Vulnerabilities of Voice Assistants at the Edge: From Defeating Hidden Voice Attacks to Audio-based Adversarial Attacks

Yingying (Jennifer) Chen

Professor, Electrical and Computer Engineering Department
Associate Director, WINLAB
Director, Data Analysis and Information Security (DAISY) Lab
Rutgers University, New Brunswick, NJ, USA
yingche@scarletmail.rutgers.edu
http://www.winlab.rutgers.edu/~yychen/

IEEE ICNP Workshop AIMCOM2
October 13, 2020

RUTGERS

WINLAB

Wireless Information Network Laboratory (WINLAB)

- ❑ Industry-university research center founded in1989
  - ❖ Focus on wireless technology
- ❑ Hosting world-class researchers
  - ❖ 20 faculties from different departments
  - ❖ 45 PhD students
- ❑ Active research directions:
  - ❖ Mobile ad hoc networks (MANET) for tactical applications
  - ❖ Mesh network protocols
  - ❖ Delay tolerant networks (DTN)
  - ❖ Software defined networks
  - ❖ Mobile content delivery
  - ❖ Wireless network security

Open-Access Research Testbed for Next-Generation Wireless Networks (ORBIT)

ORBIT nodes

USRP radio board

Control room

- ❑ 400 - USRP open access research testbed
- ❑ Funded by NSF since 2003 with $12M
- ❑ Research Applications:
  - ❖ 5G mm wave
  - ❖ Mobile edge cloud and future mobile Internet
  - ❖ Healthcare IT and Internet of Things (IoT)
  - ❖ Mobile sensing and user behavior recognition
  - ❖ Network coding and spectrum management
  - ❖ Vehicular networking

Cloud Enhanced Open Software Defined Mobile Wireless Testbed for City-Scale Deployment (COSMOS)

- Funded by NSF PAWR for $22M in 2018 for deploying 5G network testbed
- Led by Rutgers and collaborating with Columbia University, New York University and University of Arizona
- Focus on 5G technologies
  - Ultra-high bandwidth and low latency wireless communication
- Tightly coupled with edge cloud computing

- Deployment in New York City
- 9 Large sites and 40 Medium sites
- 200 small nodes to support edge computing

- Fiber connection to Rutgers, GENI/I2, NYU
- Interaction with smart community

- **Research Applications:**
  - Ultra-high bandwidth, low latency, and powerful edge computing
  - Future mobile Internet and mobile edge cloud
  - Healthcare IT and Internet of Things (IoT)
  - AR and VR
  - Vehicular networking

RUTGERS

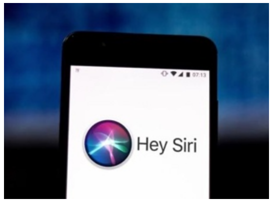
WINLAB

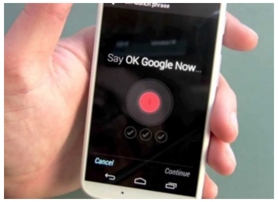
Defeating Hidden Audio Channel Attacks on Edge Voice Assistants - via Audio-Induced Surface Vibrations

Motivation

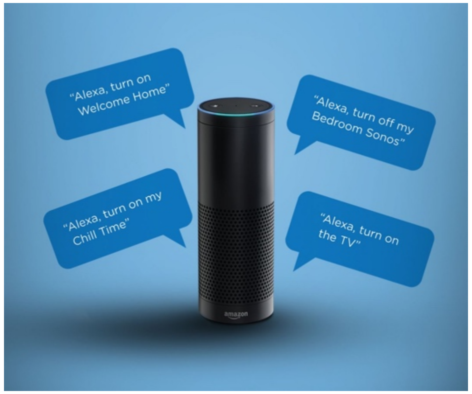
- ❑ Widely deployed voice controllable systems (VCS) at the edge
  - ❖ Convenient way of interaction
  - ❖ Integrated into many platforms

Mobile phones (e.g., Siri and Google Now)





Smart appliances



stand-alone assistants

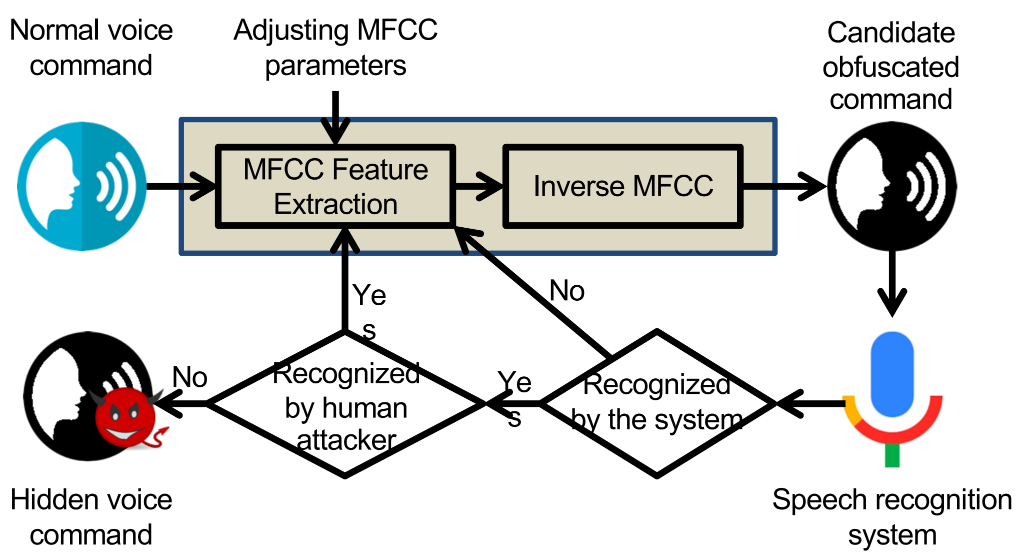
- ❑ Fundamental vulnerabilities due to the propagation properties of sound
- ❑ Emerging hidden voice commands
  - ❖ Recognizable to VCS
  - ❖ Incomprehensible to humans

