Learning to Fool the Speaker Recognition

ICASSP 2020

[1,4]Jiguo Li,
jiguo.li@vipl.ict.ac.cn

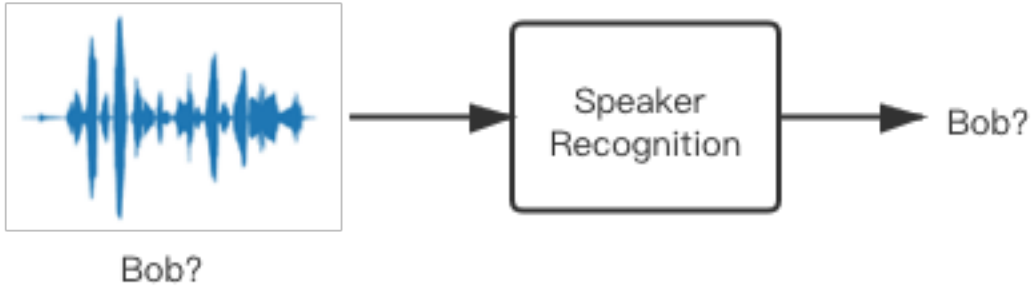
Joint work with [4]Xinfeng Zhang, [3]Jizheng Xu, [3]Li Zhang, [3]Yue Wang, [2]Siwei Ma, [2]Wen Gao

[1]Institute of Computing Technology(ICT), Chinese Academy of Sciences(CAS)
[2]Institute of Digital Media(IDM), Peking University(PKU)
[3]Bytedance.Inc
[4]University of Chinese Academy of Science(UCAS)

Biometric Systems

- Image-based Systems
  - ✓Face, fingerprint

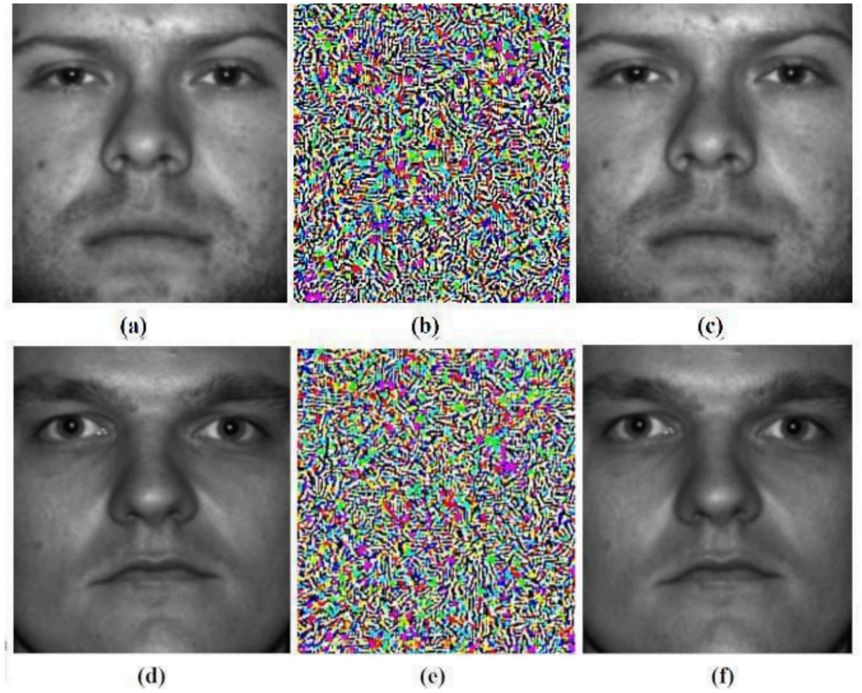
- Speech-based Systems
  - ✓Speaker recognition



