Carnegie Mellon University

INTERSPEECH 2024

To what extent can ASV systems naturally defend against spoofing attacks?

Jee-weon Jung*, Xin Wang*, Nicholas Evans, Shinji Watanabe, Hye-jin Shim, Hemlata Tak, Siddhant Arora, Junichi Yamagishi, Joon Son Chung

Motivation

- Current speaker verificaiton systems are vulnerable towards spoofing attacks
- Speech deepfake and spoofing detection field is growing

- What if the advancements in speaker verification systems naturally lead to spoofing-robust verification systems?
  - If yes, less need for speech anti-spoofing research?

Goal

- Investigate the trajectory of spoofing-robustness across speaker verification systems through time
  - If speaker verification systems are gaining spoof-robustness, estimate the speed of development
  - Confirm if different spoofing attacks pose different amount of threats

Metric

- SPF-EER
  - An estimation on how good a speaker verification system is at rejecting spoofed inputs
  - An evaluation protocol comprising ***target*** and ***spoof*** trials is used

	SV-EER	SPF-EER
Target	+	+
Non-target	-	
Spoof		-

Speaker verification systems

1. GMM-UBM
2. i-vector
3. x-vector
4. ECAPA-TDNN
5. MFA-Conformer
6. SKA-TDNN
7. RawNet3
8. WavLM-Large+ECAPA

Spoofing attacks

- 29 attacks from ASVspoof 2015 and ASVspoof 2019 logical access
  - Covers TTS and VC systems (not replay)

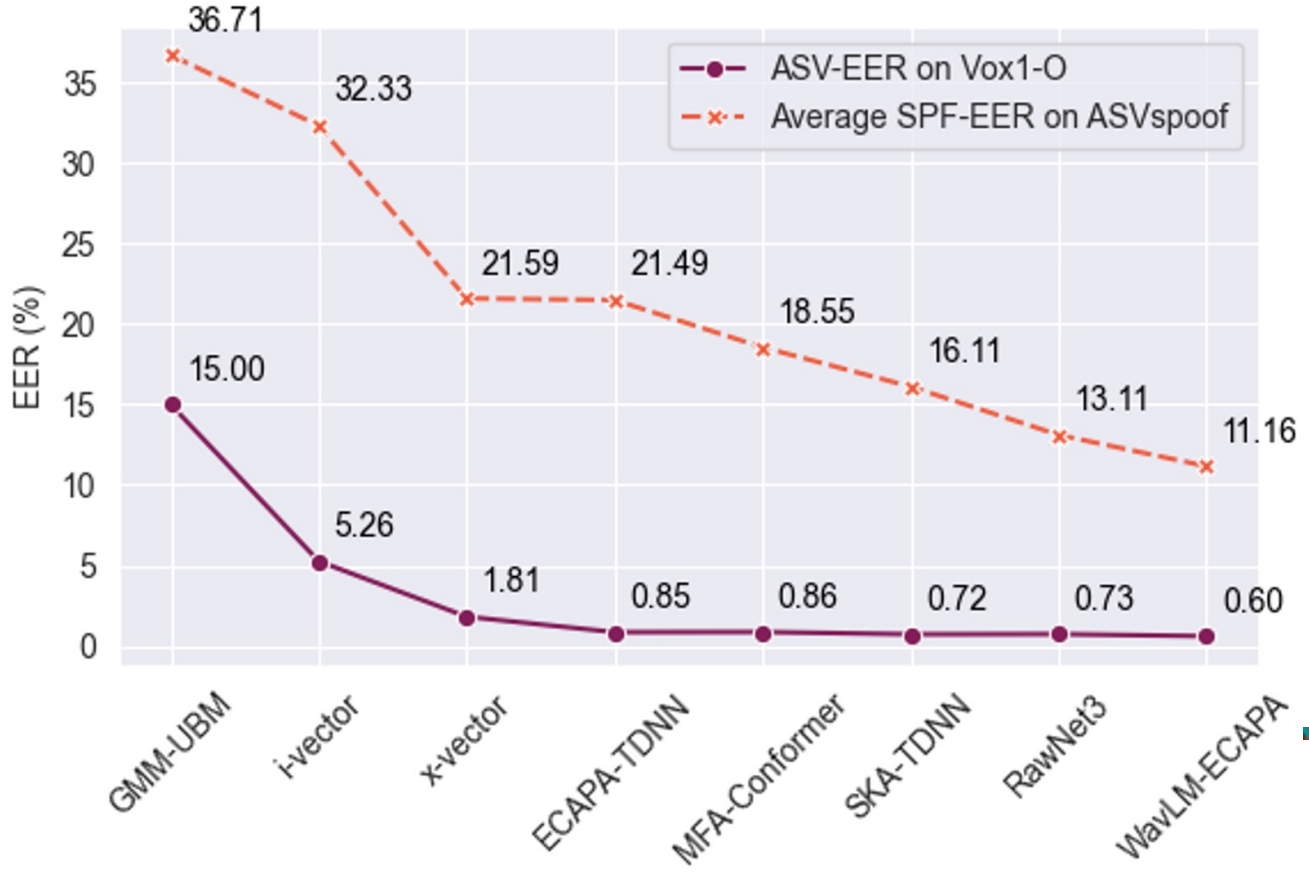
Group	ID	Type	Acoustic model	Waveform model
1	A18	VC	i-vector + PLDA	LPC
	S5	VC	GMM	MLSA
	A06	VC	GMM	spectral filtering
	A19	VC	GMM	spectral filtering
	S2	VC	Linear reg.	STRAIGHT
	S1	VC	DTW	STRAIGHT
	S6	VC	GMM + GV	STRAIGHT
	S7	VC	GMM + GV	STRAIGHT
	S3	TTS	NLP + HMM	STRAIGHT
	S4	TTS	NLP + HMM	STRAIGHT
	S8	VC	GMM-tensor	STRAIGHT
	S9	VC	DTW + Kernel reg.	STRAIGHT
2	A05	VC	VAE	WORLD
	A17	VC	VAE	waveform filtering
	A13	TTS	TTS + VC(DNN)	waveform filtering
	A09	TTS	NLP + RNN	Vocaine
	A14	TTS	TTS + VC(DNN)	STRAIGHT
	A03	TTS	NLP + DNN	WORLD
	A02	TTS	NLP + HMM-DNN	WORLD
	A07	TTS	NLP + RNN-GAN	WORLD
	A11	TTS	DNN(end2end)	Griffin-Lim
3	A08	TTS	NLP + HMM-DNN	Dilated CNN
	A01	TTS	NLP + HMM-DNN	WaveNet
	A12	TTS	NLP + RNN	WaveNet
	A15	TTS	TTS + VC(DNN)	WaveNet
	A10	TTS	DNN(end2end)	WaveRNN
4	S10			
	A04	TTS	NLP + Unit-selection	Waveform concat.
	A16			

