Parrot-Trained Adversarial Examples: Pushing the Practicality of Black-Box Audio Attacks against Speaker Recognition Models

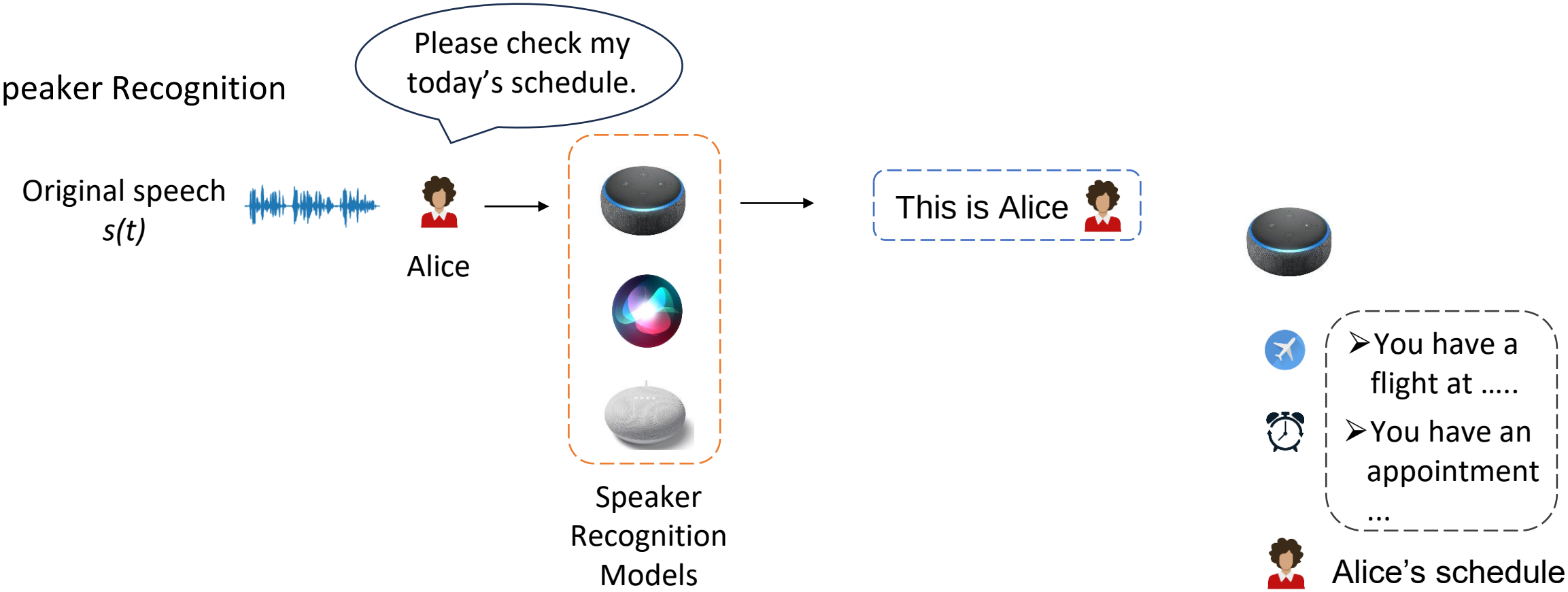
Rui Duan[1], Zhe Qu[2], Leah Ding[3], Yao Liu[1], Zhuo Lu[1]

[1]University of South Florida, USA
[2]Central South University, China
[3]American University, USA

