One-Class Learning Towards Synthetic Voice Spoofing Detection

ICASSP 2022 Singapore, China, Online

You Zhang, Fei Jiang, Zhiyao Duan

University of Rochester

Audio Information Research Laboratory

ABSTRACT

Automatic Speaker Verification (ASV) systems are vulnerable to text-to-speech (TTS), and voice conversion (VC) attacks.
Voice anti-spoofing is developed to improve the reliability of speaker verification systems against such spoofing attacks.
The fast development of speech synthesis are posing increasingly more threat.

Main issue of voice anti-spoofing systems:
- Generalization to **unseen synthetic attacks**

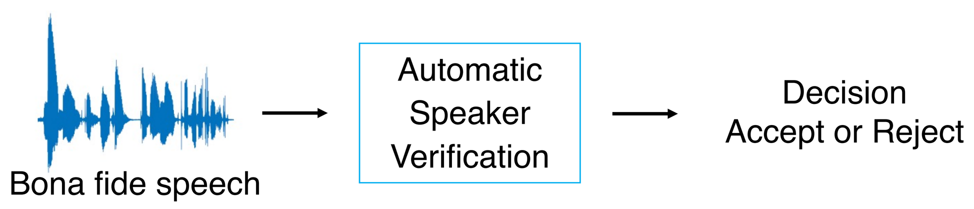
Proposed solution:
- One-Class Learning

Results:
- EER 2.19%, outperforming all single systems

Keywords:
Voice Anti-spoofing, One-Class Learning, Generalization Ability, Feature Learning, Speaker Verification, Voice Biometrics

BACKGROUND

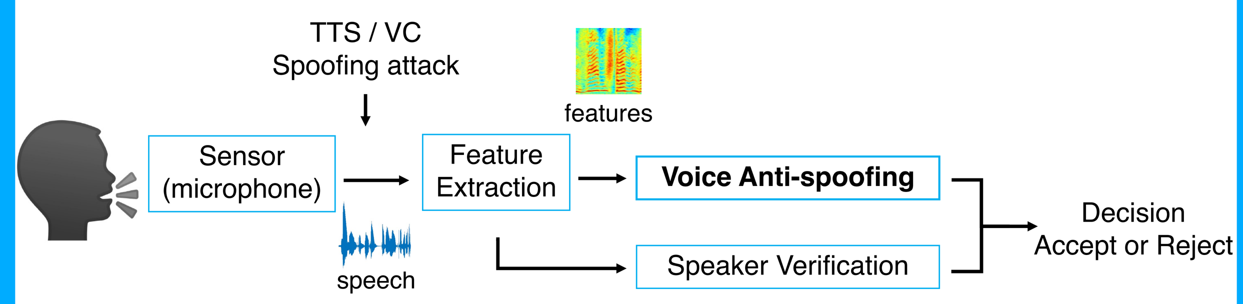
Automatic Speaker Verification (ASV)
Verify the identity of a speaker



Logical Access (LA) Spoofing Attacks

- **Text-to-speech (TTS)**
  ➢ Convert written text into audio with speech synthesis
- **Voice Conversion (VC)**
  ➢ Convert speech from source to a target speaker

