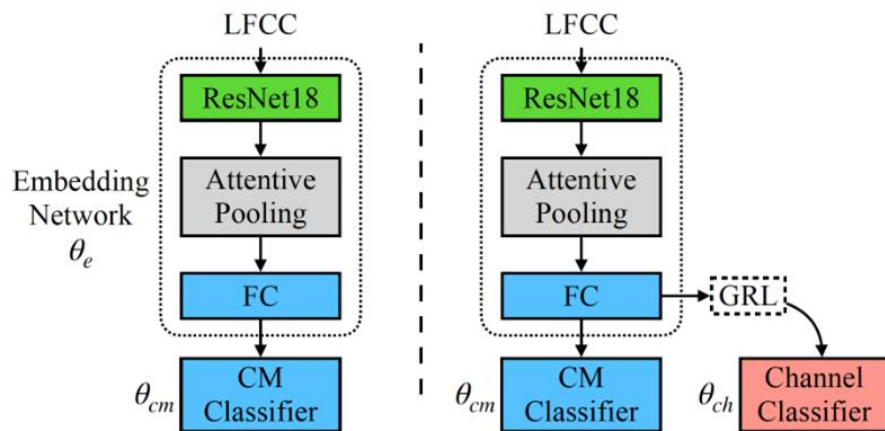
Train with **channel-augmented** data (ASVspoof2019LA and 10 effects of ASVspoof2019LA-Sim)

Multi-Task Augmentation (MT-AUG):

Add a **channel classifier**

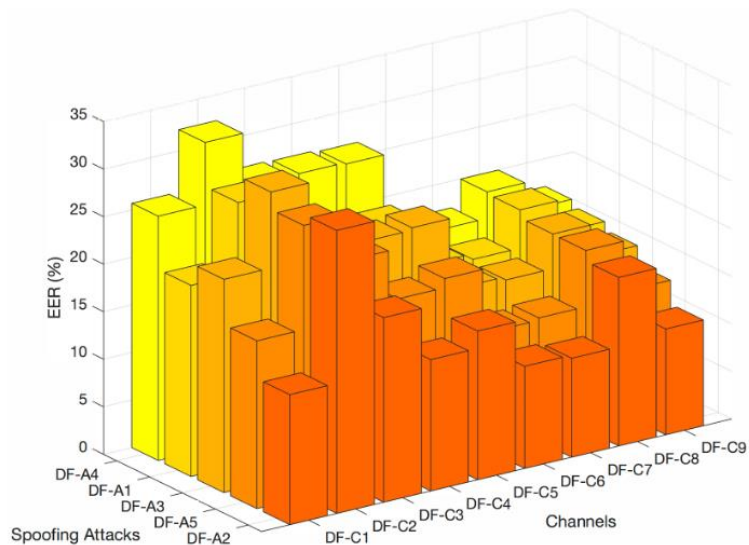
Adversarial Augmentation (ADV-AUG):

Insert a **Gradient Reversal Layer**

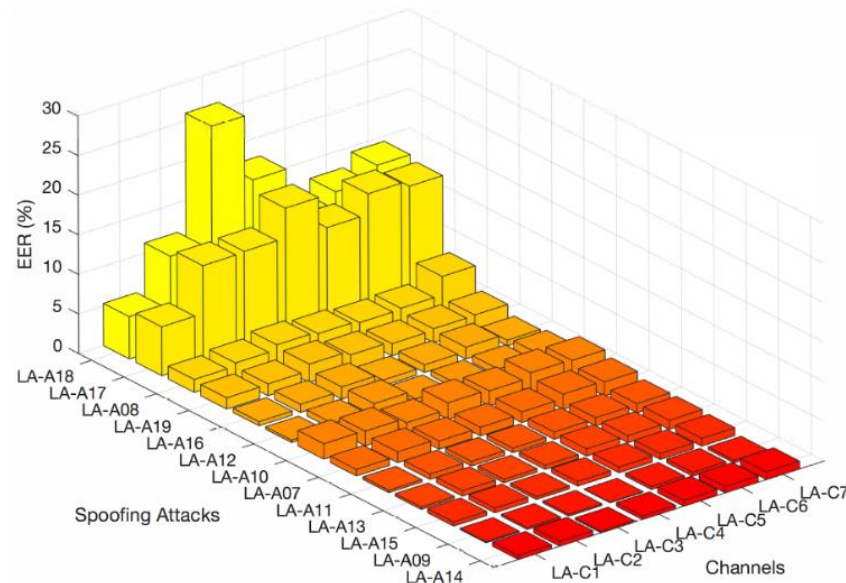


ASVspoof 2021 Challenge

DF: EER 20.33% (rank 15th)



LA: EER 5.46% (rank 10th)



Besides channel augmentation — Unseen channels

Why lack of channel robustness?