Security Risks for Biometric Systems

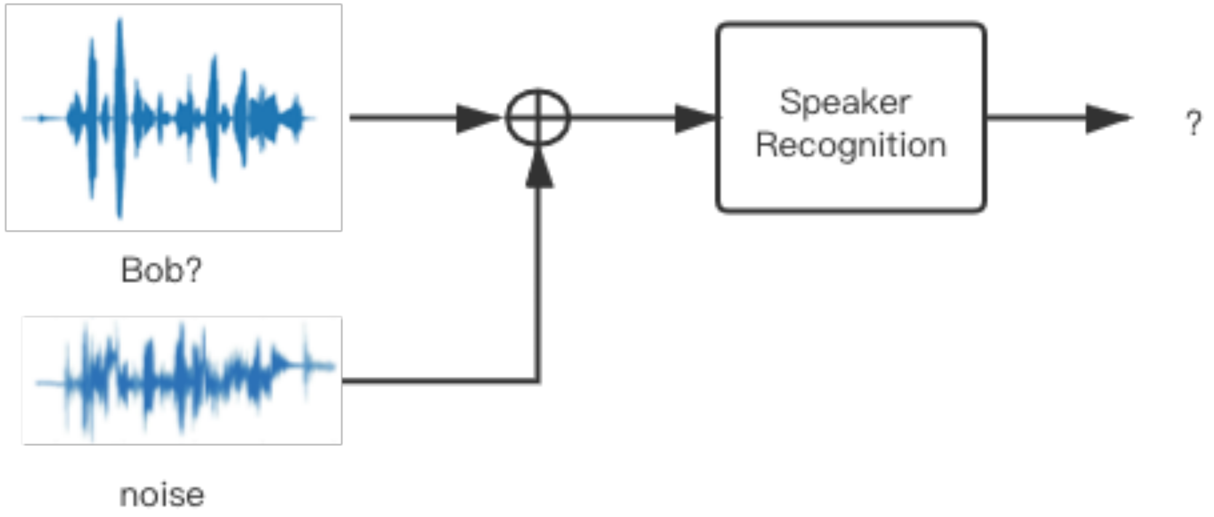
◆ Image-based Systems

✓Face attack



◆ Speech-based Systems

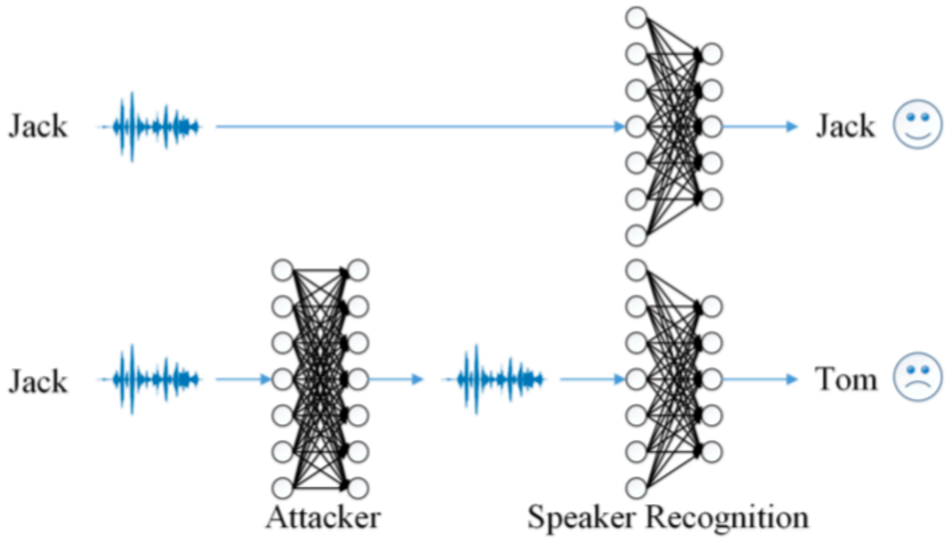
✓Speech attack



Alparslan, Yigit, et al. "Adversarial Attacks on Convolutional Neural Networks in Facial Recognition Domain." *arXiv preprint arXiv:2001.11137* (2020).

Motivation

- How to attack the speech-based biometric systems?
- Is the speech-based biometric systems vulnerable to the adversaria attack?
- Is it possible to design a biometric systems robust to the adversarial attack?