Hidden Voice Command

- ❑ **Attacks the disparities of voice recognition between human and machine**
- ❑ **Iteratively shaping their audio features to meet the requirements**:
  - ❖ Understandable to VCSs
  - ❖ Hard to be perceived by the users
- ❑ **Attack model**
  - ❖ Internal attack – embedded in media and played by the target device
  - ❖ External attack – played via a loudspeaker in the proximity

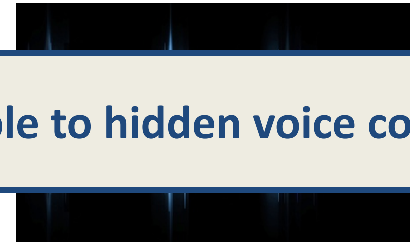
browse evil.com

call 911



Related Work

- ❑Defend acoustic attacks based on audio information
  - ❖Voice authentication models



Only relying on speech audio features is vulnerable to hidden voice commands

  - ❖Speech vocal features (e.g.,)

- ❑Speaker liveness detection

Restricted application scenarios by either requiring the microphone to be held close to mouth or additional dedicated hardware

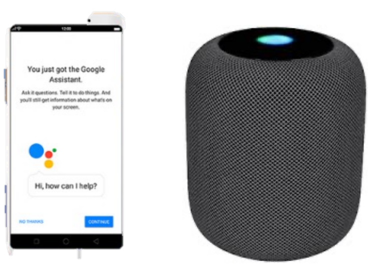
A multi-modality authentication framework is highly desirable to provide enhanced security:
Audio sending modality + vibration sensing modality

Basic Idea

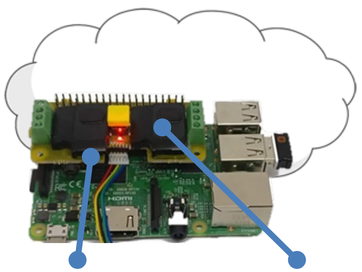
- Many VCS devices (e.g., smartphones and voice
  sensors

Basic Idea: utilizing the vibration signatures of the voice command to detect hidden voice commands

- Unique audio-induced surface vibrations captured by the motion sensor are hard to forge
- Two modes for capturing noticeable speech impact on motion sensors based on playback



Front-end playback



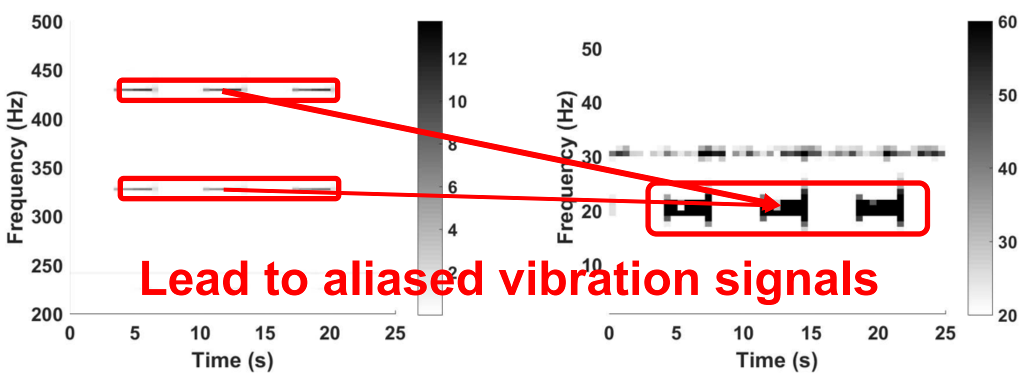
Back-end playback

Capturing Voice Using Motion Sensors

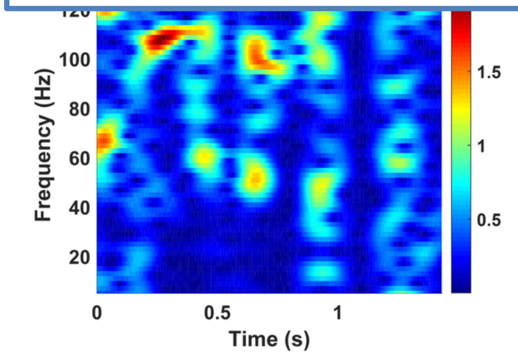
- ❑ Shared surface between loudspeaker and microphone
- ❑ Low sampling rate motion sensors (e.g., < 200Hz)
- ❑ Nonlinear vibration responses
- ❑ Distinct vibration domain

$$f_{alias} = |f - Nf_s|, N \in Z,$$

Played Audio **Vibration Responses**

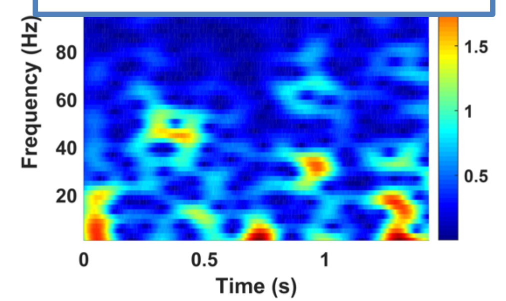


Down-sampled mic data



"show facebook.com"