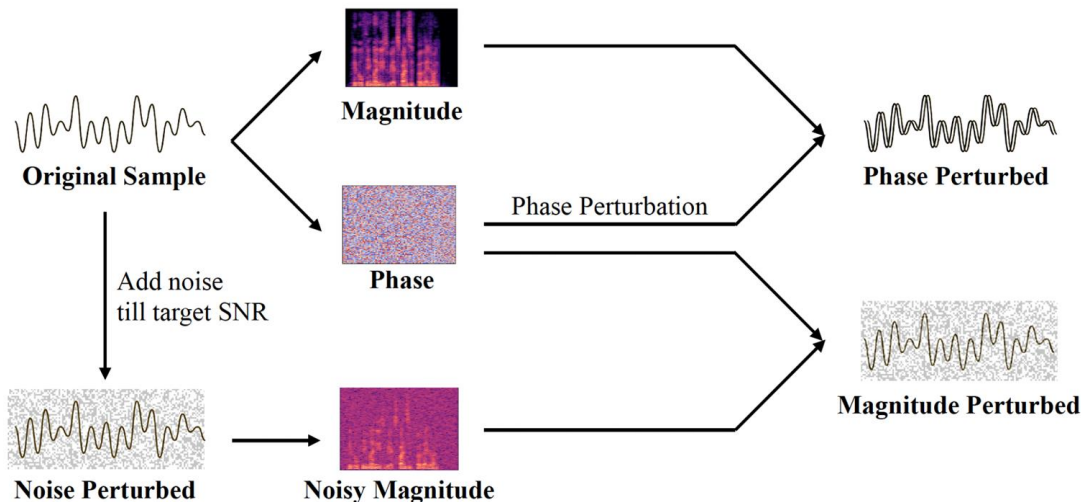
- Communication networks often employ lossy compression codec that encodes only magnitude information, therefore heavily alters phase information.
- State-of-the-art deepfake detection systems takes raw waveforms as input, which rely on phase information.

Research question:

Can phase perturbation improves channel robustness?

Magnitude and phase perturbation

We tested three CM systems on phase and magnitude perturbed data. We conclude that CM systems rely on phase information.



Yongyi Zang, You Zhang, Zhiyao Duan, "Phase perturbation improves channel robustness for speech spoofing countermeasures." in *Proc. INTERSPEECH*, pp. 3162-3166, 2023.

Speech Anti-Spoofing, ECE 277/477 - Computer Audition, Fall 2023

Phase perturbation during training improves channel robustness

EER (%)	C1	C2	C3	C4	C5	C6	C7	Pooled	
No Perturbation	4.68	5.87	14.39	5.75	5.44	7.66	10.26	9.91	
Phase	$\frac{1}{2}\pi$	4.49	6.18	8.68	5.18	5.80	6.35	8.20	7.33
	π	6.72	7.00	7.34	6.41	6.89	6.85	7.30	7.31
	$\frac{3}{2}\pi$	5.52	6.20	10.66	5.21	6.18	7.25	6.21	8.32
	2π	5.68	6.37	9.63	5.42	6.34	7.14	6.39	7.73
Magnitude	10 dB	7.36	8.57	19.88	8.79	8.39	9.53	14.36	14.54
	5 dB	10.40	11.46	30.87	11.96	11.35	14.45	19.09	17.80
	0 dB	17.41	18.23	40.99	18.07	17.98	20.63	26.21	23.64
	-5 dB	23.70	25.05	46.63	24.45	25.00	29.75	34.23	29.87
	-10 dB	34.77	34.55	46.84	34.49	34.95	36.55	38.84	37.40

Yongyi Zang, You Zhang, Zhiyao Duan, "Phase perturbation improves channel robustness for speech spoofing countermeasures." in *Proc. INTERSPEECH*, pp. 3162-3166, 2023.