NDSS
SYMPOSIUM/2024

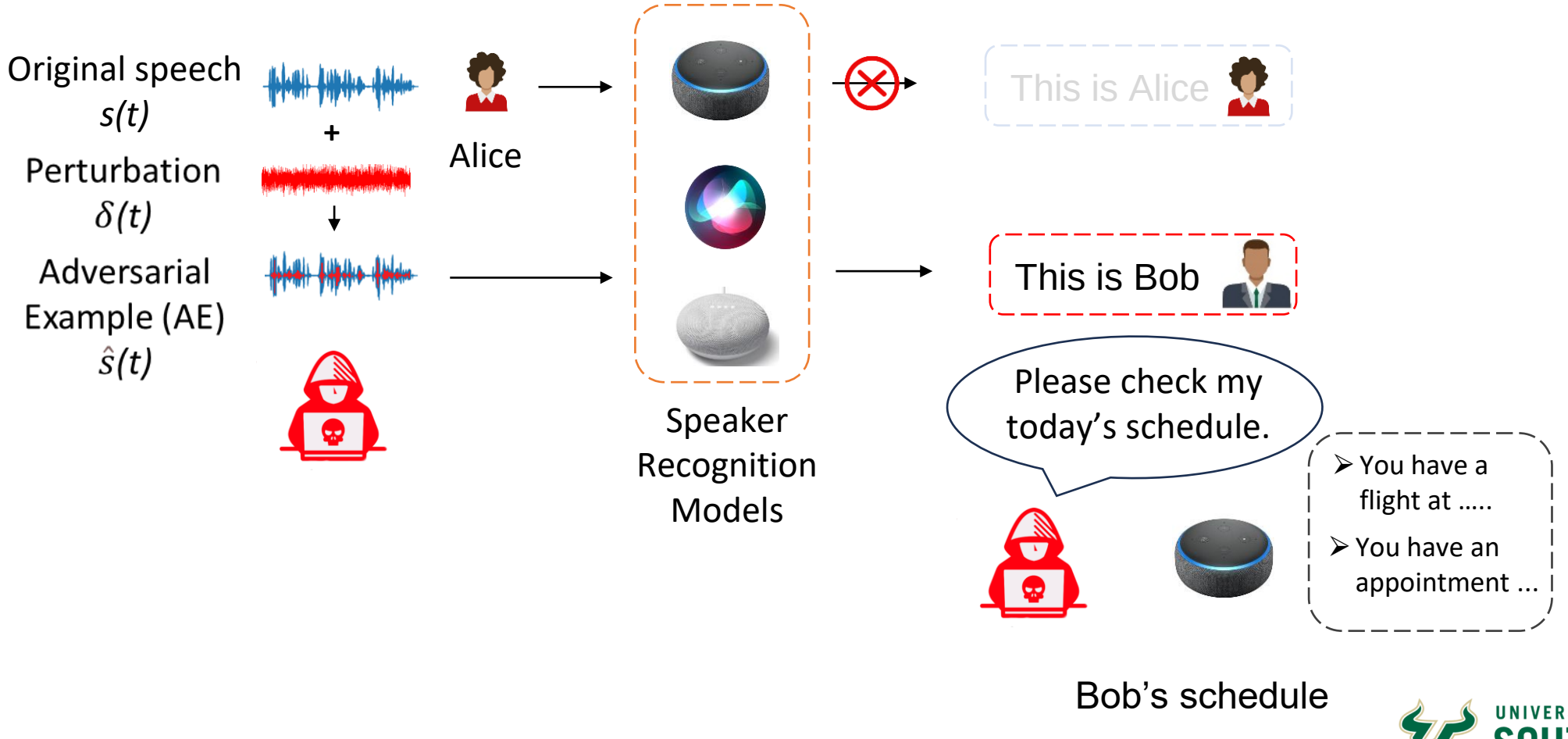
Speaker Recognition Models

❑ Speaker Recognition



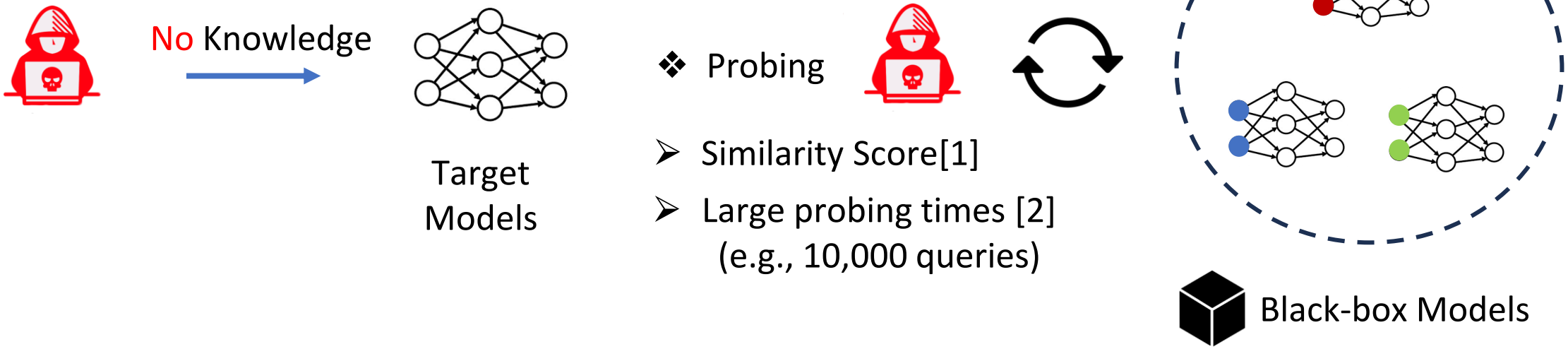
Adversarial Attack on Speaker Recognition Models