Corpora

- VoxCelebs 1&2 development sets
  - 7,205 speakers / 2.5k+ hours of speech
  - Used for training speaker verification models
- Vox1-O protocol
  - 40 speakers / 37k+ trials
  - Used for assessing speaker verification performance (SV-EER)
- ASVspoof 2019 logical access evaluation set
  - 48 speakers / 68k+ utterances

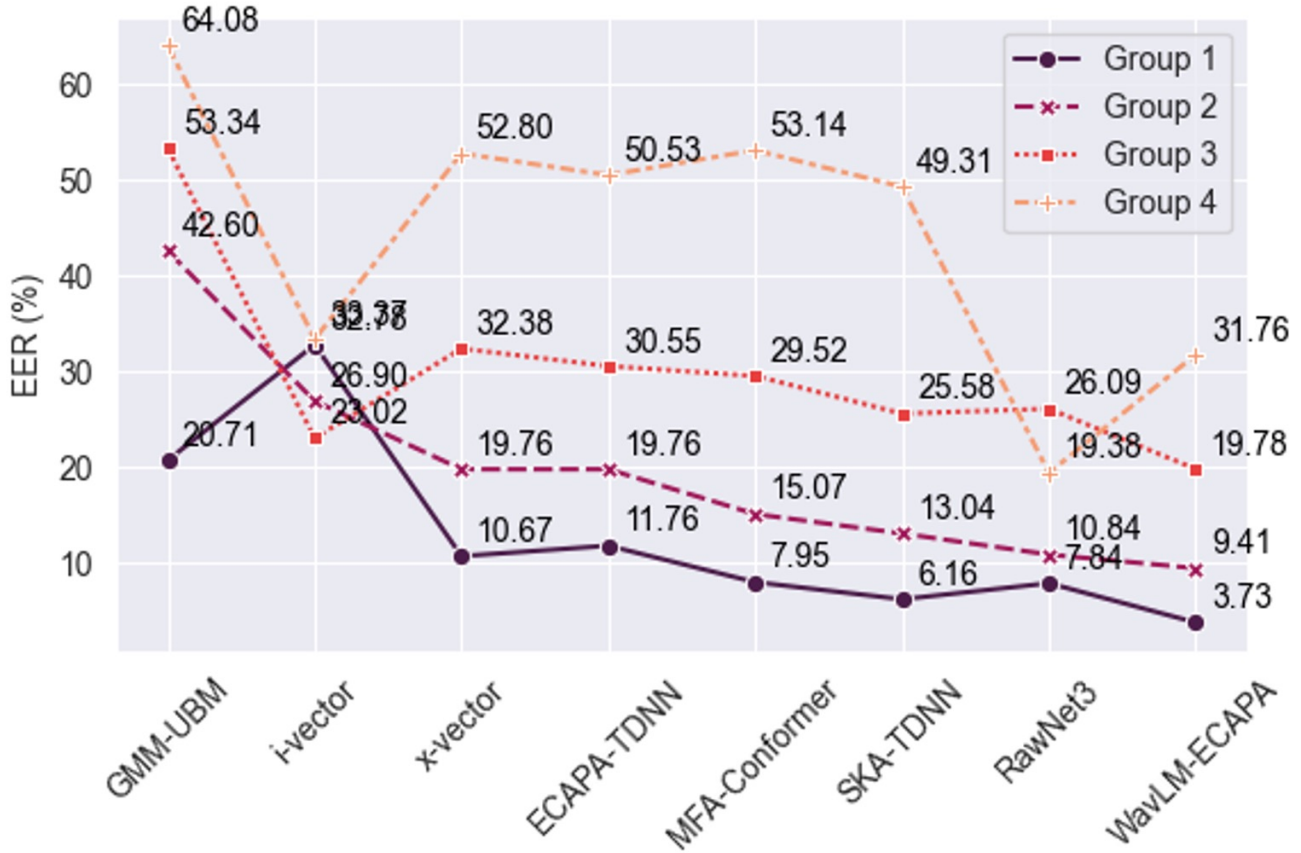
General result – speaker verification

- Speaker verification systems are achieving zero-shot spoofing-robustness
  - Yet, the development of speech generation technologies outpaces



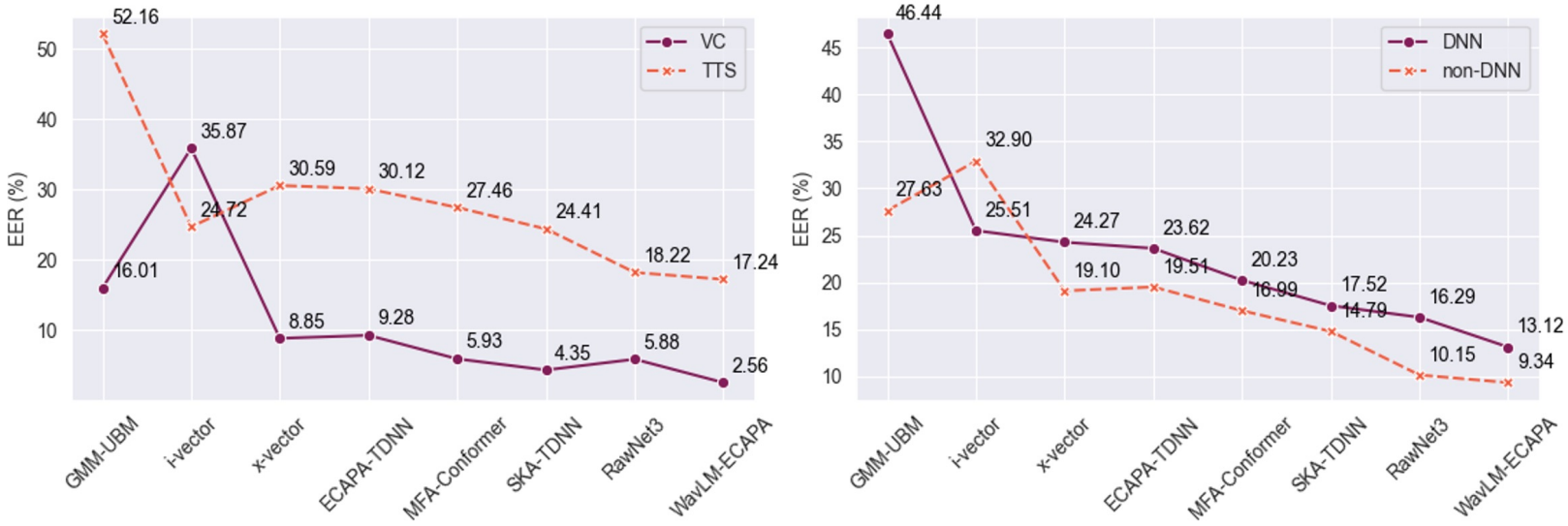
Results across different groups – speaker verification