Speech Anti-Spoofing, ECE 277/477 - Computer Audition, Fall 2023

Joint optimization of ASV and CM

Motivation:

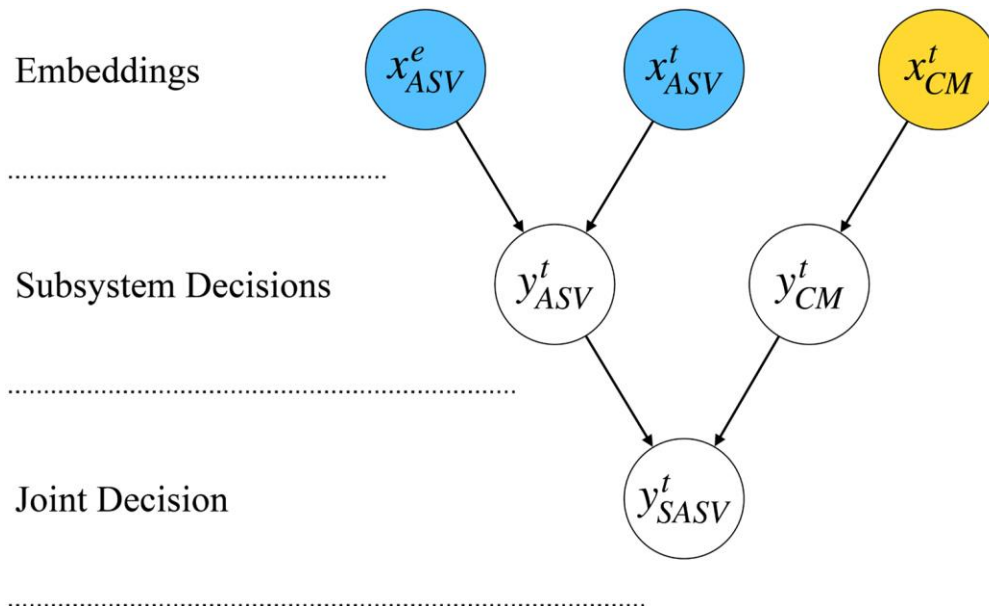
- Standalone CM system might not benefit ASV system.
- **Spoofing aware speaker verification (SASV)** challenge 2022

Evaluation metrics	Target	Non-target	Spoof
SASV-EER	+	-	-
SV-EER	+	-	
SPF-EER	+		-

Research question:

How to **jointly optimize** speaker verification and anti-spoofing?

Probabilistic fusion framework



$$P(y_{SASV}^t = 1 | x_{ASV}^e, x_{ASV}^t, x_{CM}^t) = P(y_{ASV}^t = 1 | x_{ASV}^e, x_{ASV}^t) P(y_{CM}^t = 1 | x_{CM}^t).$$

Product rule with strategies

Direct inference strategy

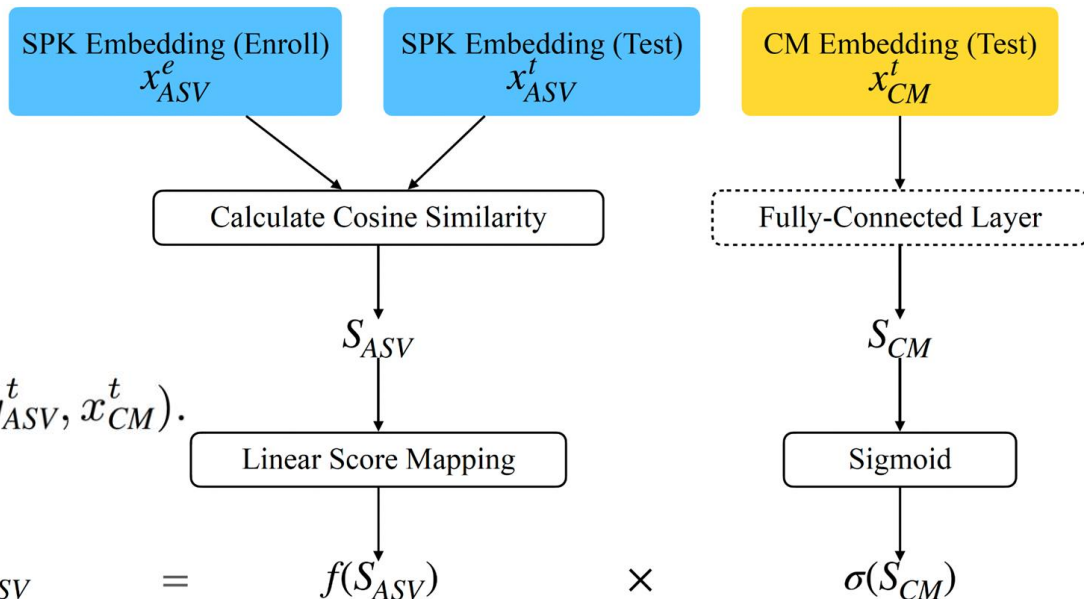
$$\mathcal{S}_{SASV} = f(\mathcal{S}_{ASV}) \times \sigma(\mathcal{S}_{CM})$$

