Generalizing Voice Presentation Attack Detection to Unseen Synthetic Attacks

You (Neil) Zhang

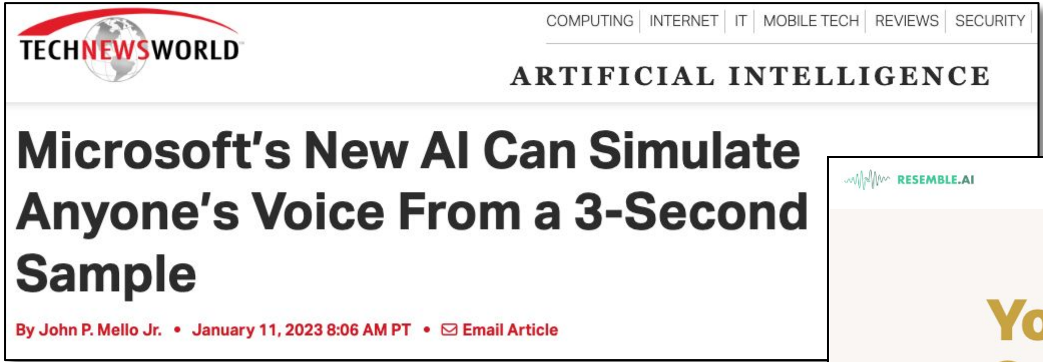
University of Rochester

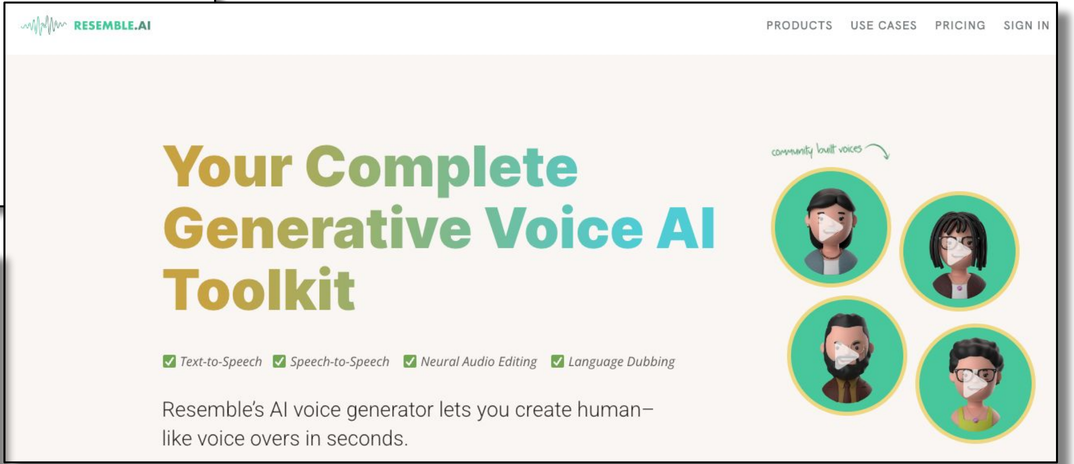
Feb 06, 2023

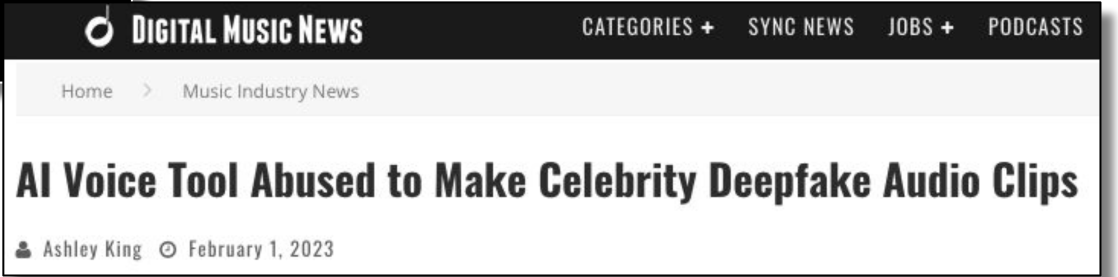
Outline

One-Class Learning Towards Synthetic Voice Spoofing Detection (SPL'21)

SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing (ICASSP'23)



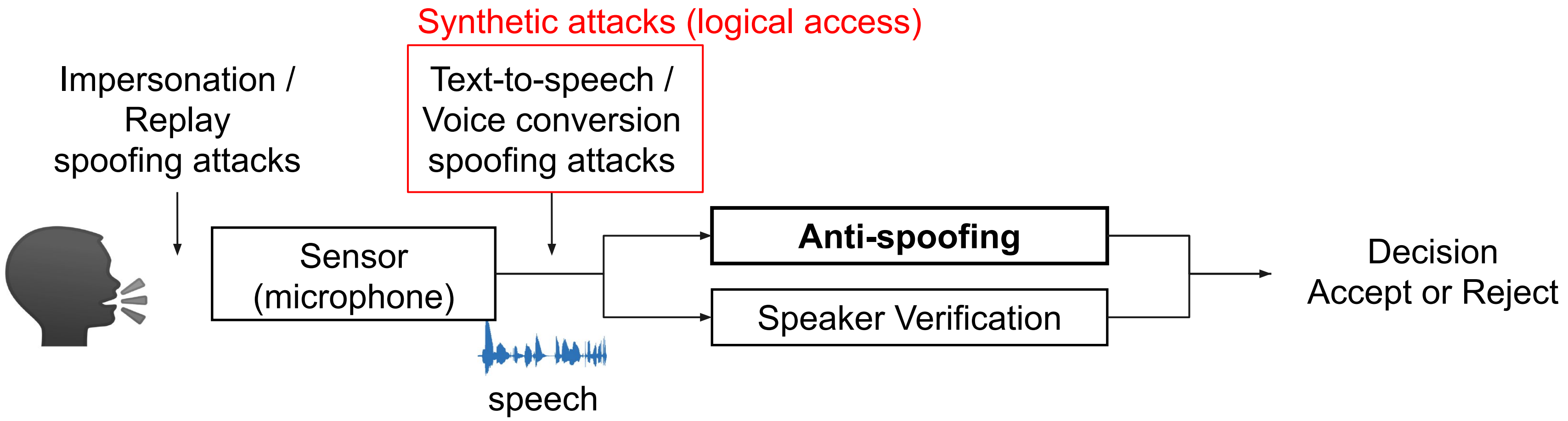




[top left](#)
[top right](#)
[bottom left](#)
[bottom right](#)

Presentation Attack Detection

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



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!

One-Class Learning Towards Synthetic Voice Spoofing Detection

You Zhang, *Fei Jiang, Zhiyao Duan*

University of Rochester, NY, USA

Y. Zhang, F. Jiang and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," in *IEEE Signal Processing Letters*, vol. 28, pp. 937-941, 2021, doi: 10.1109/LSP.2021.3076358.