- SSL-based model achieves the be performance on average, but does not guarantee better spoofing-robustness across all groups
- RawNet3 was most effective against Group 4 attacks
- i-vector has mixed tendency on different groups



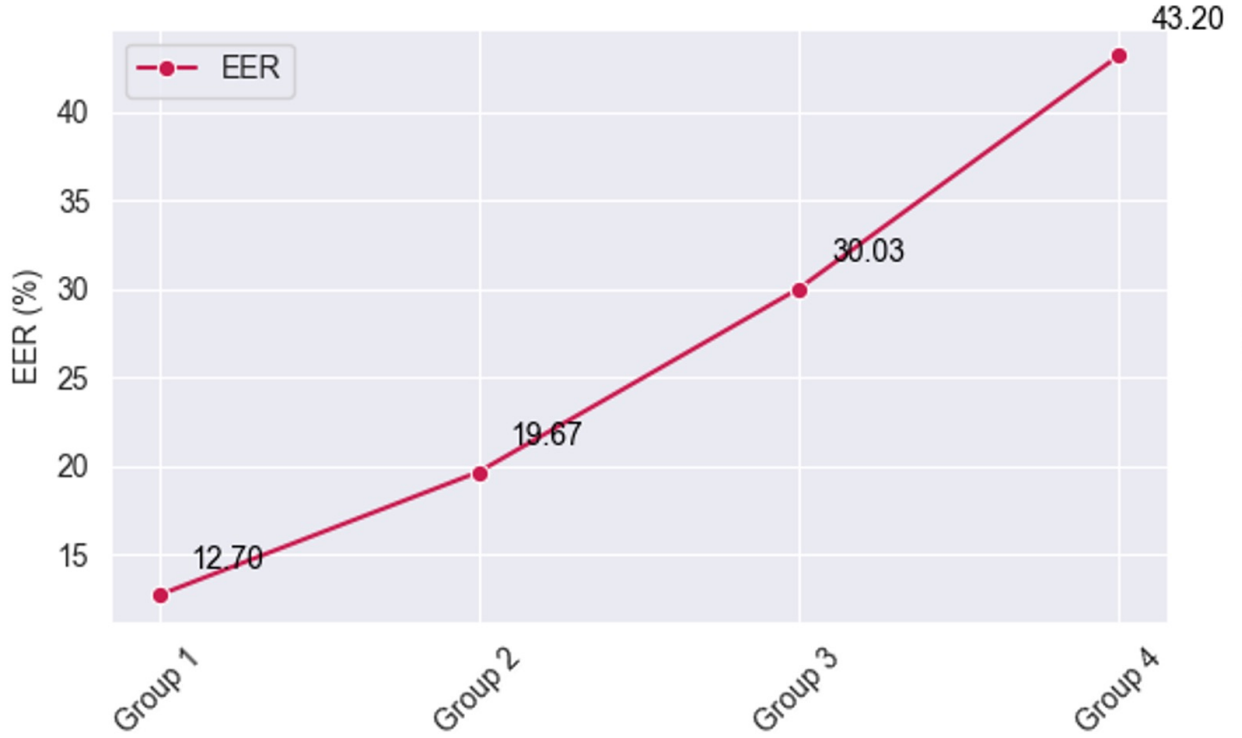
Results on TTS/VC & DNN/non-DNN

- VC attacks are easier to detect for speaker verification systems
- DNN-based attacks are more harder to detect