Fine-tuning strategy

$$P(y_{SASV}^t = 1 | x_{ASV}^e, x_{ASV}^t, x_{CM}^t) \\ = P(y_{ASV}^t = 1 | x_{ASV}^e, x_{ASV}^t) P(y_{CM}^t = 1 | y_{ASV}^t, x_{CM}^t).$$

Optimize the FC layer according to \mathcal{S}_{SASV}

$$\mathcal{S}_{SASV} = f(\mathcal{S}_{ASV}) \times \sigma(\mathcal{S}_{CM})$$



Singing voice deepfake detection (SVDD)

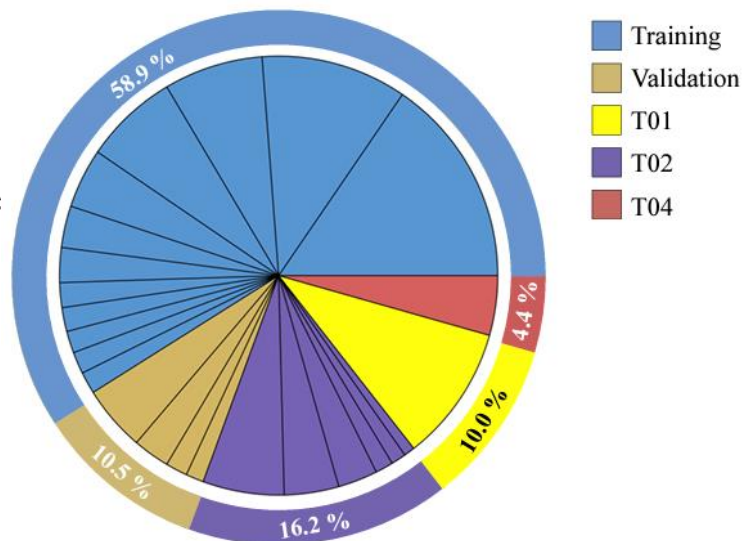


<https://www.youtube.com/watch?v=DhIO1sAni9U>

SingFake dataset

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		1089 / 543
T01	Test set for seen singer Stefanie Sun	1		370 / 1208
T02	Test set for unseen singers	6		1685 / 1006
T03	T02 over 4 communication codecs	6		6740 / 4024
T04	Test set for Persian musical context	17		353 / 166

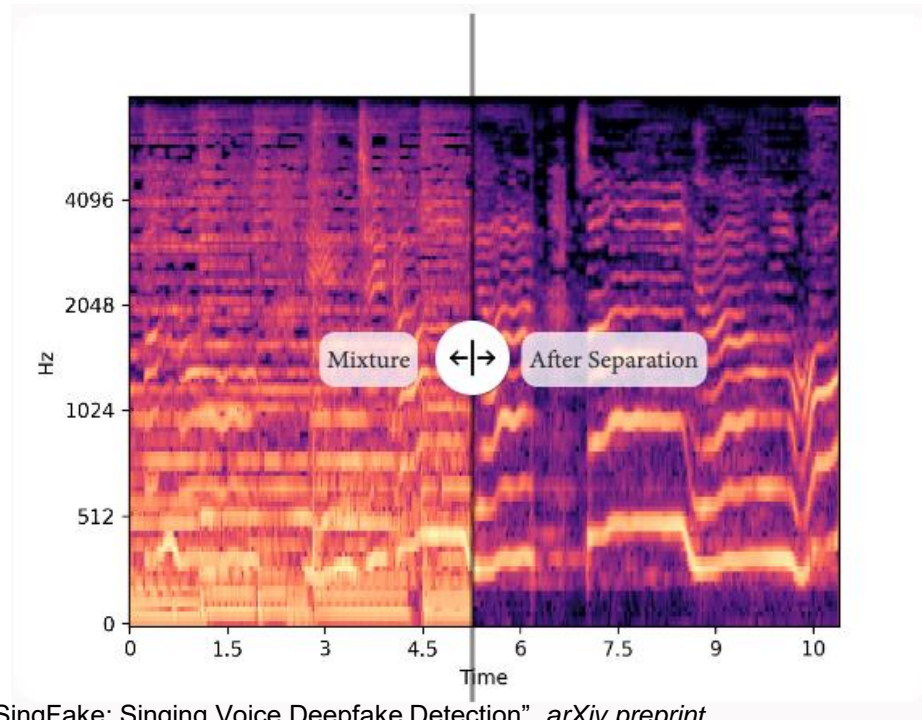


Yongyi Zang*, You Zhang*, Mojtaba Heydari, and Zhiyao Duan. "SingFake: Singing Voice Deepfake Detection", *arXiv preprint arXiv:2309.07525*, 2023. (* equal contribution)