Our attack results

Non-targeted attack

	real	fake
dr1/fcjf0/si1027.ogg		
dr1/fcjf0/sx37.ogg		
dr2/faem0/si762.ogg		
dr8/fbcg1/sx82.ogg		

Targeted attack

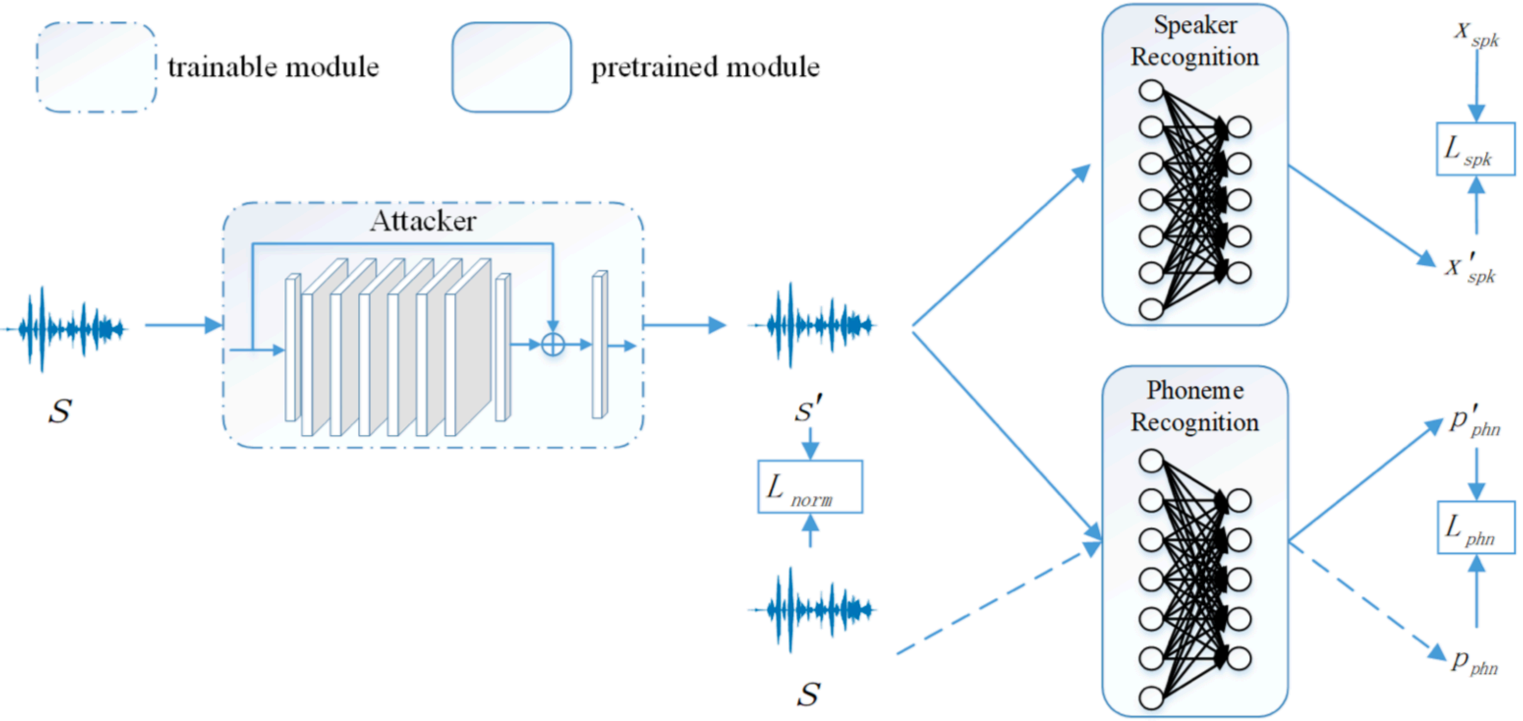
	real	target0	target100	target200
dr1/fcjf0/si1027.ogg				
dr1/fdaw0/si1046.ogg				
dr2/faem0/si762.ogg				
dr8/fbcg1/si982.ogg				