Accelerometer data



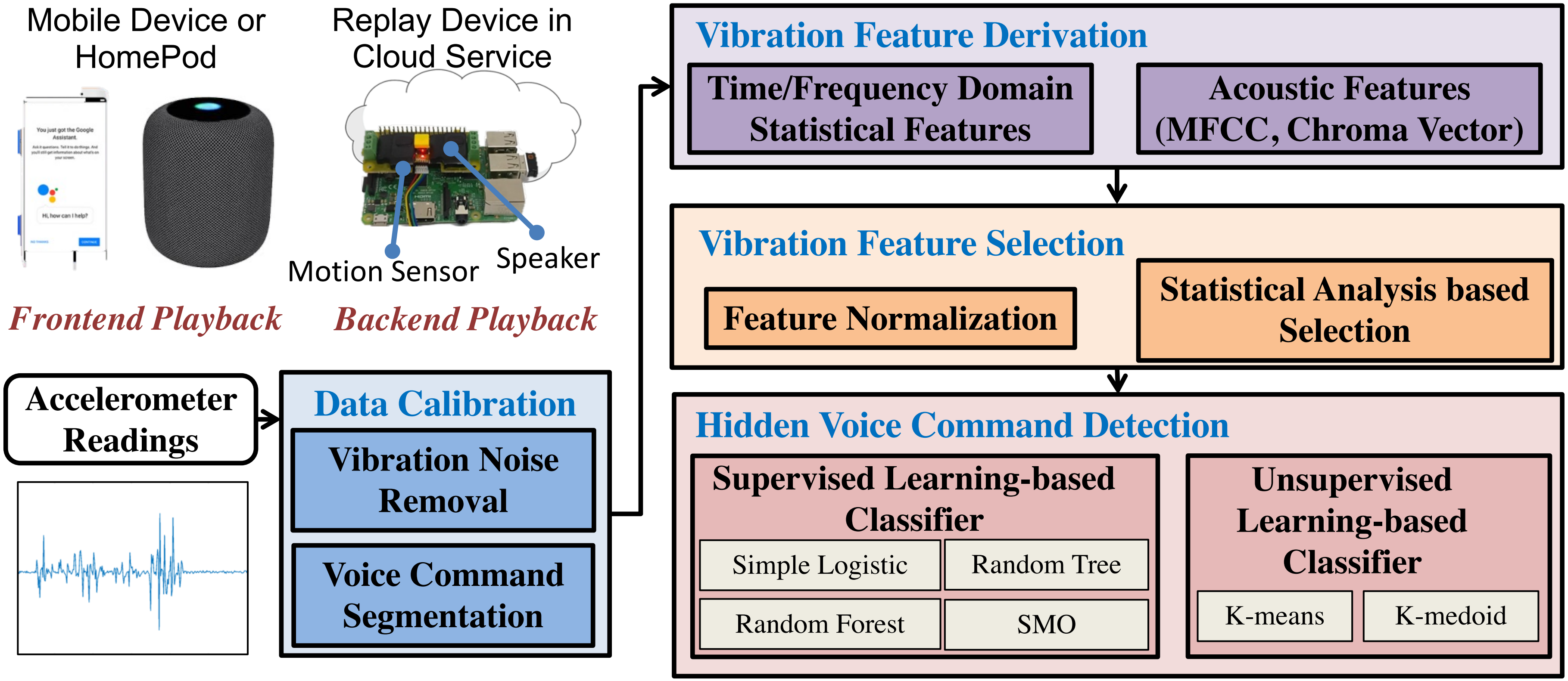
Why Vibration?

- ❑Existing speech/voice recognition methods based on audio domain voice vocal features
- ❑Hidden voice commands designed to duplicate these audio domain features by iteratively modify a voice command
- ❑Audio-induced surface vibrations
  - ❖An additional sensing domain, distinct to audio

The vibration domain approach can work in conjunction with the audio domain approach to more effectively detect the hidden voice commands.

physical vibrations, motion sensors)