❑ Speaker Recognition



Existing Black-box attacks

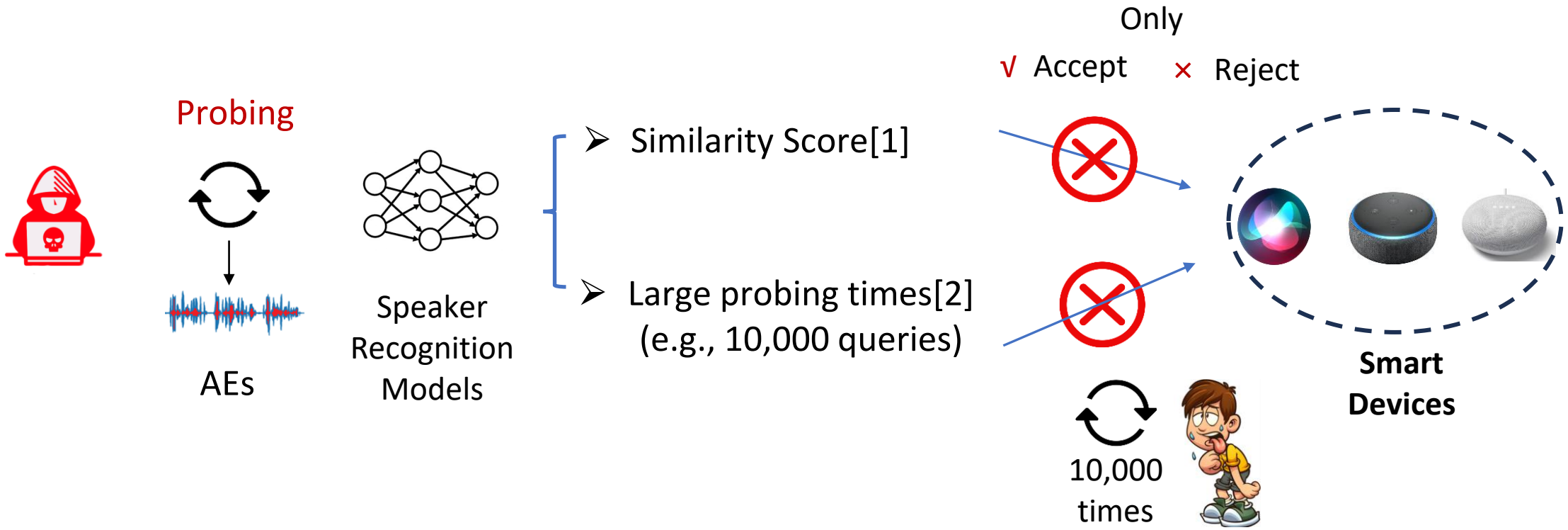
❑ Black-box attacks in digital line



Reference:
[1] Chen et al. "Who is real bob? adversarial attacks on speaker recognition systems." 2021 IEEE Symposium on Security and Privacy (SP).
[2] Zheng et al. "Black-box adversarial attacks on commercial speech platforms with minimal information." 2021 ACM CCS.