More results: https://smallflyingpig.github.io/speaker-recognition-attacker/main

Proposed Attack Framework

◆Our Framework

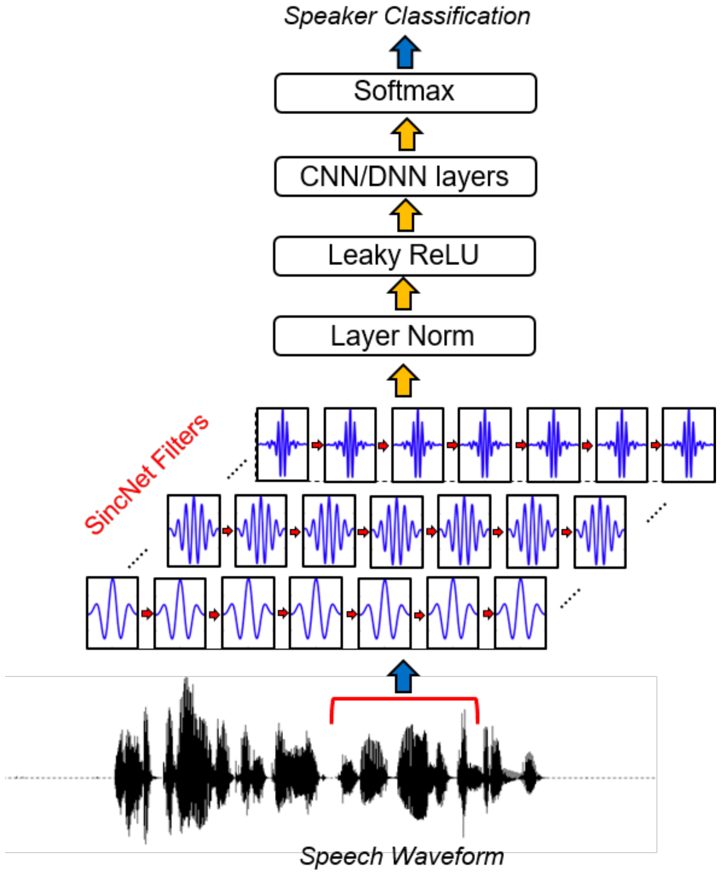
- ✓An attacker for all samples
- ✓Optimize the speech via phoneme recognition module



Proposed Attack Framework

◆Speaker/Phoneme Recognition Model: Sincnet[1]

- ✓Frequency filters in the first layer
- ✓Process on the raw waveform
- ✓More interpretable



Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

Proposed Attack Framework

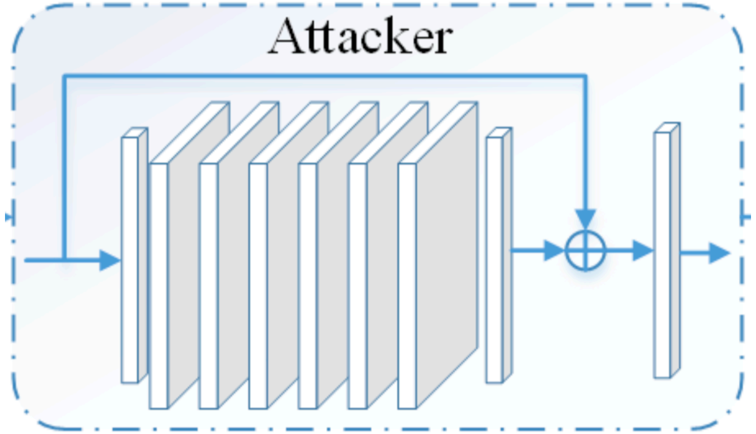
◆Attacker: a Residual Block

✓Referring to Adversarial Transformer Networks (ATNs)[1]

✓Additive perturbations

✓The scale of the perturbation is controllable

✓Training once for all testing samples



[1] Baluja, Shumeet, and Ian Fischer. "Learning to Attack: Adversarial Transformation Networks." *AAAI*. Vol. 1. 2018.