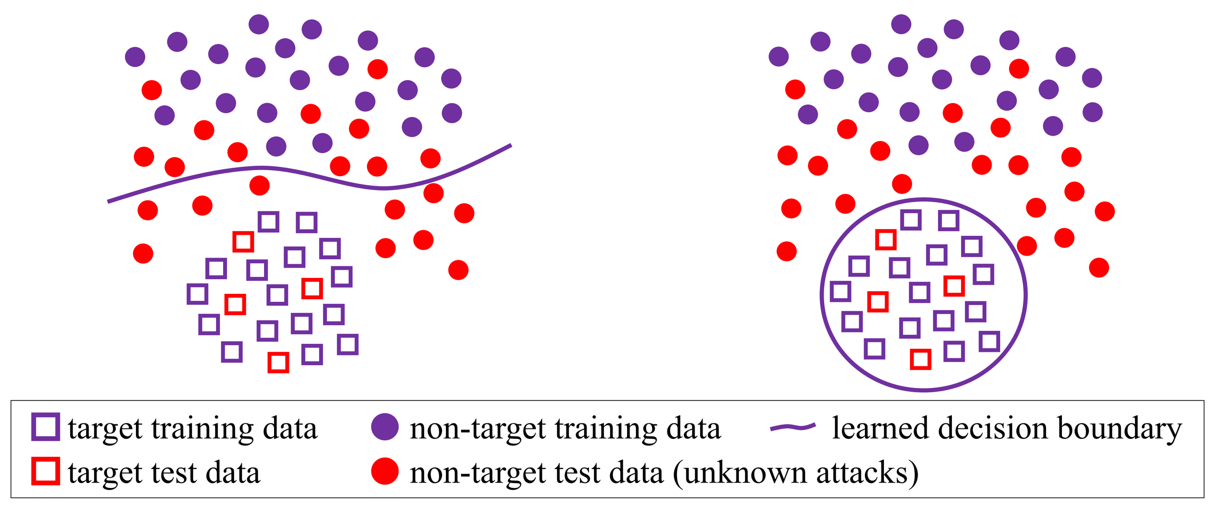
Voice Anti-spoofing / Spoofing Countermeasure (CM)
Detect spoofing attacks



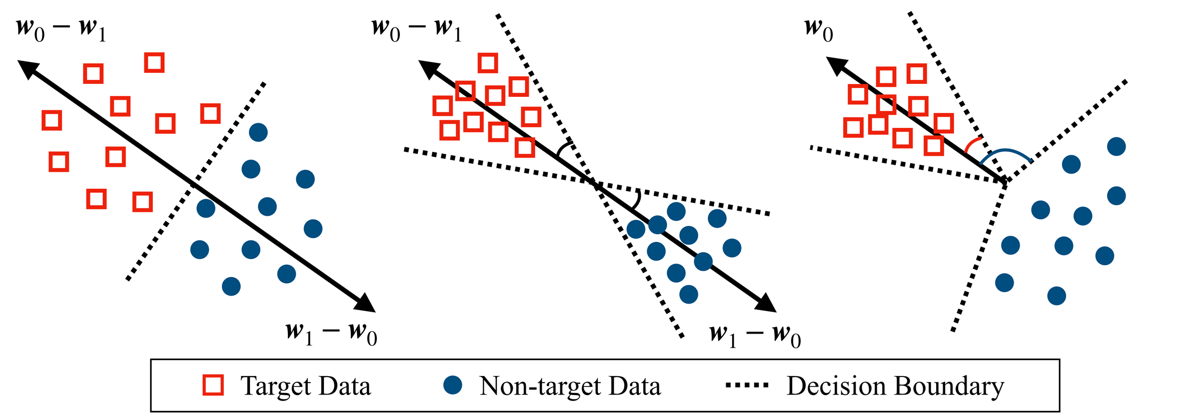
METHOD

One-Class Learning

The **distribution mismatch** between training and test for the spoofing attacks class, makes the problem a good fit for **one-class classification** [1].



We propose a loss function called **one-class Softmax** (OC-Softmax) to learn a feature space in which the **bona fide** speech embeddings have a **compact boundary** while **spoofing** data are kept away from the bona fide data by **a certain margin**.



(a) Original Softmax (b) AM-Softmax (c) OC-Softmax (Proposed)

The proposed **OC-Softmax** can be formulated as:

$$\mathcal{L}_{OCS} = \frac{1}{\underset{\text{\# samples}}{N}} \sum_{i=1}^{N} \log\left(1 + e^{\overset{\text{scale factor}}{\alpha}(\underset{\text{margin}}{m_{y_i}} - \overset{\text{center vector}}{\hat{\boldsymbol{w}}_0} \underset{\text{embedding}}{\hat{\boldsymbol{x}}_i})(-1)^{\overset{\text{label}}{y_i}}}\right).$$

Preliminary: Binary classification loss functions
Vanilla Softmax:

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

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

Additive Margin Softmax (AM-Softmax) [2]:

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

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

RESULTS

Dataset: **ASVspoof 2019 LA**

	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

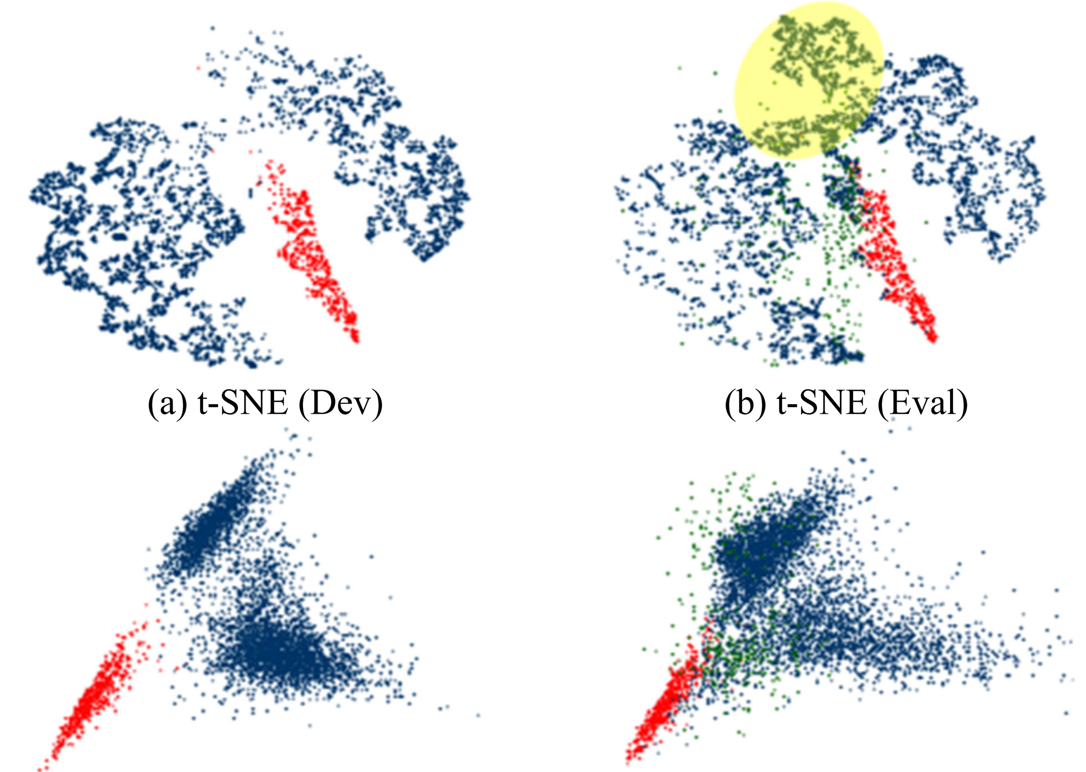
We evaluate the performance with equal error rate (EER) and minimum tandem detection cost function (t-DCF).

Comparing with binary classification loss functions:

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**

Our proposed OC-Softmax achieves the best results.

Feature embedding **visualization**:



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

The bona fide speech has the same distribution in both sets, while the spoofing attacks show different distributions.
The figure verifies our problem formulation and shows the effectiveness of our proposed OC-Softmax.

Comparing with other existing single systems:

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**

CONCLUSIONS

- 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.
- One-class learning could **improve** the **generalization ability** of anti-spoofing system against **unknown spoofing attacks**.
- The proposed system trained with **OC-Softmax** outperforms all existing single systems.

ACKNOWLEDGMENTS

NSF

VOICE BIOMETRICS GROUP

REFERENCES

[1] Shehroz S. Khan and Michael G. Madden, "A survey of recent trends in one class classification," in *Proc. Irish Conference on Artificial Intelligence and Cognitive Science*, 2009, pp. 188–197.
[2] Feng Wang, Jian Cheng, Weiyang Liu, Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

CITATION

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.

CODE

FOLLOW-UP WORKS

You Zhang, Ge Zhu, Fei Jiang, Zhiyao Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems", in *Proc. Interspeech*, 2021, pp. 4309-4313.

Xinhui Chen, You Zhang, Ge Zhu, Zhiyao Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021", in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 75-82.