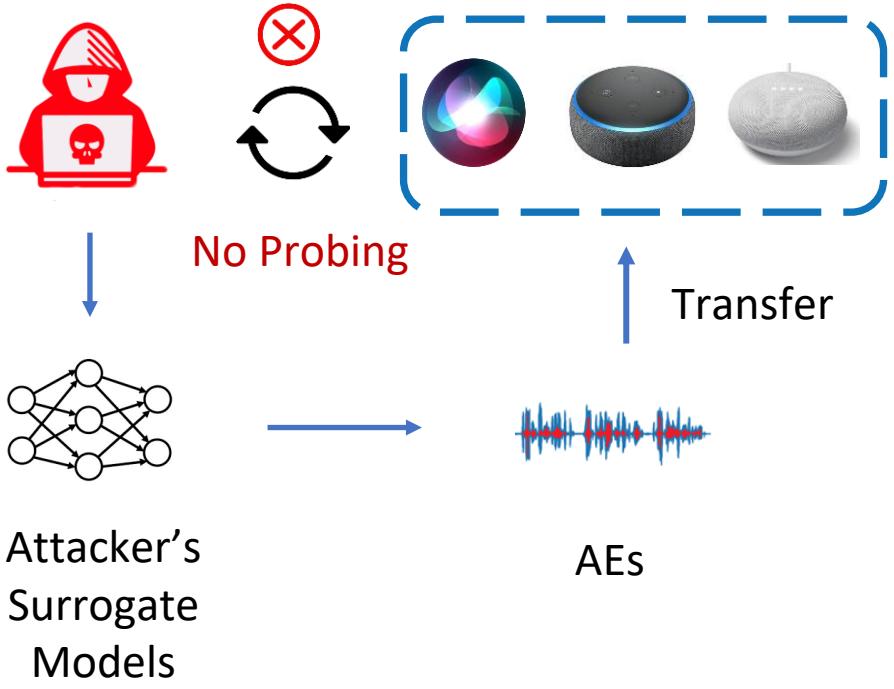
Motivation

❑ Existing Black-box attack is limited to Over-the-air scenario



Potential Solution

❑ Transfer attack

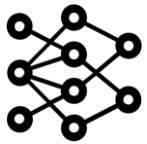


Challenges

- Minimal attack knowledge
  - Only know one short speech of target speaker

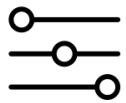
8 seconds

Training datasets



Model Architecture

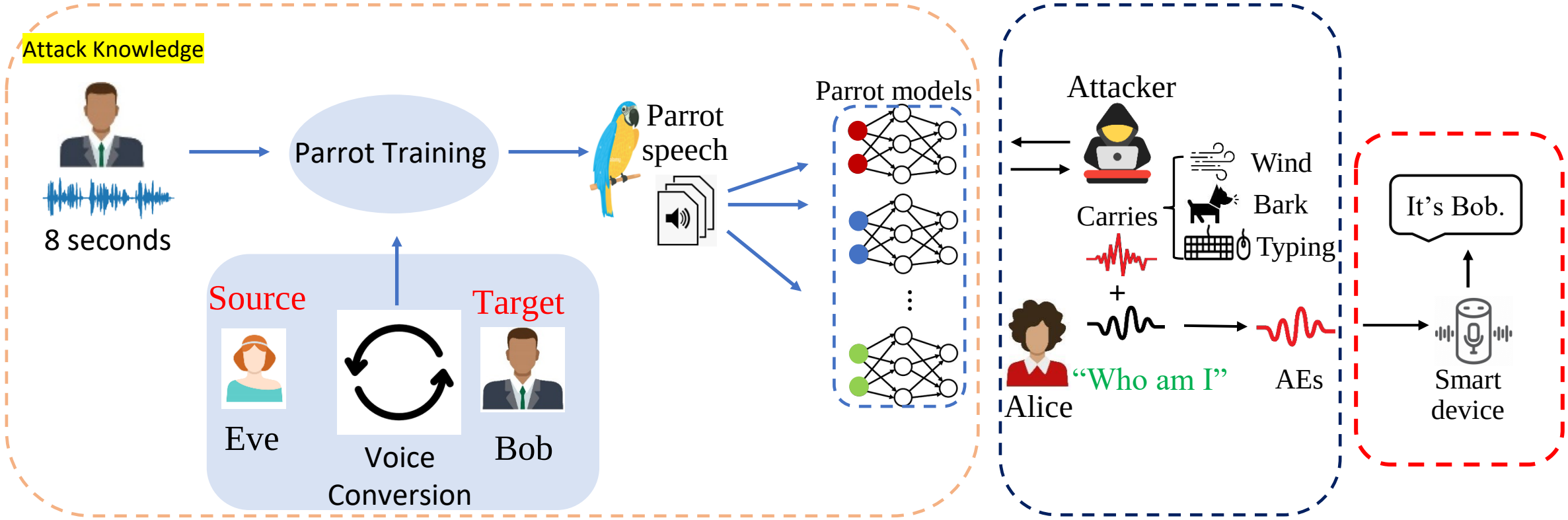
...