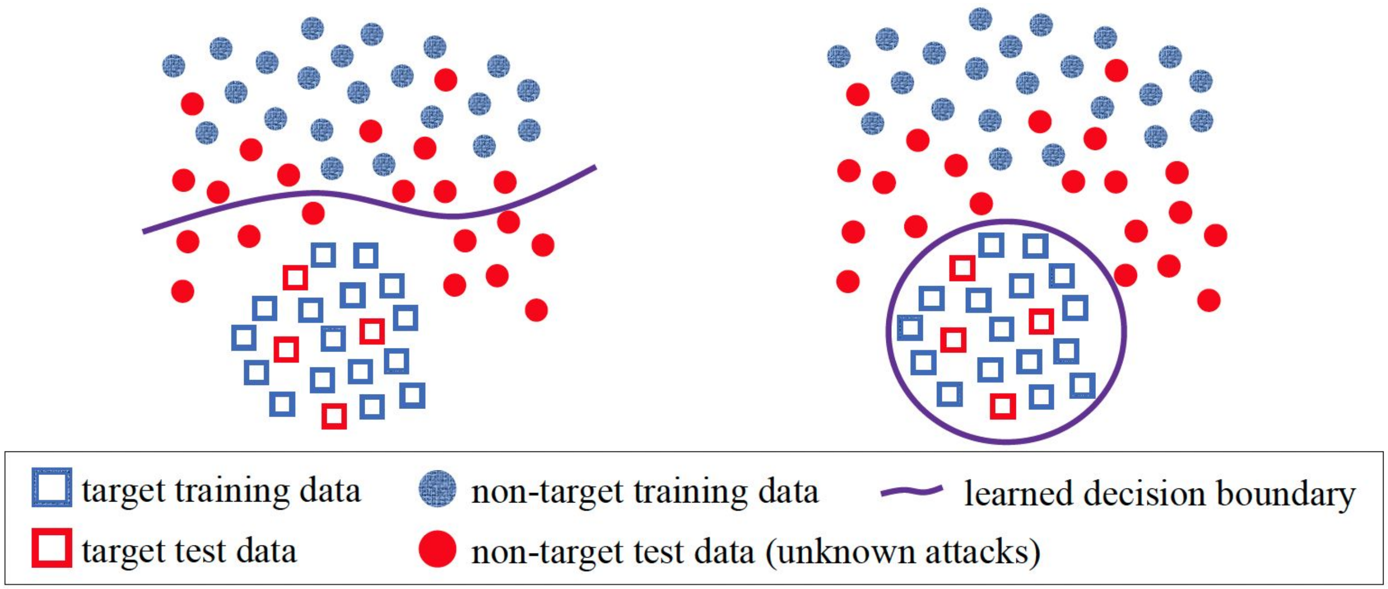
Definition of one-class

- “One-class classification (OCC) algorithms aim to build classification models when the negative class is either absent, poorly sampled or not well defined.”

- “In **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. (to be published)

One-class learning

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

Training: OC-Softmax loss (Proposed)

scale factor, center vector, label

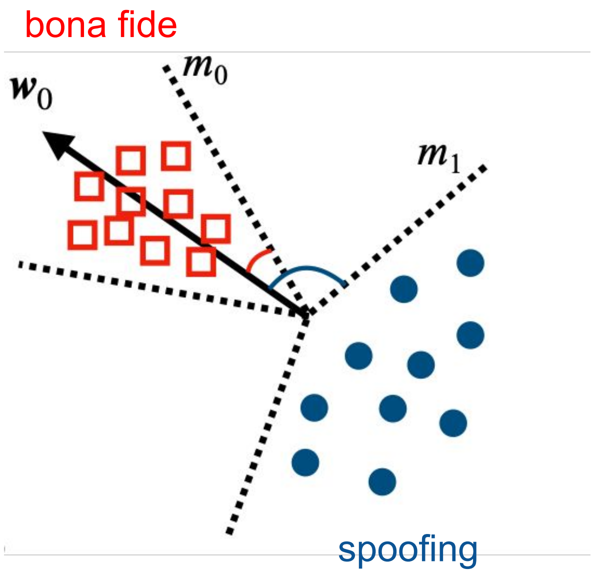
$$\mathcal{L}_{OCS} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1 + e^{\alpha(m_{y_i} - \hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i)(-1)^{y_i}}\right).$$