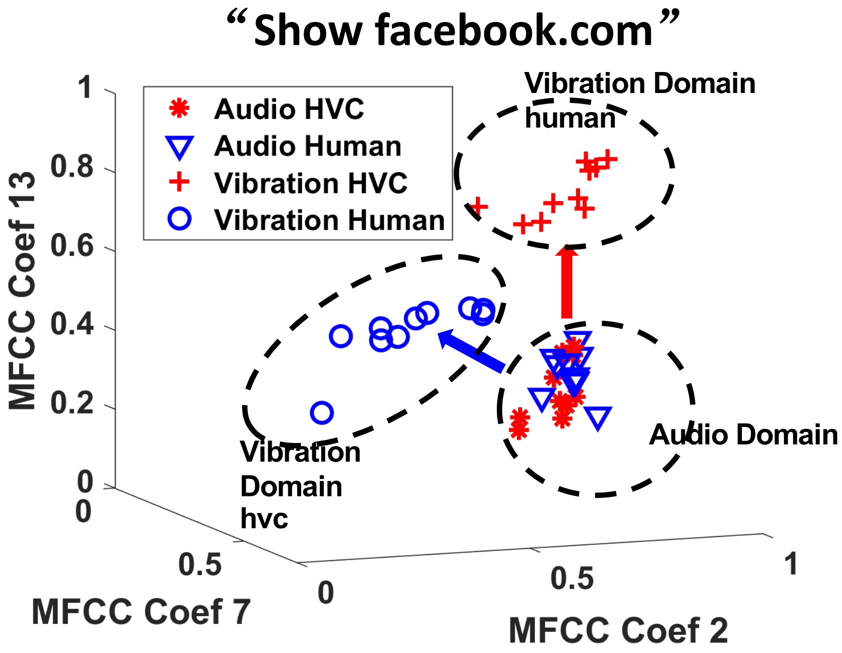
System Overview



Vibration Feature Derivation

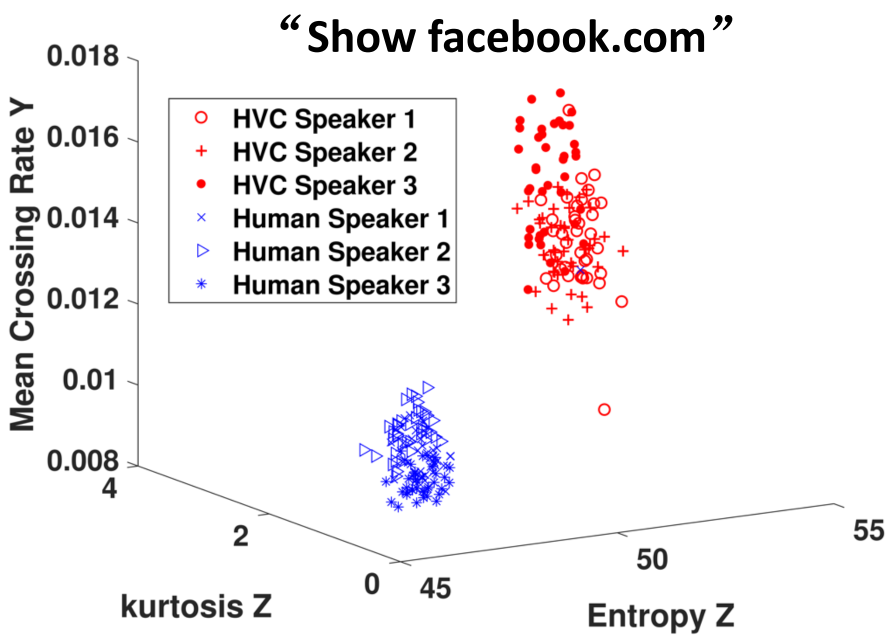
- ❑Unique and hard to forge
  - ❖Statistical features in time and frequency domains
  - ❖Deriving Acoustic Features from Motion Sensor Data
    - ➢MFCC
    - ➢Chrome vectors

- ❑Nonlinear relationship between audio features and vibration features



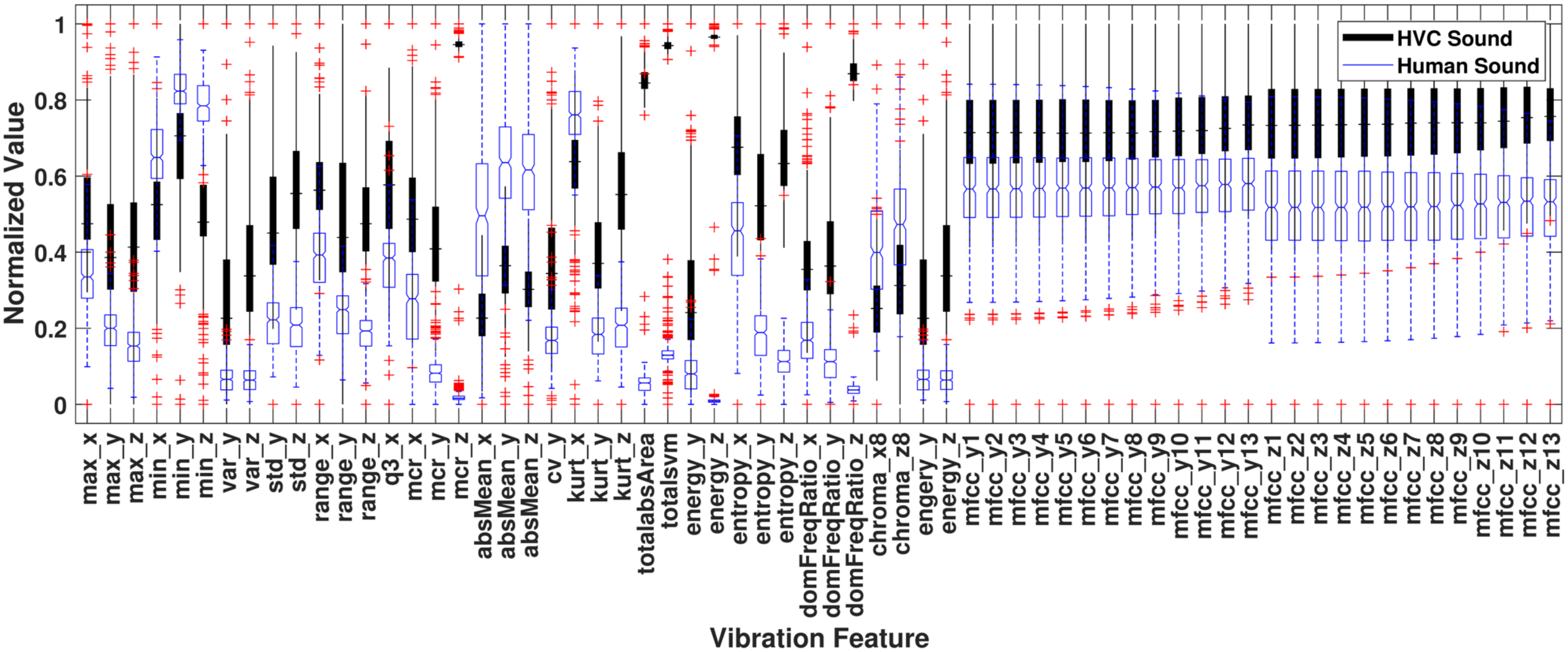
Vibration Feature Derivation

- Unique and hard to forge vibration features
  - Statistical features in time and frequency domains
  - Deriving Acoustic Features from Motion Sensor Data
    - MFCC
    - Chrome vectors
- Nonlinear relationship between audio features and vibration features
- Feature Selection Based on Statistical Analysis