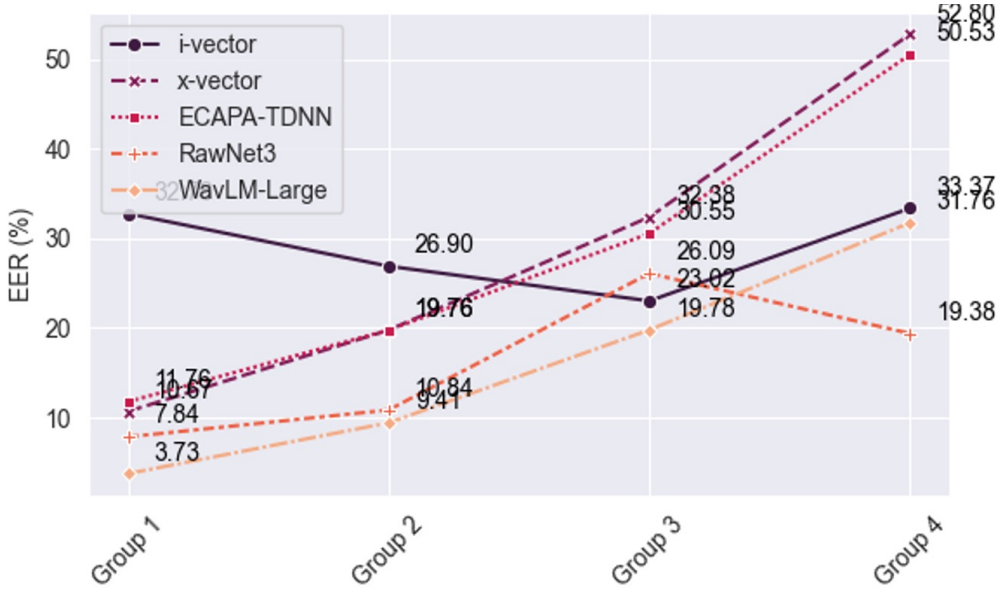
General result – viewpoint of attacks

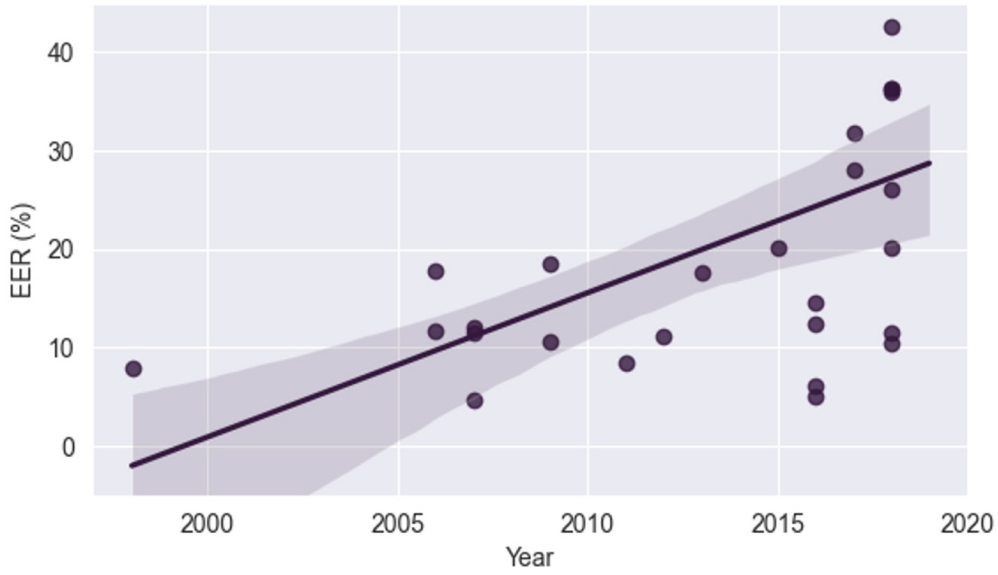
- Group 1 is the easiest and Group 4 is the hardest to detect



Chronological results on attacks

- More recent attacks are harder to detect





Takeaways

- Speaker verification systems are gaining zero-shot robustness against spoofing attacks
- The pace of advancement is slower than that of speech generation technology
- More recent attacks are harder to detect

- We need more effort on speech deepfake detection/anti-spoofing and spoofing-robust automatic speaker verification (SASV)!!