margin, embedding

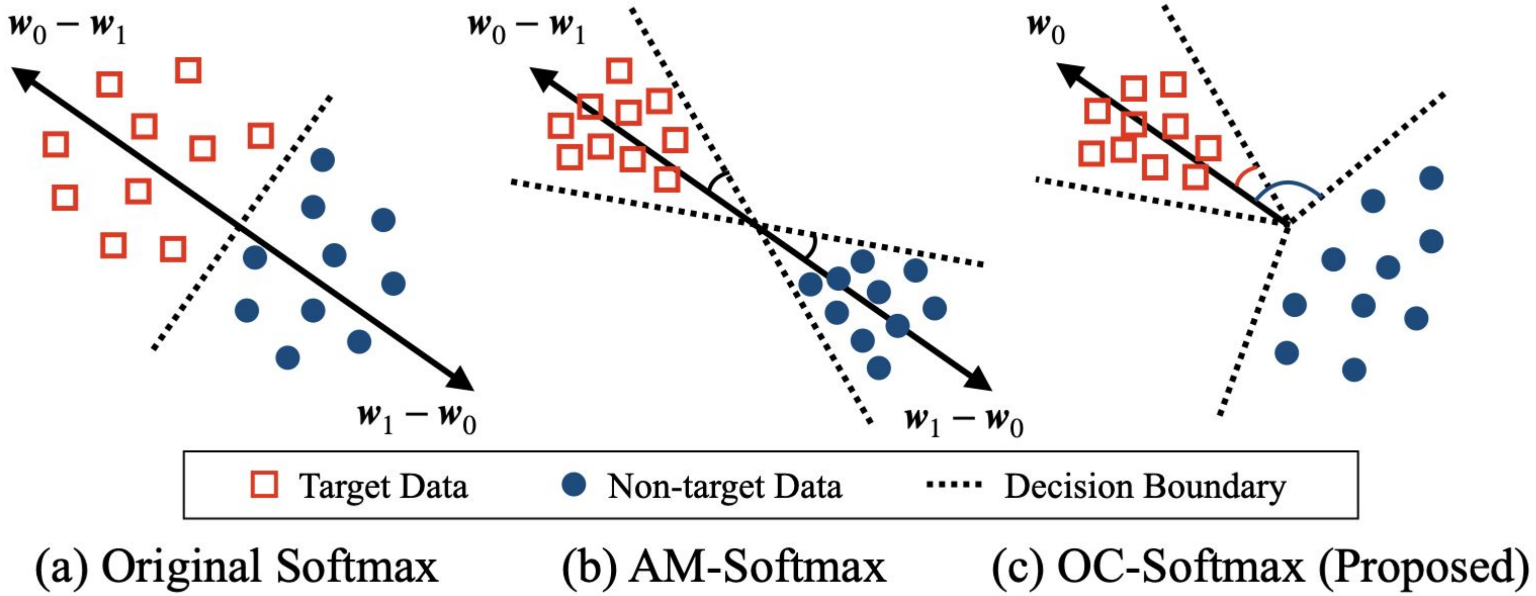
samples

Inference: cosine similarity

$$\mathcal{S}_{OCS} = \hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i.$$



Comparing OC-Softmax with binary classification



Comparing OC-Softmax with binary classification

Softmax:

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^T\boldsymbol{x}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{(\boldsymbol{w}_{1-y_i}-\boldsymbol{w}_{y_i})^T\boldsymbol{x}_i}\right),$$

AM-Softmax:

$$\mathcal{L}_{AMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}}{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}+e^{\alpha\hat{\boldsymbol{w}}_{1-y_i}^T\hat{\boldsymbol{x}}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha\left(m-(\hat{\boldsymbol{w}}_{y_i}-\hat{\boldsymbol{w}}_{1-y_i})^T\hat{\boldsymbol{x}}_i\right)}\right),$$

OC-Softmax:

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

Dataset

ASVspoof 2019 Logical Access (TTS + VC)

- Bona fide speech (VCTK dataset)
- 6 known attacks (appear in the training set)
- 13 unknown attacks (only appear in the evaluation set)