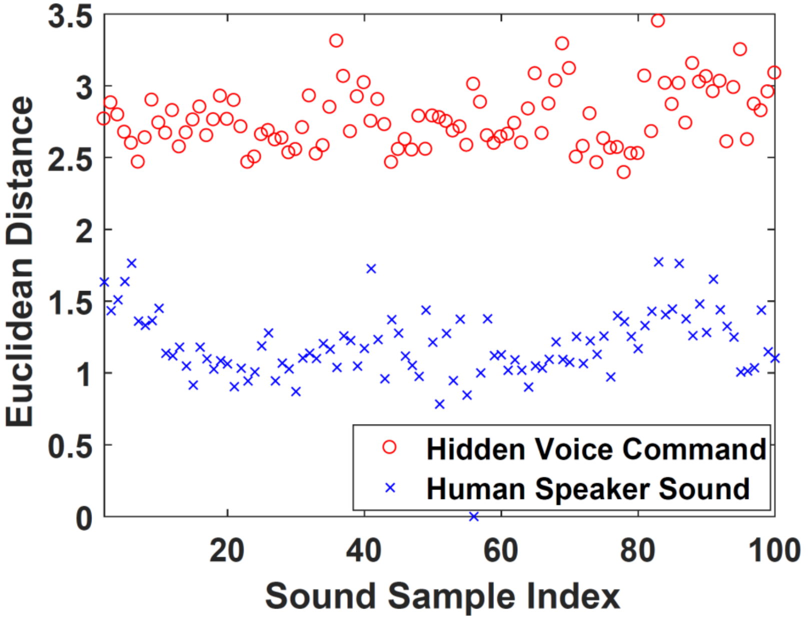
$$s = \frac{\bar{F}_{hid} - \bar{F}_{hum}}{max(\frac{\sqrt{\sum(F_{hid}(i) - \bar{F}_{hid})^2}}{n}, \frac{\sqrt{\sum(F_{hum}(j) - \bar{F}_{hum})^2}}{n})}$$

Feature Selection Based on Statistical Analysis



Hidden Voice Command Detection

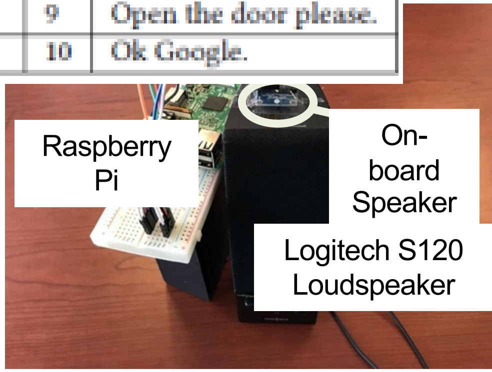
- ❑ Supervised Learning-based method
  - ❖ Simple Logistic
  - ❖ Support Vector Machine
  - ❖ Random Forest
  - ❖ Random Tree
- ❑ Unsupervised learning-based method
  - ❖ k-means/k-medoids based methods
  - ❖ Calculating the Euclidean distance of the voice command samples to the cluster centroid
  - ❖ Not require much training



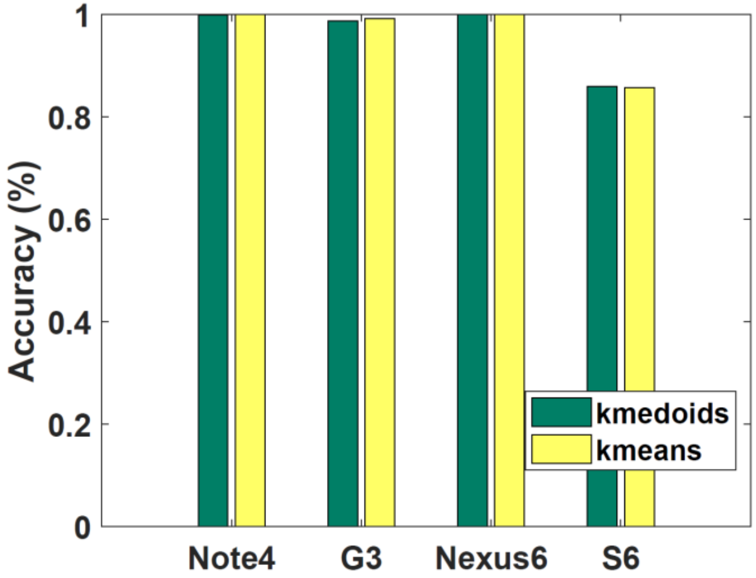
Experimental Setup

- ❑ Front-end playback setup
  - ❖ 4 different smartphones
  - ❖ On table
  - ❖ Held by hand
  - ❖ Placed on sofa
- ❑ Backend playback setup
  - ❖ Imitated cloud service device
  - ❖ Prototype on Raspberry Pi
- ❑ 10 voice commands, 5 speakers
- ❑ 13,000 vibration data traces
  - ❖ 6500 benign commands
  - ❖ 6500 hidden voice commands