Proposed Attack Framework

◆Adversarial training/Optimization

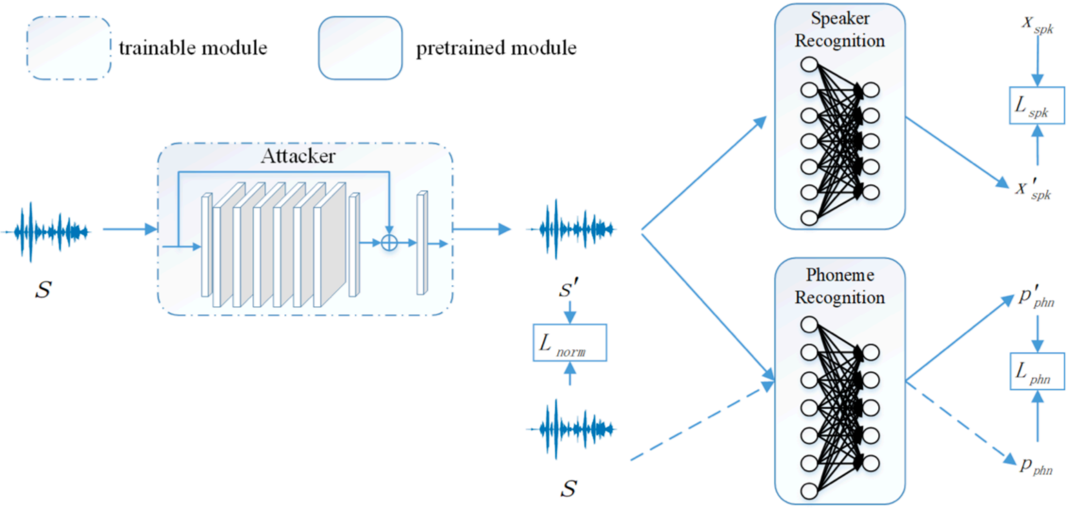
✓Non-targeted attack

$$L_{total} = L_{spk} + \lambda_{phn} L_{phn} + \lambda_{norm} L_{norm}$$

$$L_{spk} = \begin{cases} x'_{spk}[I_{1st}] - x'_{spk}[I_{2nd}], I_{1st} = y_{spk} \\ 0, \quad else \end{cases}$$

$$L_{phn} = KL(p_{phn} || p'_{phn})$$

$$L_{norm} = [\max(s - s' - m, 0)]^2$$



✓Targeted attack

$$L_{spk} = \begin{cases} x'_{spk}[I_{1st}] - x'_{spk}[y_{target}], I_{1st} \neq y_{target} \\ 0, \quad else \end{cases}$$

Experimental Results

◆Datasets and Metrics

✓Dataset

Dataset	Label	Speaker number	Samples (train+test)
TIMIT	Speaker+phoneme	462	3694(2309+1385)

✓Metric

- ➢Sentence Error Rate(SER): used for non-targeted attack
- ➢Prediction Target Rate(PTR): used for targeted attack
- ➢Signal-noise Ratio(SNR)
- ➢Perceptual Evaluation of Speech Quality(PESQ): 0.5~4.5

Experimental Results

◆Can our proposed model attack the pretrained speaker recognition model?

- ✓Non-targeted attack
- ✓SER 90.5% with SNR 59.01 dB
- ✓SER 90.5% with PESQ 4.28

λ_{phn}	λ_{norm}	SER(%)↑	SNR(dB)↑	PESQ↑
-	-	1.52*	-	-
0	0	99.7	18.56	1.09
0	1000	96.5	56.39	3.72
0	2000	86.7	57.79	3.61
1	1000	**99.2**	57.20	4.20
5	1000	93.9	58.00	4.25
10	1000	90.5	**59.01**	**4.28**

Experimental Results

◆Can our proposed model attack the pretrained speaker recognition model?