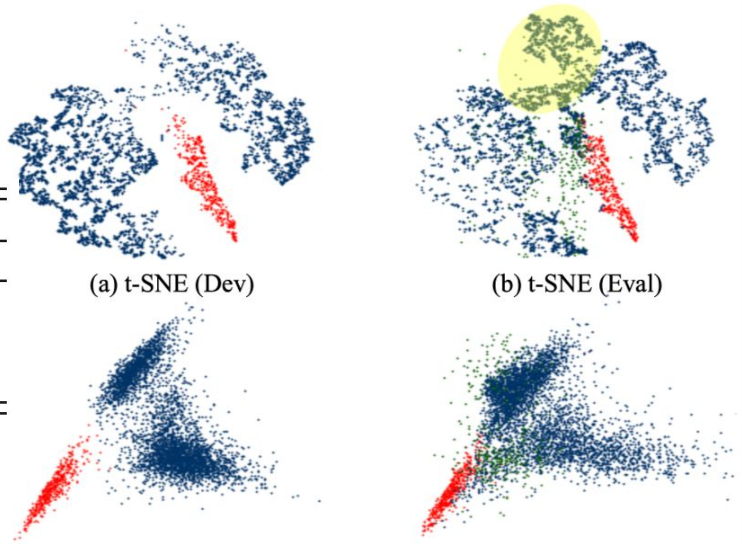
	Bona fide	Spoofed	
	# utterance	# utterance	attacks
Training	2,580	22,800	A01 - A06
Development	2,548	22,296	A01 - A06
Evaluation	7,533	63,882	A07 - A19

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.



(a) t-SNE (Dev) (b) t-SNE (Eval) (c) PCA (Dev) (d) PCA (Eval)

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

Comparison with single systems

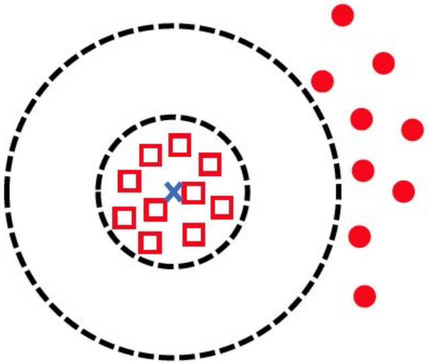
Achieved the state-of-the-art performance

System	EER (%)	min t-DCF
CQCC + GMM [3]	9.57	0.237
LFCC + GMM [3]	8.09	0.212
Chettri et al. [22]	7.66	0.179
Monterio et al. [14]	6.38	0.142
Gomez-Alanis et al. [16]	6.28	-
Aravind et al. [18]	5.32	0.151
Lavrentyeva et al. [21]	4.53	0.103
ResNet + OC-SVM	4.44	0.115
Wu et al. [17]	4.07	0.102
Tak et al. [19]	3.50	0.090
Chen et al. [15]	3.49	0.092
Proposed	**2.19**	**0.059**

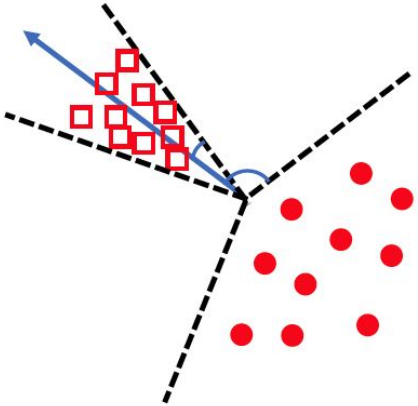
Other one-class losses

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

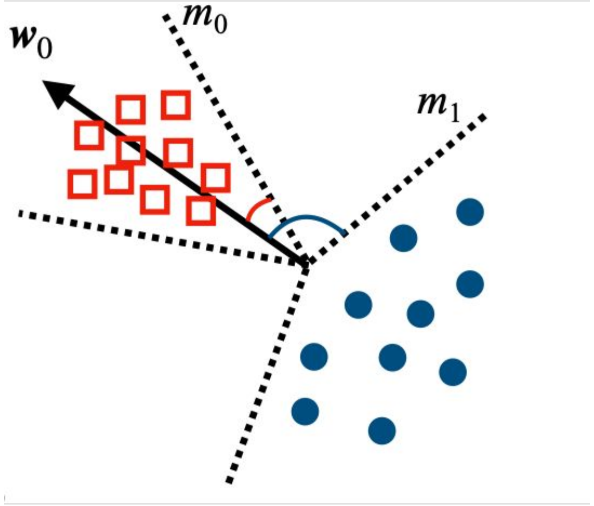


(b) Cosine similarity-based

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. (to be published)