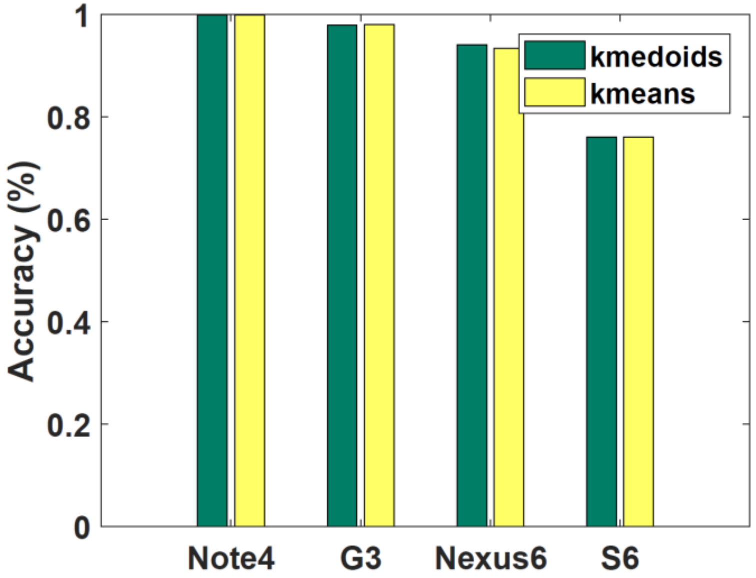
1	What's my current location?	6	Call 911.
2	Open Bank of America.	7	Open youtube.com.
3	Turn on airplane mode.	8	Show facebook.com.
4	Play country music.	9	Open the door please.
5	What's my schedule today?	10	Ok Google.



Performance Evaluation Unsupervised-learning



Front-end playback setup



Back-end playback setup

Up to 99% accuracy for both frontend and backend setups to differentiate normal commands from hidden voice commands

Performance Evaluation

- Partial playback to reduce delay

Front-end playback setup

	Note 4	G3	Nexus 6	S6
Replay all	100%	99.10%	100%	85.70%
Replay 1s	100%	89.10%	99.90%	85.60%
Replay 0.5s	99.90%	85.20%	95.90%	85%

Back-end playback setup

	Note 4	G3	Nexus 6	S6
Replay all	99.90%	97.90%	93.40%	76%
Replay 1s	92.9	99.10%	92.40%	75.90%
Replay 0.5s	88.5	90.20%	90.50%	73.80%

- Various mobile device usage scenarios of frontend playback setup

	Table	Held in hand	Placed on sofa	80%vol. on table	2x speed on table
Kmed	100%	87.30%	100%	100%	88.30%
Kmea	100%	87.30%	100%	100%	85.20%

Take-aways

- Demonstrate that hidden voice commands can be detected by their speech features in the vibration domain
- Derive the unique vibration features (statistical features in the time and frequency domains and speech features to distinguish hidden voice commands from normal commands
- Develop both supervised and unsupervised learning-based systems to detect hidden voice commands
- Implemented the proposed system in two modes: frontend playback and backend playback

Practical Adversarial Attacks Against Speaker Recognition Systems

What’s Speaker Recognition?

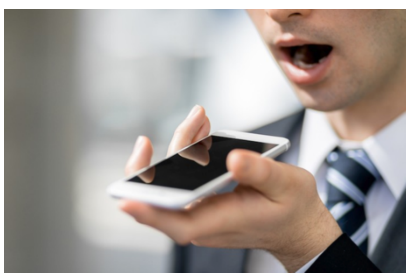
❑Speaker Recognition (SR)

Who is this?

Enrolled Speakers	Score	Result
	95	✔
	40	✖
	60	✖

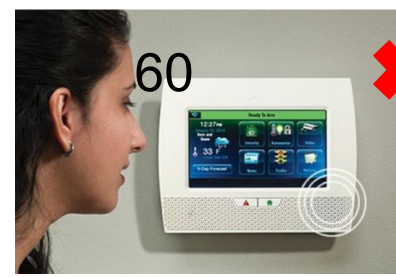
❑Applications

❖ Smartphone



❖ Telephone Banking

❖ Access Control



Bixby

CHASE

WELLS FARGO

VOICE BIOMETRICS GROUP

RUTGERS

WINLAB

Attack Chances on Speaker Recognition

- Trend in Speaker Recognition
  - Adopting *Deep Neural Networks (DNNs)* for better performance [1]

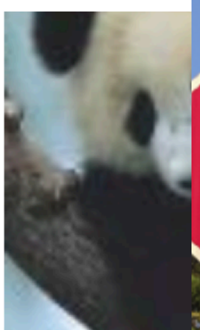
- DNNs are vulnerable to *adversarial examples* [2, 3]

