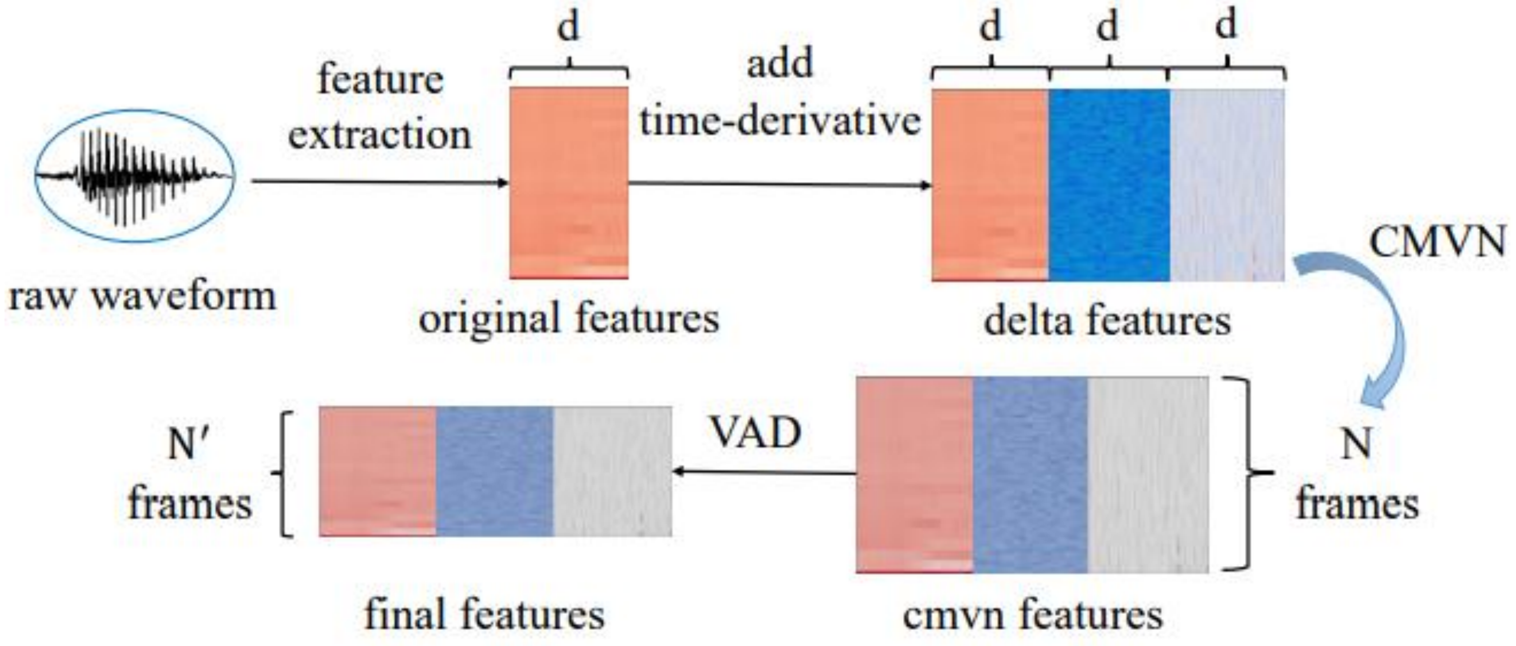
Defense: feature-level transformation-based defense

Categories: robust training; *transformation*; detection

Motivations:
- relies on handcrafted acoustic features
- perturbation added to waveforms will propagate to acoustic features
- existing defenses operate at the waveform-level