Takeaways

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



SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing



Siwen Ding[1], ***You Zhang***[2], *Zhiyao Duan*[2]

[1]Columbia University, NY, USA
[2]University of Rochester, NY, USA

Siwen Ding, You Zhang, and Zhiyao Duan. “SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing”, *arXiv preprint* arXiv:2211.02718, 2022. (submitted to ICASSP 2023, under review)

Research question

Motivation:

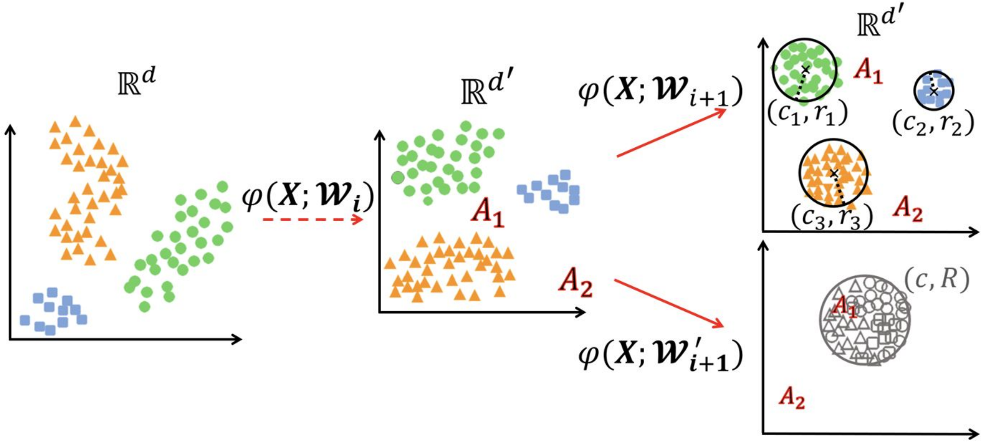
- In our previous work, we compact the embedding space of the bona fide speech into one cluster.
- However, due to **the variety of timbre and speaking traits of different speakers**, the bona fide speech of different speakers naturally forms multiple clusters in the embedding space.

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

Related work

Deep Multi-sphere Support Vector (SDM'20)

If data is naturally multi-cluster, merging them into one cluster could be harmful for detecting anomaly.



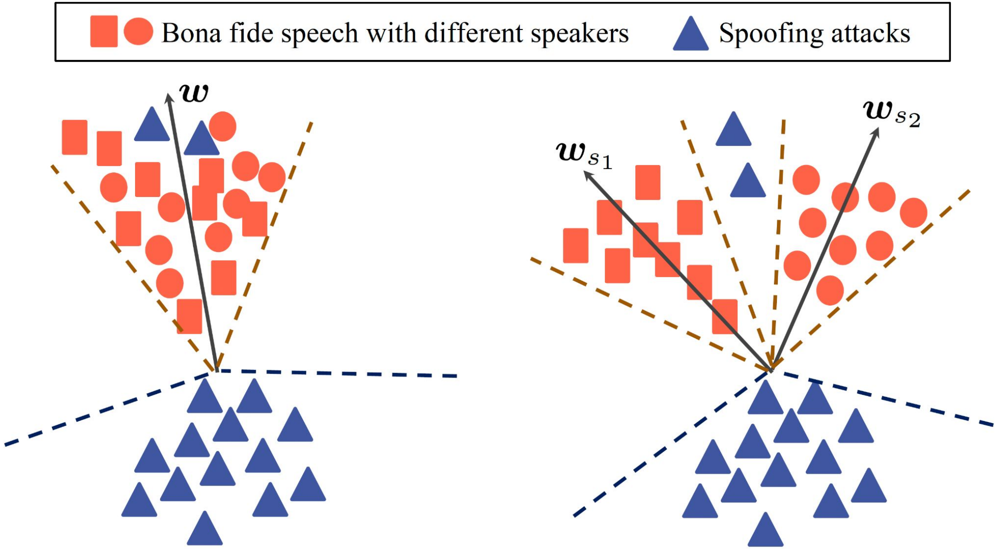
(Figure 2 in Ghafoori et al.)