Parameters

Overview of Parrot Training Attack

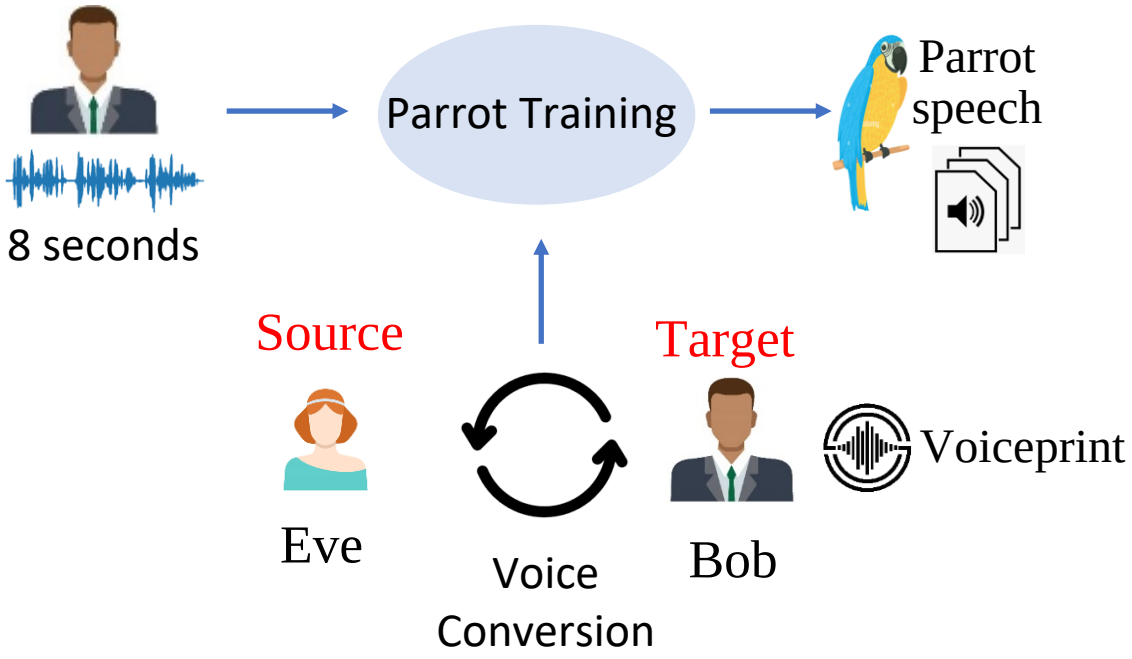
❑ Workflow of Parrot Training attack



Parrot Training Attack

❑ Build surrogate model to approximate black-box model

➢ **Training datasets:**

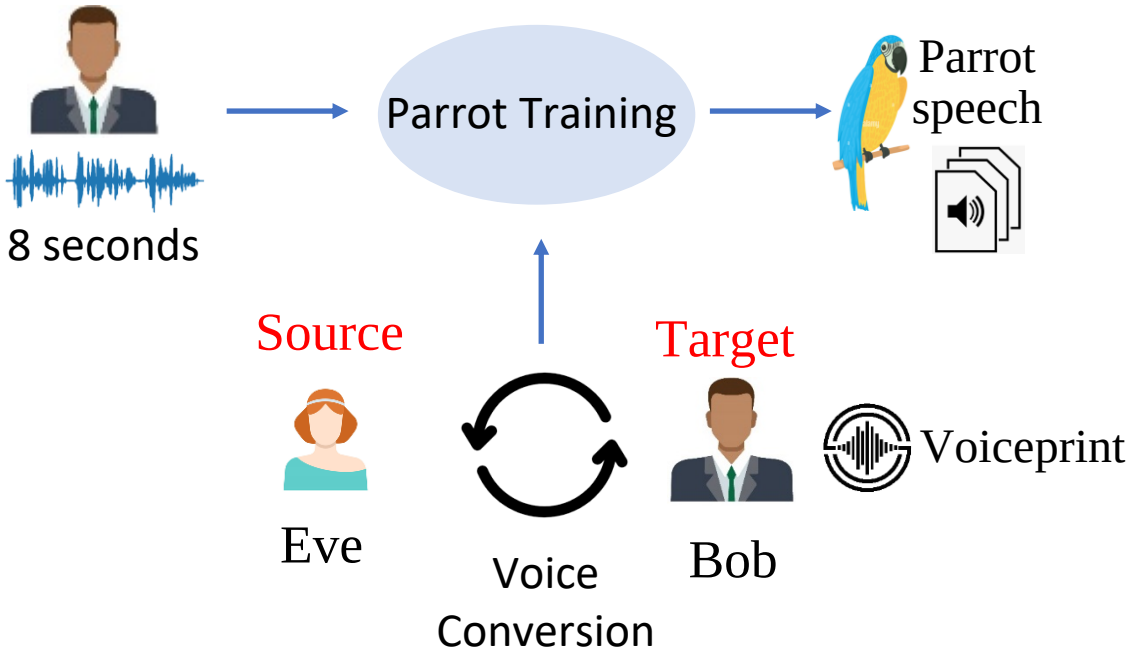


One-shot Voice Conversion