Speech anti-spoofing heavily degrades 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", *arXiv preprint arXiv:2309.07525*, 2023. (* equal contribution)

Performance of training on the SingFake data

Table 3. Evaluation results for SVDD systems on all testing conditions in our SingFake dataset (EER (%))

Method	Setting	Train	T01	T02	T03	T04
AASIST	Mixture	4.10	7.29	11.54	17.29	38.54
	Vocals	3.39	8.37	10.65	13.07	43.94
Spectrogram+ResNet	Mixture	4.97	14.88	22.59	24.15	48.76
	Vocals	5.31	11.86	19.69	21.54	43.94
LFCC+ResNet	Mixture	10.55	21.35	32.40	31.85	50.07
	Vocals	2.90	15.88	22.56	23.62	39.27
Wav2Vec2+AASIST (Joint-finetune)	Mixture	1.57	4.62	8.23	13.62	42.77
	Vocals	1.70	5.39	9.10	10.03	42.19

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", *arXiv preprint arXiv:2309.07525*, 2023. (* equal contribution)

References

- [1] **You Zhang**, Fei Jiang, and Zhiyao Duan, "One-class Learning Towards Synthetic Voice Spoofing Detection", *IEEE Signal Processing Letters*, vol. 28, pp. 937-941, 2021. [[link](#)][[code](#)][[video](#)]
- [2] **You Zhang**, Ge Zhu, Fei Jiang, and Zhiyao Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems", in *Proc. Interspeech*, pp. 4309-4313, 2021. [[link](#)][[code](#)][[video](#)]
- [3] Xinhui Chen*, **You Zhang***, Ge Zhu*, and Zhiyao Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021", in *Proc. ASVspoof 2021 Workshop*, pp. 75-82, 2021. (* equal contribution) [[link](#)][[code](#)][[video](#)]
- [4] **You Zhang**, Fei Jiang, Ge Zhu, Xinhui Chen, and Zhiyao Duan, "Generalizing Voice Presentation Attack Detection to Unseen Synthetic Attacks and Channel Variation", *Handbook of Biometric Anti-spoofing (3rd Ed.)*, Springer, 2023. [[link](#)][[code](#)]
- [5] **You Zhang**, Ge Zhu, and Zhiyao Duan, "A Probabilistic Fusion Framework for Spoofing Aware Speaker Verification", in *Proc. Odyssey*, 2022. [[link](#)][[code](#)]
- [6] Siwen Ding, **You Zhang**, and Zhiyao Duan. "SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. [[link](#)][[code](#)]
- [7] Yongyi Zang, **You Zhang**, and Zhiyao Duan. "Phase Perturbation Improves Channel Robustness for Speech Spoofing Countermeasures", in *Proc. Interspeech*, pp. 3162-3166, 2023. [[link](#)][[code](#)]
- [8] Yongyi Zang*, **You Zhang***, Mojtaba Heydari, and Zhiyao Duan. "SingFake: Singing Voice Deepfake Detection", *arXiv preprint arXiv:2309.07525*, 2023. (* equal contribution) [[link](#)][[code](#)][[webpage](#)]

Future directions

Generalizing to diversified spoofing attacks

- Replay + TTS + VC + Adversarial + PartialSpoof

Robustness

- Additive noise, channel variation, quality of TTS/VC systems (In-the-wild)

Explainability

- The artifacts or the cues that distinguish bona fide from spoofed speech

Visually-informed speech anti-spoofing

- Audio-visual deepfake detection

Takeaways

Introduction to speech anti-spoofing

Generalization ability to unseen synthetic attacks

- One-class learning: OC-Softmax, SAMO

Robustness to channel variation:

- Channel-robust training strategies, phase perturbation

Joint optimization with speaker verification: Probabilistic fusion framework

Beyond speech anti-spoofing: Singing voice detection

Thank you! Questions?