Ghafoori, Z., & Leckie, C. (2020). Deep multi-sphere support vector data description. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 109-117). Society for Industrial and Applied Mathematics.

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



(a) One-class learning

(b) SAMO

Speaker attractors

Define: a speaker-specific anchor in the embedding space

Compute: average the embeddings of each speaker’s bona fide speech

- **Training:** attract bona fide speech embeddings of the same speaker
- **Inference:** cosine similarity between test utterance and enrolled utterance or attractors of training speakers

speaker attractor

embedding

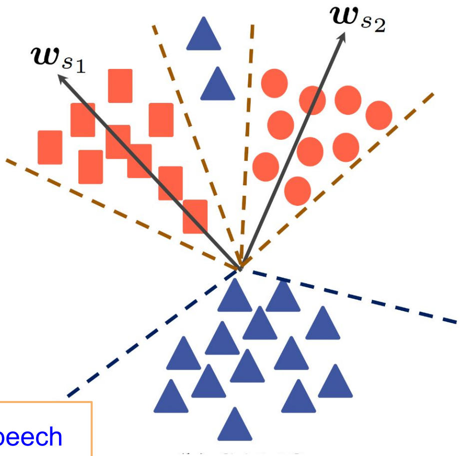
speaker label of *i*-th utterance

$$S_{SAMO} = \begin{cases} \hat{\boldsymbol{w}}_{s_i} \hat{\boldsymbol{x}}_i & \text{if } s_i \text{ is enrolled} \\ \max\limits_{s}(\hat{\boldsymbol{w}}_s \hat{\boldsymbol{x}}_i),\ s \in \mathcal{S}_{train} & \text{otherwise} \end{cases},$$

speakers in the training set

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



scale factor

label

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

samples

margin

where d_i is calculated by

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

bona fide speech

spoofing attacks

SAMO Training algorithm

- Compact the bona fide utterances spoken by the same speaker
- Push away spoofing utterances from all speaker attractors

Algorithm 1: SAMO Training Algorithm

Require: T: Total number of epochs
M: speaker attractor update interval (# epochs)

```
Initialize network F with random weights
Initialize speaker attractors w_s as one-hot vectors
for i ← 1 to T do
    if i mod M = 0 then
        Update w_s as the average bona fide embedding for
          each speaker s ∈ S_train
    end if
    Update F by L_SAMO with mini-batches        ▷ Eq. (3)
end for
return Optimized network F and speaker attractors w_s
```

Dataset

ASVspoof2019 LA target-only portion

- Same train/dev/eval splits
- Keep only the target speakers in dev/eval sets

Table 1. Summary of the dataset used in our experiments, which is the target-only portion of the ASVspoof2019 LA corpus.

Partition	# Speakers	# Enrollment Utts	Bona Fide	Spoofing Attacks	
			# Utts	# Utts	Attack Types
Train	20	-	2580	22800	A01∼A06
Dev	10	142	1484	22296	A01∼A06
Eval	48	696	5370	63882	A07∼A19

Comparison with state-of-the-art methods

Table 2. Comparison of our proposed SAMO with Softmax and OC-Softmax on the target-only portion of the ASVspoof2019 LA evaluation set. All the systems use AASIST [9] backbone. The average (best) results across 3 trials with random training seeds are shown.