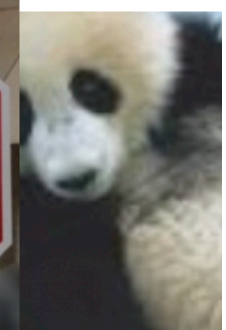
Recognized as **Panda**

Recognized as **Stop**

Recognized as **Speed Limit 45**

Recognized as **Gibbon**



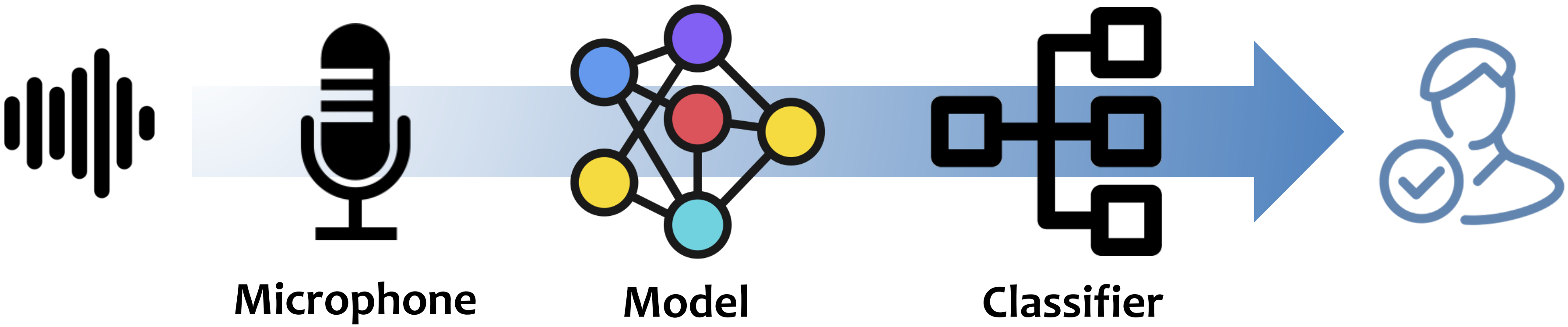


Benign Input **Benign Input** **Perturbation** **Adversarial Example** **Adversarial Example**

[1] Mitchell McLaren, Yun Lei, and Luciana Ferrer. 2015. Advances in deep neural network approaches to speaker recognition. In IEEE ICASSP 2015.
[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572 (2014).
[3] Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

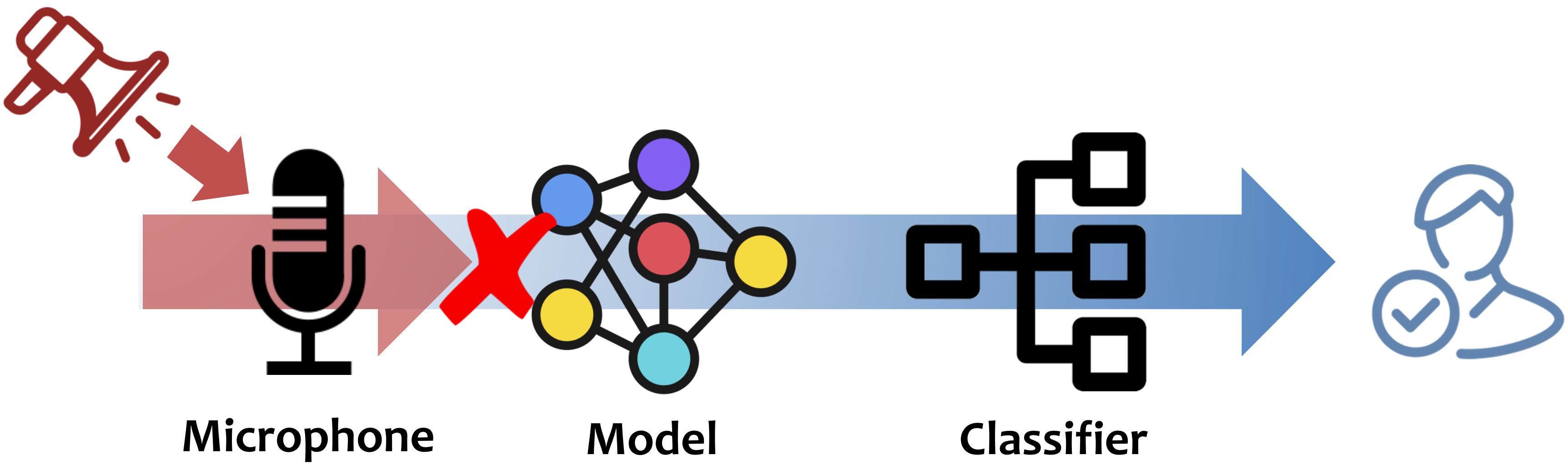
Limitation of Existing Attacks

- Speaker Recognition Pipeline



Limitation of Existing Attacks

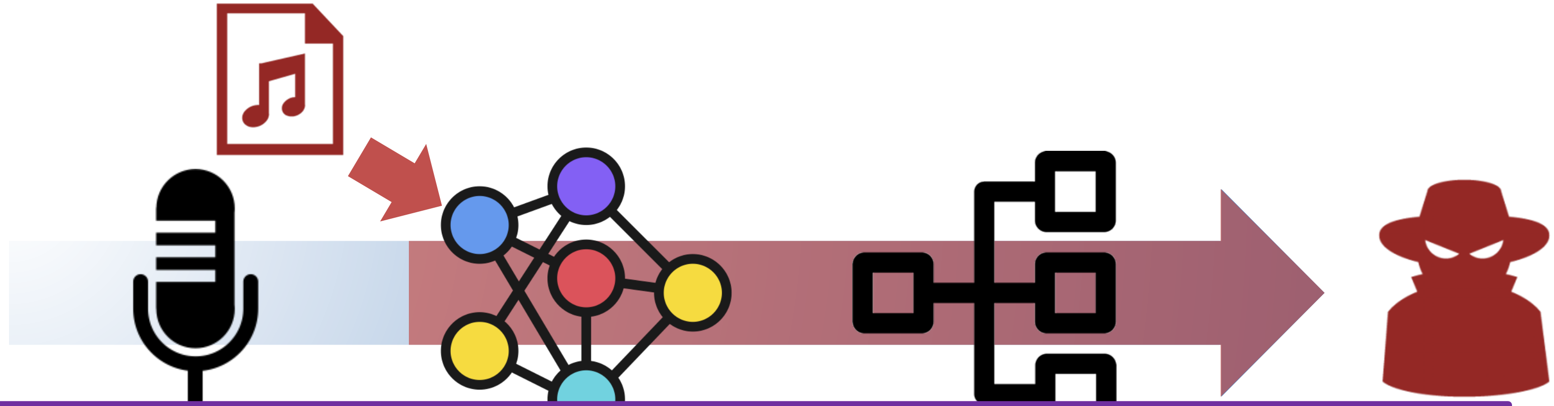
- Conventional Attacks
  - Replay attack, synthesis attack, voice conversion attack
  - **Pros**: injected via physical channel
  - **Cons**: can be defended by modern SR models [4, 5]



