Overview

Voiceprint recognition (VPR):

- recognizing a person by speeches
- applications: financial transactions, app logging in
- once broken → property damage, information leakage

Research goal: building secure and robust VPR systems

Various attack vectors against VPR:

- Conventional: replay, speech synthesis, voice conversion
- Recent: adversarial attack, backdoor attack

1st Project: **Systematic**, **practical**, and **physical** adversarial attack against VPR systems

➢ **Systematic: applied to arbitrary source-to-target on all tasks**

- VPR involves three tasks: open-set identification (OSI), close-set identification (CSI), and speaker verification (SV).
- Source/targeted speakers could be enrolled speakers or imposters, depending on the attack goals (e.g., unauthorized access, DoS).
- Cross entropy and margin loss cannot be used for all 10 settings.

Task	ID	Source Speaker	Target	Goals
OSI	C1	enrolled speaker s	enrolled speaker $t \neq s$	S1-1, S3-2, S4-2
	C2	unenrolled speaker	enrolled speaker	S1-2
	C3	enrolled speaker	imposter	S2-1, S2-2
	C4	enrolled speaker	untargeted	S2-3, S3-1, S4-1, S5-1
	C5	unenrolled speaker	untargeted	S1-3
CSI	C6	enrolled speaker s	enrolled speaker $t \neq s$	S1-1, S3-2, S4-2
	C7	unenrolled speaker	enrolled speaker	S1-2
	C8	enrolled speaker	untargeted	S2-3, S3-1, S4-1, S5-1
SV	C9	enrolled speaker	imposter	S2-1, S2-2, S2-3 S3-1, S4-1, S5-1
	C10	unenrolled speaker	enrolled speaker	S1-2, S1-3

The attack settings on VPR. S1-1~S5-1 denote different attack goals.

Design various loss functions for each setting. Findings:

✓ In some settings, cross entropy and margin loss perform worse than newly designed loss, e.g., 0% attack success rate of cross entropy for the untargeted attack on OSI.

✓ The optimal loss varies with the setting.

✓ OSI is more difficult to attack than CSI.

➢ **Practical: black-box**

- OSI and SV predict based on a threshold θ.
- Attack succeeds only when the score exceeds θ.
- θ is unknown to attackers in black-box.

Proposing threshold estimation algorithm f:

✓ f outputs the estimated threshold $\tilde{\theta}$.

✓ $\tilde{\theta} > \theta$: attack succeeds

✓ $\tilde{\theta} \approx \theta$: reduce attack overhead

TABLE IV: Results of threshold estimation

ivector			GMM		
θ	$\acute{\theta}$	Time (s)	θ	$\acute{\theta}$	Time (s)
1.45	**1.47**	**628**	**0.091**	**0.0936**	**157**
1.57	1.60	671	0.094	0.0957	260
1.62	1.64	686	0.106	0.1072	269
1.73	1.75	750	0.113	0.1141	289
1.84	1.87	804	0.119	0.1193	314

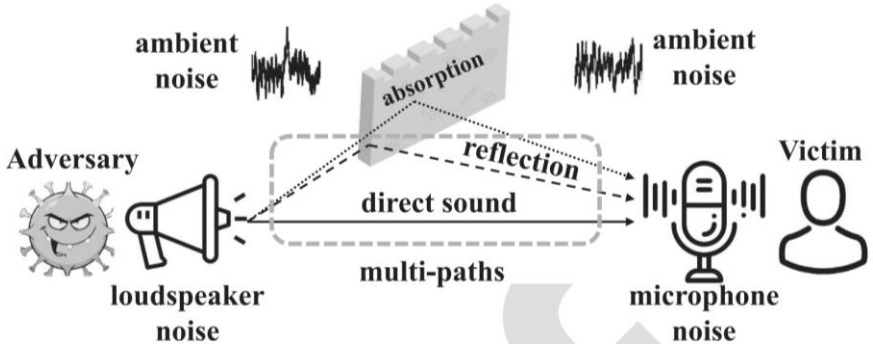
1st Project: **Systematic**, **practical**, and **physical** adversarial attack against VPR systems

➢ **Physical: effective when played over-the-air**

- Adversarial voices become ineffective when being played over-the-air

Robust over-the-air attack

✓ Study the distortions occurring during over-the-air



The acoustic model of over-the-air attack

✓ Design transformation functions to simulate the distortion

✓ Incorporate them during the crafting process

☐ Results: successfully attack commercial VPR Microsoft Azure and TalentedSoft in both digital and physical worlds.

☐ Publications:

1. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. **Guangke Chen**, Sen Chen, Lingling Fan, Xiaoning Du, Fu Song, Yang Liu. In S&P (Oakland) 2021.

2. AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. **Guangke Chen**, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Yang Liu. IEEE Transactions on Dependable and Secure Computing (TDSC).