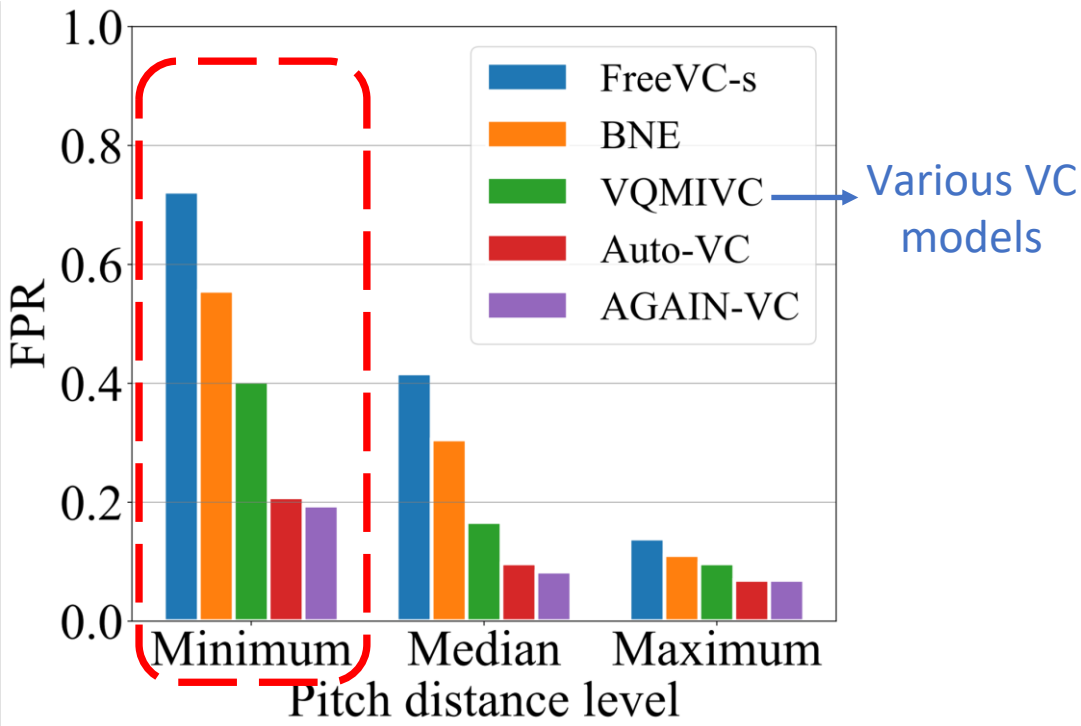
Parrot Training Attack

❑ Build surrogate model to approximate black-box model

➢ **Training datasets:**



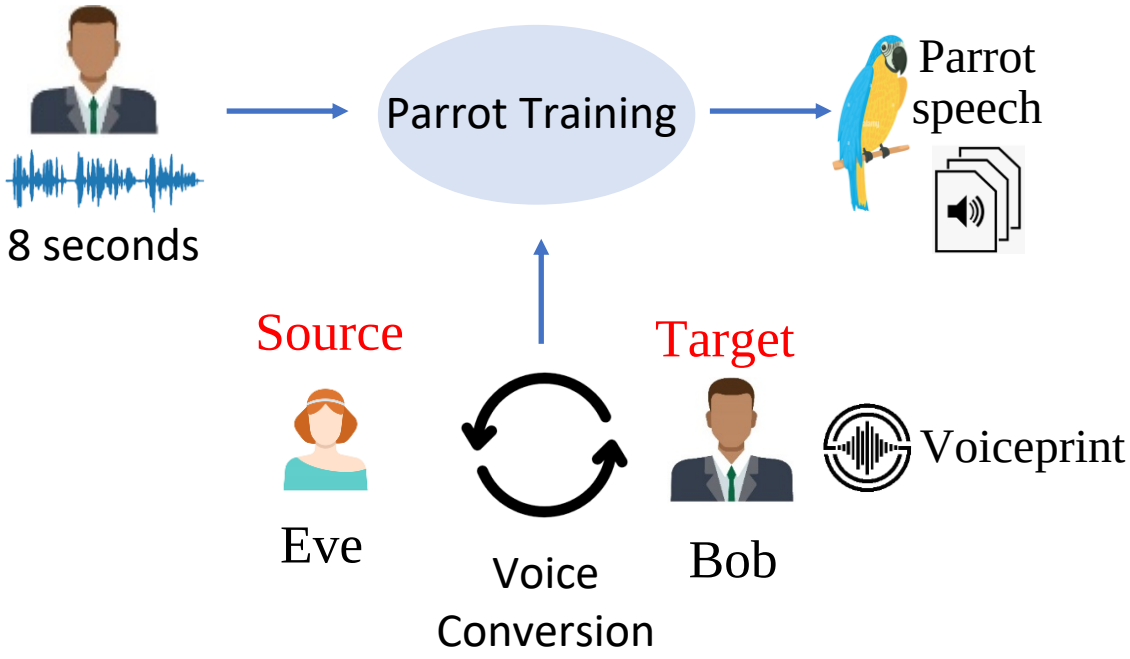
1. Selection of Source Speaker:
- Pitch feature



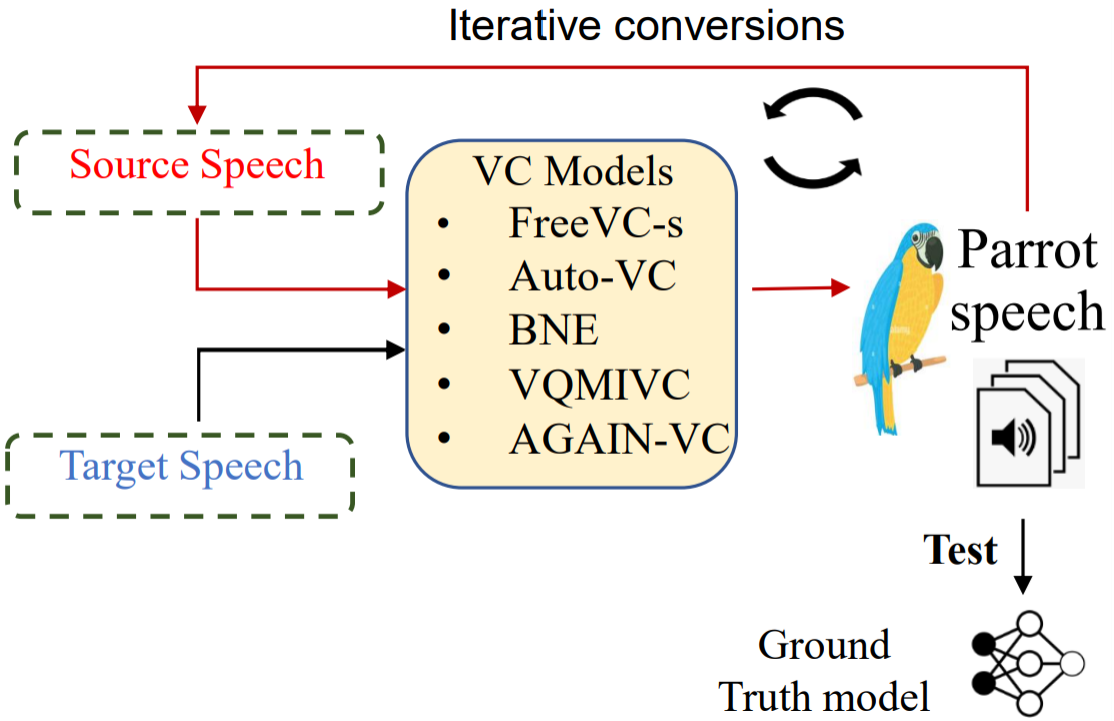
Parrot Training Attack

❑ Build surrogate model to approximate black-box model

➢ **Training datasets:**



2. Iterative conversions:
- 5 times conversion



Parrot Training AEs

- How to generate a good AE based on the PT-models?
  - **Attack algorithm-AE generation :** 
    - Perception
    - Transferability