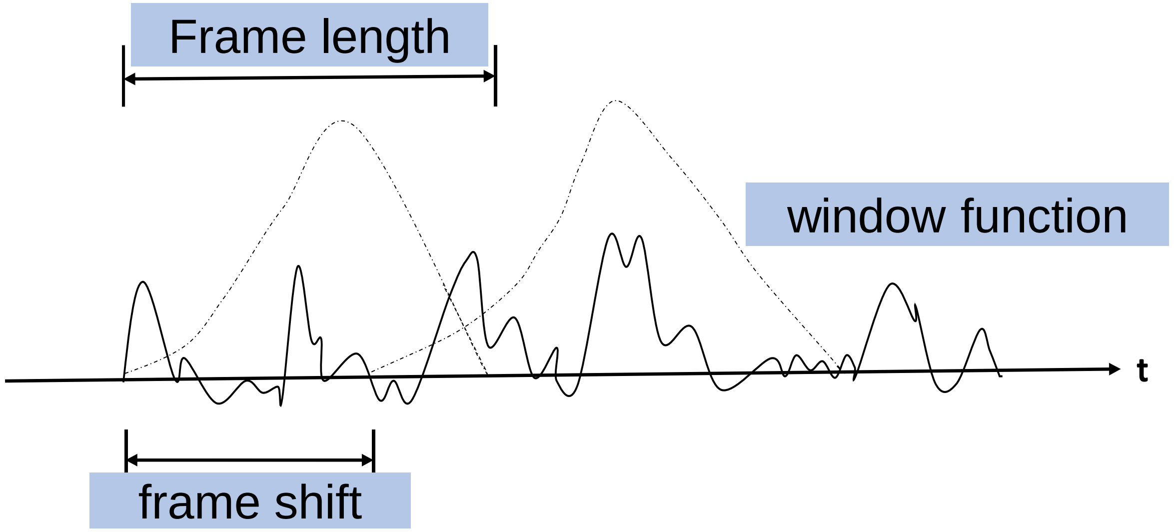


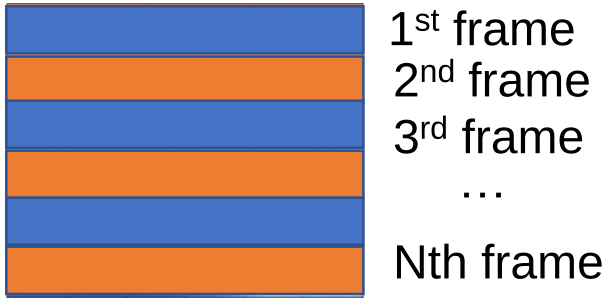
Defense: feature-level transformation-based defense

Our defense: Feature Compression (FeCo)

Motivation:

- large redundancy between adjacent frames
- Compressing N frames to K frames (K<<N) can
  - disrupt perturbation
  - reduce search space of attackers
  - incur little impact on benign examples





acoustic feature

Defense: feature-level transformation-based defense

Our defense: Feature Compression (FeCo)

Method:
Feature compression by clustering methods

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: $\quad C_i \leftarrow \{\mathbf{a}_k \mid b_k = i\}$ ▷ compute the i-th cluster
5: $\quad \mathbf{m}_i \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{a} \in C_i} \mathbf{a}$ ▷ compute the representative vector
6: $\mathcal{M}' \leftarrow [\mathbf{m}_1, \cdots, \mathbf{m}_K]$ ▷ concatenate the representative vectors
7: **return** $\mathcal{M}'$

---

clustering methods:

- rely on temporal dependency (e.g., ivector-PLDA): kmeans
- not rely on temporal dependency (e.g., DeepSpeaker): warped-kmeans

Defense: experiments against non-adaptive attacks

non-adaptive attacks:
unaware and not consider defense when crafting adversarial examples

accuracy on normal voices A_b
accuracy on adversarial voices A_a
trade-off $R_1 = \frac{2 \times A_b \times A_a}{A_b + A_a}$

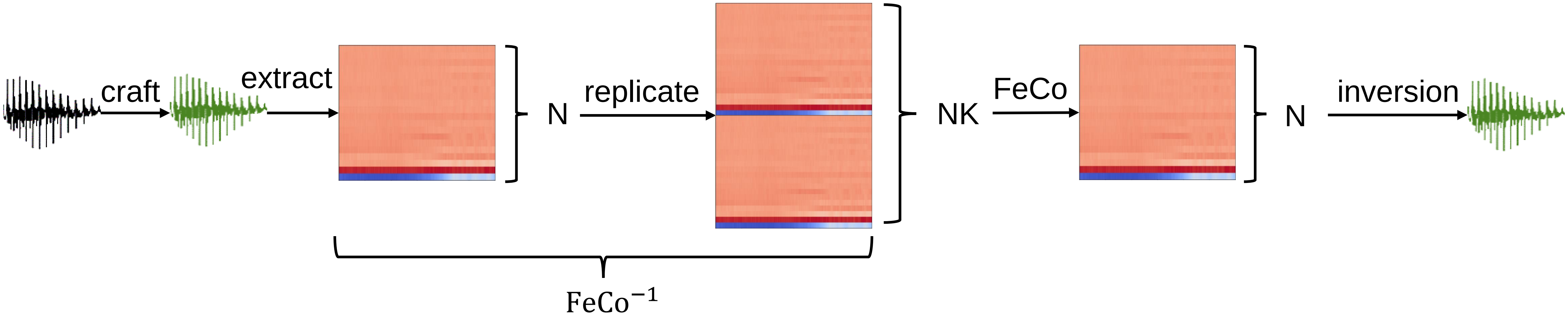
Defense	R_1 Score	A_b	A_a													
			L_∞ white-box attacks							L_2 white-box attacks			black-box attacks			
			FGSM	PGD			CW_∞			CW_2			Score-based (L_∞)		Decision-only
				10	20	100	10	20	100	0	0.2	0.5	FAKEBOB	SirenAttack	Kenansville
Baseline	15.6	99.7	48.4	0.4	0.1	0	0	0	0	3.4	0	0	6.9	28.4	22.2
FeCo-o(wk)-ts	78.8	95.4	72.4	59.1	60.7	65.5	58.8	58.4	63.6	93.7	91.1	81.1	84.6	50.5	33.9
FeCo-o(wk)-rd	70.7	99.1	73.7	32.3	34.7	46.3	21.1	22.4	32	97.2	90.9	66.5	90.1	74.2	32.3