2nd Project: Securing VPR against adversarial attacks

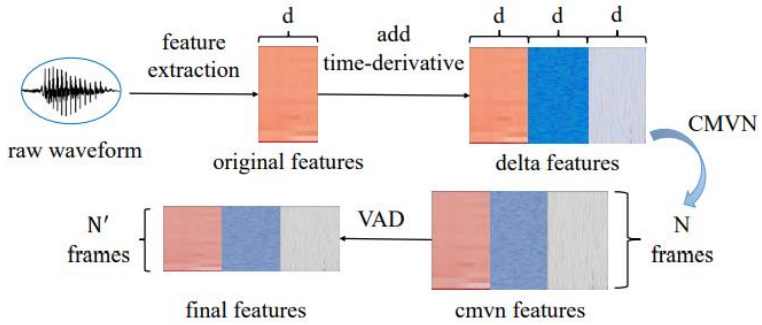
➢ **Study of existing defenses against adaptive attack**

Findings:

- Variable bit rate (VBR) speech compressions have different performance from constant bit rate (CBR) ones.
- VBR is more resilient against black-box attack.
- CBR cannot be evaded by BPDA in white-box attack .
- Evading VBR incurs negligible increase in perturbation.
- Randomized defenses remain resilient against adaptive black-box attacks.

➢ **Feature-level defense: Feature Compression (FeCo)**

- Different from vision systems, VPR still heavily relies on hand-crafted acoustic features.



The typical acoustic feature extraction module of VPR

- Existing defenses overlook this difference and only mitigate the adversarial perturbation at the waveform-level.
- We propose a feature-level approach that mitigates adversarial perturbation by compressing the acoustic features.

2nd Project: Securing VPR against adv. attack

Feature Compression (FeCo)

- Motivation: large redundancy between adjacent frames of an audio
- Compressing N frames to K frames (K<<N) can
  - ✓ disrupt the added perturbation
  - ✓ reduce the search space of the attacker
  - ✓ incur little impact on benign example

Algorithm 1 FeCo

Input: feature matrix $\mathcal{M} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$; cluster ratio $0 < cl_r < 1$; cluster oracle $\mathcal{O}$ = kmeans or warped-kmeans
Output: compressed feature matrix $\mathcal{M}'$

1: $K \leftarrow \lceil N \times cl_r \rceil$ ▷ K = number of clusters
2: $[b_1, \cdots, b_N] \leftarrow \mathcal{O}(\mathcal{M}, K)$ ▷ $\mathbf{a}_i$ is assigned to the b_i-th cluster
3: **for** $(i = 1; i \leq K; i++)$ **do**
4: $C_i \leftarrow \{\mathbf{a}_k \mid b_k = i\}$ ▷ compute the i-th cluster
5: $\mathbf{m}_i \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{a} \in C_i} \mathbf{a}$ ▷ compute the representative vector
6: $\mathcal{M}' \leftarrow [\mathbf{m}_1, \cdots, \mathbf{m}_K]$ ▷ concatenate the representative vectors
7: **return** $\mathcal{M}'$

- Under two tailored adaptive attacks, when combined with adversarial training, FeCo:
  - ✓ increases the robustness accuracy by **13.62%**
  - ✓ increases attack cost by two orders of magnitude

TABLE 7: Results (A_a, SNR, PESQ) on Standard, Vanilla-AdvT, and AdvT+Transformation

	R1 Score	A_b	L_∞ white-box attacks					L_2 white-box attacks			black-box attacks		
			FGSM	PGD-10	PGD-100	CW_∞-10	CW_∞-100	CW_2-1			FAKEBOB	SirenAttack	Kenansville
			A_a	A_a	A_a	A_a	A_a	A_a	SNR	PESQ	A_a	A_a	A_a
Standard	6.54	99.06%	19.61%	0%	0%	0%	0%	0%	55.87	4.47	0.35%	0.38%	0.03%
Vanilla-AdvT	61.48	95.67%	75.20%	58.19%	53.83%	58.95%	55.56%	0%	36.96	3.91	85.63%	86.73%	0.03%
AdvT+QT	67.68	95.74%	88.19%	72.12%	64.08%	73.20%	65.43%	0.7%	46.59	3.86	79.84%	88.81%	0.31%
AdvT+AT	71.11	95.57%	71.10%	61.10%	59.22%	61.47%	59.89%	9.3%	36.21	3.90	94.69%	95.39%	39.80%
AdvT+AS	58.35	93.59%	82.72%	53.83%	43.12%	54.10%	45.24%	0%	35.46	3.45	83.55%	87.08%	0.03%
AdvT+MS	54.66	92.76%	65.85%	49.77%	44.13%	50.33%	46.66%	0%	37.85	3.66	76.38%	77.24%	0.17%
AdvT+DS	56.41	95.32%	70.14%	51.44%	44.06%	52.13%	45.41%	0%	36.23	3.91	79.91%	85.04%	0.69%
AdvT+FeCo-o(k)	88.03	97.81%	95.06%	93.65%	85.50%	94.14%	86.11%	96.0%	29.89	2.53	98.08%	97.42%	39.94%

: The top-1 is highlighted in blue excluding Standard. The results in green background indicate that the transformation worsens adversarial training

The comparison between waveform- and feature-level defenses

- **SpeakerGuard Platform**
  - A fully Pytorch-written security analysis platform for VPR
  - Mainstream VPRs, voice datasets, white- and black-box attacks
  - Widely-used evasion techniques for adaptive attacks
  - Evaluation metrics of listening, diverse audio defense solutions
- ☐ Publications:

Towards Understanding and Mitigating Audio Adversarial Examples for Speaker Recognition. Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, Jiashui Wang. IEEE Transactions on Dependable and Secure Computing (TDSC).