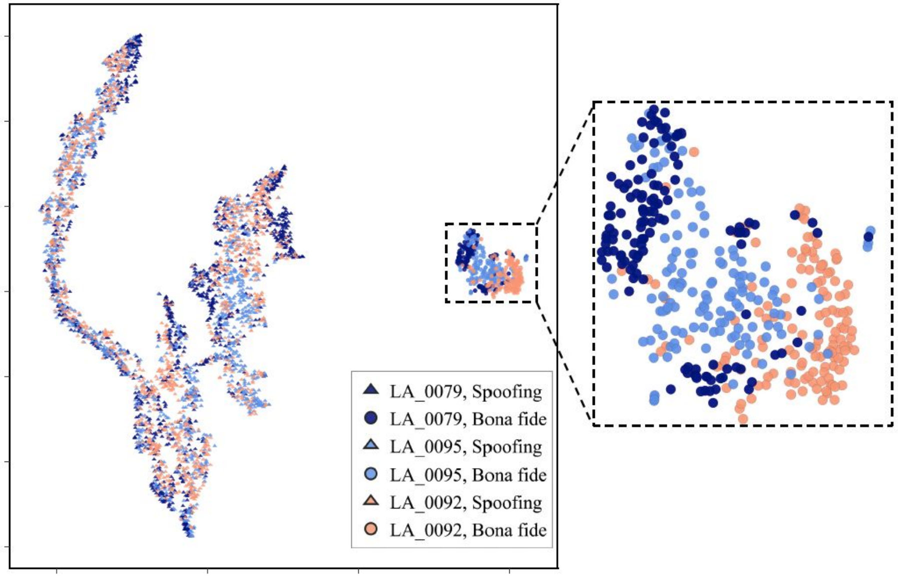
Method	EER(%)	min t-DCF
Softmax	1.74 (1.25)	0.0583 (0.0425)
OC-Softmax	1.25 (1.17)	0.0415 (0.0393)
SAMO (test w/o enrollment)	1.09 (0.91)	0.0363 (0.0306)
SAMO (test w/ enrollment)	**1.08 (0.88)**	**0.0356 (0.0291)**

SAMO further improves the performance, indicating the advantage brought by the multi-center modeling of bona fide speech.

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.



Ablation studies

Effects of leveraging bona fide speech data for speaker attractors

Effects of the speaker attractor update interval *M* (# epochs)

Table 3. Ablation experiments for SAMO. Results of test scenarios without and with enrollment data are both presented.

Setup	Configuration	Test w/o enroll (w/ enroll) EER(%)	Test w/o enroll (w/ enroll) min t-DCF
1	SAMO	**1.09 (1.08)**	**0.0363 (0.0356)**
2	one-hot and fixed attractors	47.93 (49.50)	0.9999 (0.9980)
3	w/o speaker attractor update	1.54 (1.55)	0.0504 (0.0503)
4	update every epoch (M=1)	1.33 (1.33)	0.0442 (0.0437)
5	update every 10 epochs (M=10)	2.36 (2.77)	0.0792 (0.0868)

Future work

With a larger variety of speakers in the training set, the benefit of

SAMO could be demonstrated even more since the speaker attractors

will better represent the bona fide embedding space.

Extend the SAMO idea to model other speech attributes, such as device and codec variations.

Other works on anti-spoofing

- Channel Robustness
  - **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]
  - 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]
- Joint Optimization with ASV
  - **You Zhang**, Ge Zhu, and Zhiyao Duan, "A Probabilistic Fusion Framework for Spoofing Aware Speaker Verification", in *Proc. Odyssey*, 2022. [link][code]

Take-aways

One-class learning could improve the **generalization ability** of anti-spoofing system against unseen spoofing attacks.

It aims to **compact the bona fide speech representation** in the embedding space, and push away spoofing attacks by a certain margin.

Speaker attacker multi-center one-class learning could further improve the generalization ability while **clustering** bona fide speech around a number of speaker attractors and **pushing away** spoofing attacks from all the attractors.

Thank you! Questions?