Defense: experiments against adaptive attacks

adaptive attacks:
have complete knowledge of defenses

1st adaptive attacker:
end-to-end differentiable;
overcome randomness by expectation over transformation (EOT):
In each step, independently sample FeCo multiple times and average the losses

2nd adaptive attacker:
Replicate feature attack (Replicate)



Defense: experiments against adaptive attacks

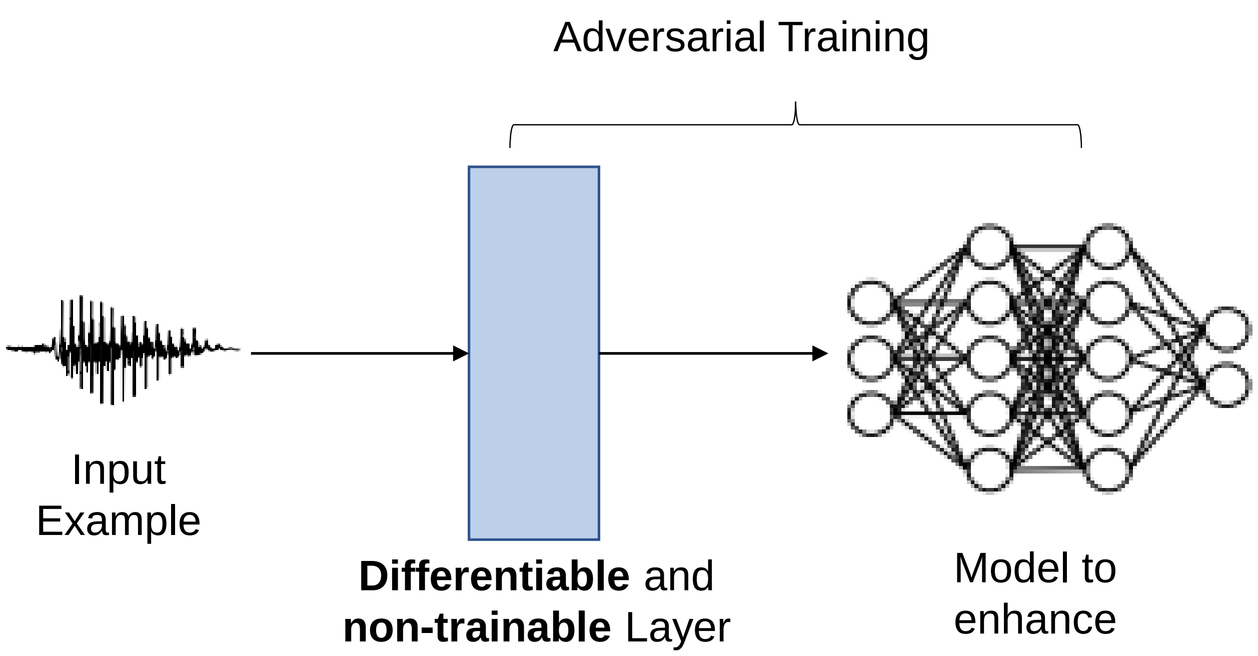
adaptive attacks:
have complete knowledge of defenses