- ✓Targeted attack
- ✓Average success rate 72.1%
- ✓Average SNR 57.64dB
- ✓Average PESQ 3.48

Target ID	PTR(%)↑	SNR(dB)↑	PESQ↑
0	91.4	57.55	3.36
100	89.3	56.83	3.16
200	63.3	58.42	3.69
300	58.7	56.92	3.52
400	57.6	58.36	3.68
avg	72.1	57.64	3.48

Experimental Results

◆Does our design work? (the phoneme recognition model)

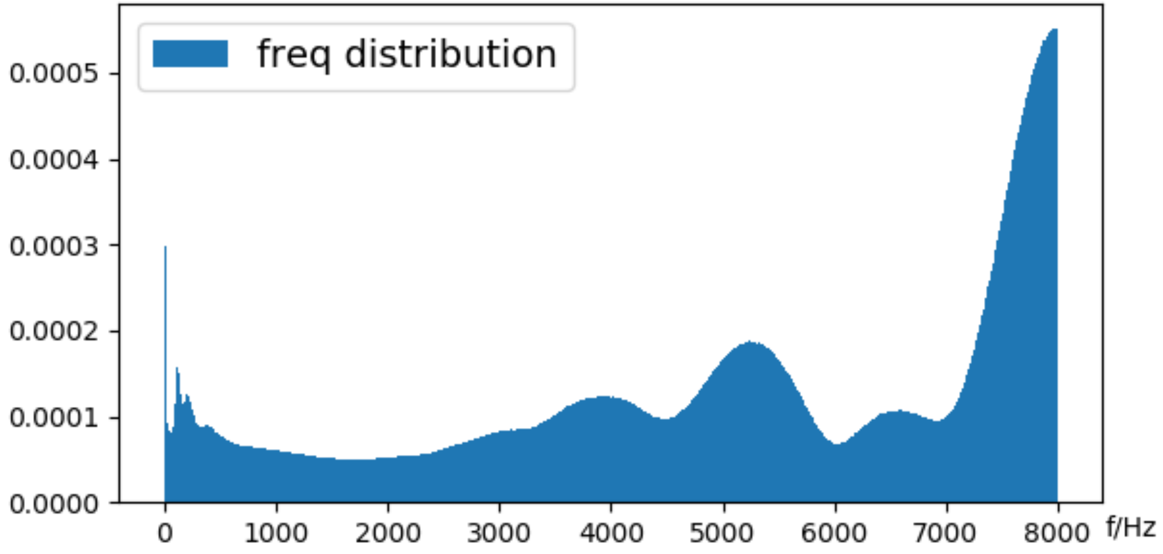
✓With fixed λ_{norm}, larger λ_{phn}results a higher SNR and PESQ

✓The phoneme brunch works for obtaining a trade-off between SER and SNR/PES

λ_{phn}	λ_{norm}	SER(%)↑	SNR(dB)↑	PESQ↑
-	-	1.52*	-	-
0	0	99.7	18.56	1.09
0	1000	96.5	56.39	3.72
0	2000	86.7	57.79	3.61
1	1000	**99.2**	57.20	4.20
5	1000	93.9	58.00	4.25
10	1000	90.5	**59.01**	**4.28**

Experimental Results

- Other findings
  - ✓The perturbations concentrate on high frequency
  - ✓Can we design robust speaker recognition models focusing on the low frequency? (future works)



Perturbations distribution

The questions

- How to attack the speech-based biometric systems?
  - ✓Our proposed framework successfully attacked the SOTA speaker recognition model
- Is the speech-based biometric systems vulnerable to the attacker?
  - ✓Yes
- Is it possible to design a biometric systems robust to the adversarial attack?
  - ✓The future works

Thanks

Q & A

Codes, data and more results: https://smallflyingpig.github.io/speaker-recognition-attacker/main
Paper early access: https://ieeexplore.ieee.org/document/9053058