Parrot Training AEs

❑ How to generate a good AE?

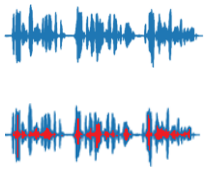
➢ **Attack algorithm-AE generation :** 🤔

- ➢ Perception
- ➢ Transferability

❖ Human perception model

- We recruited 30 volunteers

Original



Perturbed

- 1 : least similarity
- ⋮
- 7 : most similarity

UNIVERSITY of SOUTH FLORIDA

Parrot Training AEs

❑ How to generate a good AE?

➢ **Attack algorithm-AE generation :**

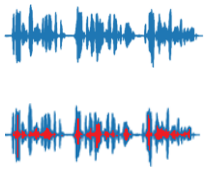
- ➢ Perception
- ➢ Transferability

→ Feature-twisted environmental sound ✓

❖ Human perception model

- We recruited 30 volunteers

Original



Perturbed

1 : least similarity

⋮

7 : most similarity

❖ Transferability

- Noise carrier (PGD, FGSM) Carlini et al., 2018
- Feature-twisted carrier (change the pitch or rhythm) Yu et al., 2023

- Environmental sound carrier Deng et al., 2022

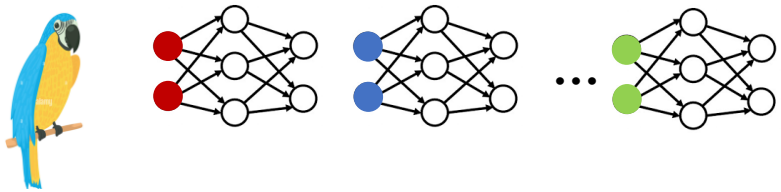
Ensemble learning- Parrot Training Attack

❑ Further enhance transferability

➢ **Model architecture: Ensemble learning**

❖ Different model architecture
-TDNN and CNN

❖ Various parameters
-Different speakers



Experimental Results

❑ Over-the-air:

Amazon Echo

Siri

Google Home

Smack: Improved 264% (attack success) and 11% (human perception score).

QFA2SR: Improved nearly 96% (attack success) and 45% (human perception score).

Smart Devices	Methods	FakeBob		Occam		Smack		QFA2SR		PT-AEs	
	Tasks	ASR	SRS	ASR	SRS	ASR	SRS	ASR	SRS	ASR	SRS
Intra-gender											
Average	-	**4.2%**	**2.15**	**12.5%**	**2.35**	**16.7%**	**4.51**	**31.3%**	**2.75**	**58.3%**	**4.77**
Inter-gender											
	Tasks	ASR	SRS	ASR	SRS	ASR	SRS	ASR	SRS	ASR	SRS
Average	-	**2.1%**	**1.59**	**8.3%**	**1.73**	**12.5%**	**3.82**	**22.9%**	**2.33**	**47.9%**	**4.45**

Experimental Results

❑ With different distances:

Distance	0.25 m	0.5 m	1.0 m	2.0 m	4.0 m
Intra-gender	60.4%	58.3%	52.1%	35.4%	20.8%
Inter-gender	47.9%	47.9%	37.5%	27.1%	14.5%

Demo link: *https://www.youtube.com/watch?v=6Pcca7uQQ4M*

UNIVERSITY of SOUTH FLORIDA

Conclusion

- We use the minimal attack knowledge (e.g., 8 seconds speech) to build our surrogate models via parrot training.

8 seconds

- We systematically evaluate the existing methods from both transferability and human perception.

- We evaluate PT-AEs in the over-the-air scenario with smart devices and compared with recent works.

UNIVERSITY of SOUTH FLORIDA

Thank You!