1st adaptive attacker: EOT
2nd adaptive attacker: Replicate

Defense	Adaptive Techniques	L_∞ white-box attacks					L_2 white-box attacks									black-box attacks		
		FGSM	PGD-10	PGD-100	CW_∞-10	CW_∞-100	CW_2-0			CW_2-2			CW_2-50			FAKEBOB	SirenAttack	Kenansville
		A_a	A_a	A_a	A_a	A_a	A_a	SNR	PESQ	A_a	SNR	PESQ	A_a	SNR	PESQ	A_a	A_a	A_a
FeCo-o(k)	EOT	54.1%	0%	0%	0%	0%	90.4%	56.20	4.14	88.0%	53.54	4.05	1.2%	18.38	1.57	92.17%	96.4%	31.0%
	Replicate-W	68.0%	39.4%	49.0%	39.3%	49.9%	82.7%	-	-	78.7%	-	-	58.6%	-	-	87.8%	83.9%	20.0%
	Replicate-F	72.4%	7.9%	15.6%	7.3%	14.5%	92.8%	-	-	88.6%	-	-	36.7%	-	-	98.1%	93.2%	22.6%

Defense: incorporating adversarial training

adversarial training:
augment training data with adversarial examples



Defense: incorporating adversarial training

Vanilla-AdvT: sole adversarial training

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%

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

Reason: the larger randomness of FeCo enables models to encounter more diverse adversarial examples during training

Defense: incorporating adversarial training

Tuning #attack steps N, attack step size, EOT size R

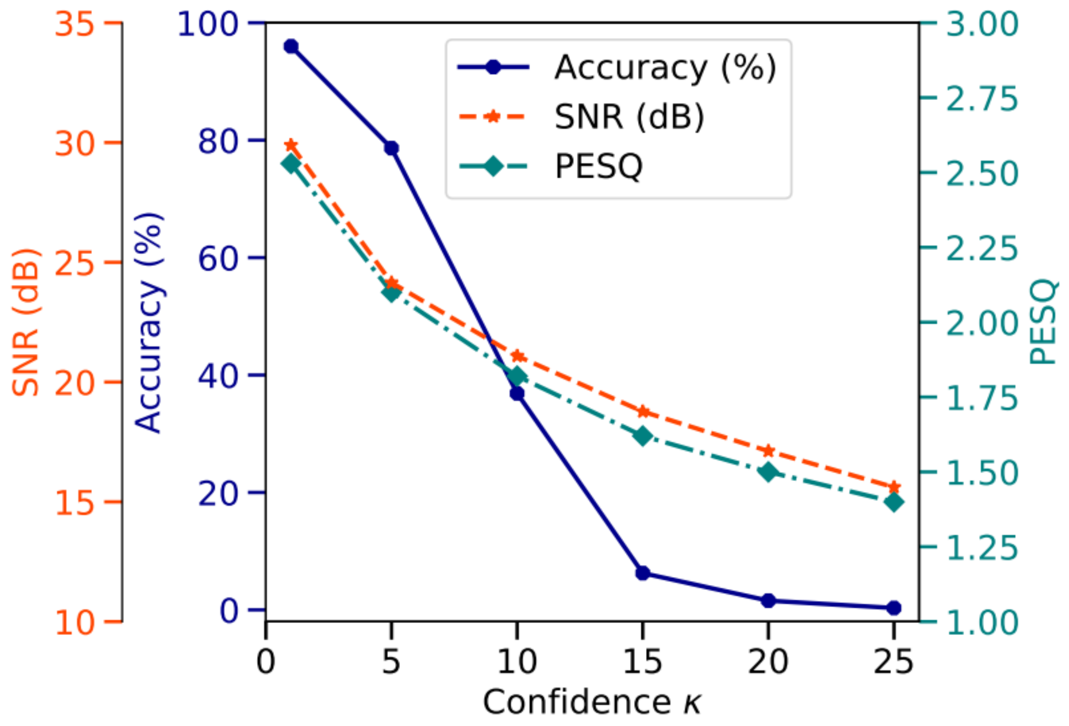
accuracy of AdvT+FeCo-o plateaus at 60.62% with $R = 275, N = 100, \alpha = \frac{\varepsilon}{20}$
accuracy of Vanilla-AdvT plateaus at 47.0% with $R = 1, N = 100, \alpha = \varepsilon/40$.

- improve adversarial accuracy: from 47.0% to 60.62%
- increase attack cost: from 100×1 to 100×275

Defense: incorporating adversarial training

Tuning #attack steps N, attack step size, EOT size R

- worsen imperceptibility:



- no free lunch: degrade the inference efficiency

SpeakerGuard: security analysis platform

SpeakerGuard:

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
- Diverse audio defense solutions
- Evaluation metrics of listening