[4] Hong Yu, Zheng-Hua Tan, Yiming Zhang, Zhanyu Ma, and Jun Guo. 2017. DNN filter bank cepstral coefficients for spoofing detection. IEEE Access 5 (2017), 4779–4787.
[5] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah. 2012. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In IEEE APSIPA ASC 2012. 1–5.

Limitation of Existing Attacks

- Adversarial Attack
  - Leverage adversarial examples
  - **Pros**: strong, can fool state-of-the-art model
  - **Cons**: success in digital domain, sensitive to over-the-air distortions



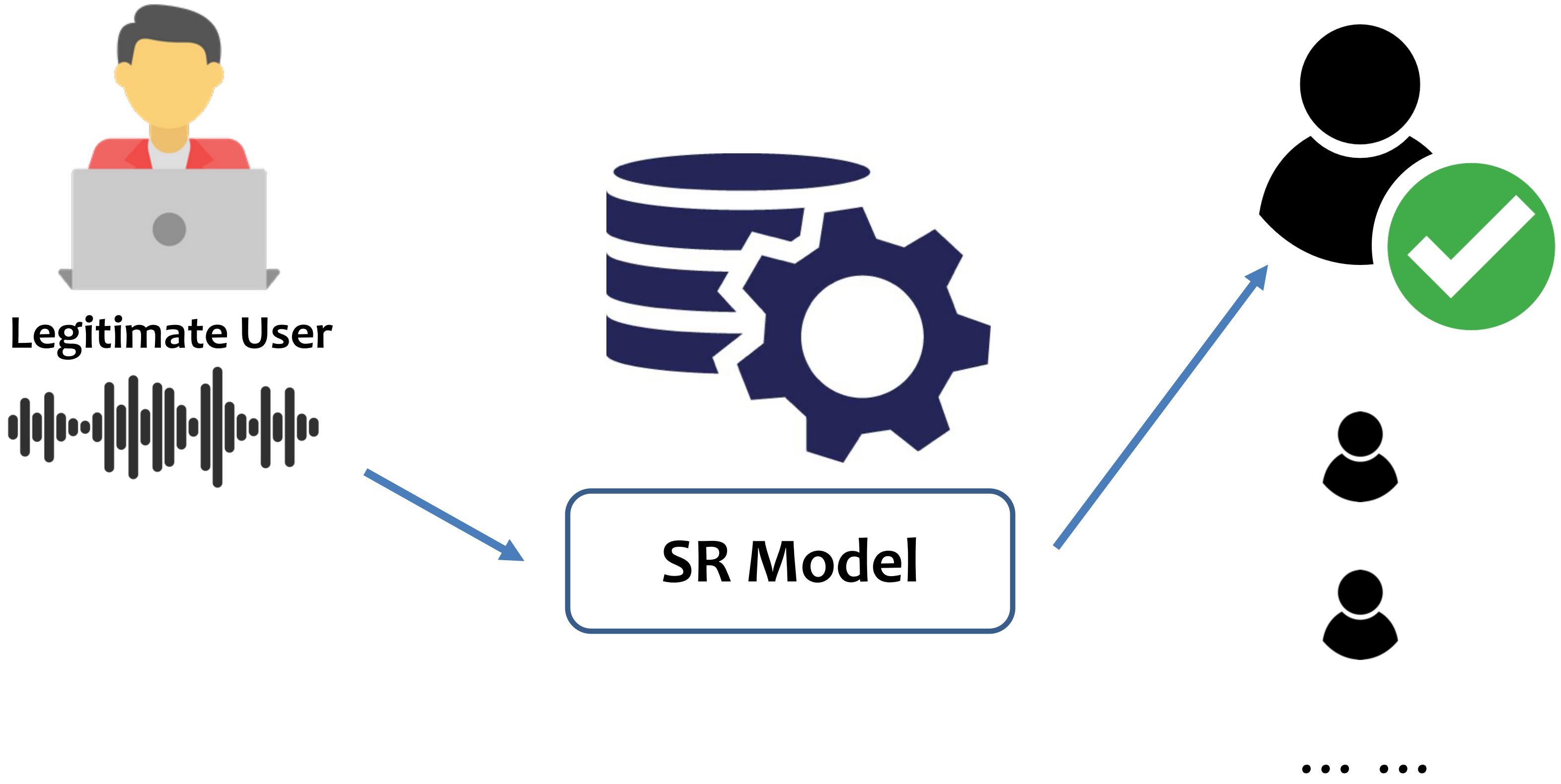
Our goal: Design a practical over-the-air adversarial attack against state-of-the-art speaker recognition system

RUTGERS WINLAB

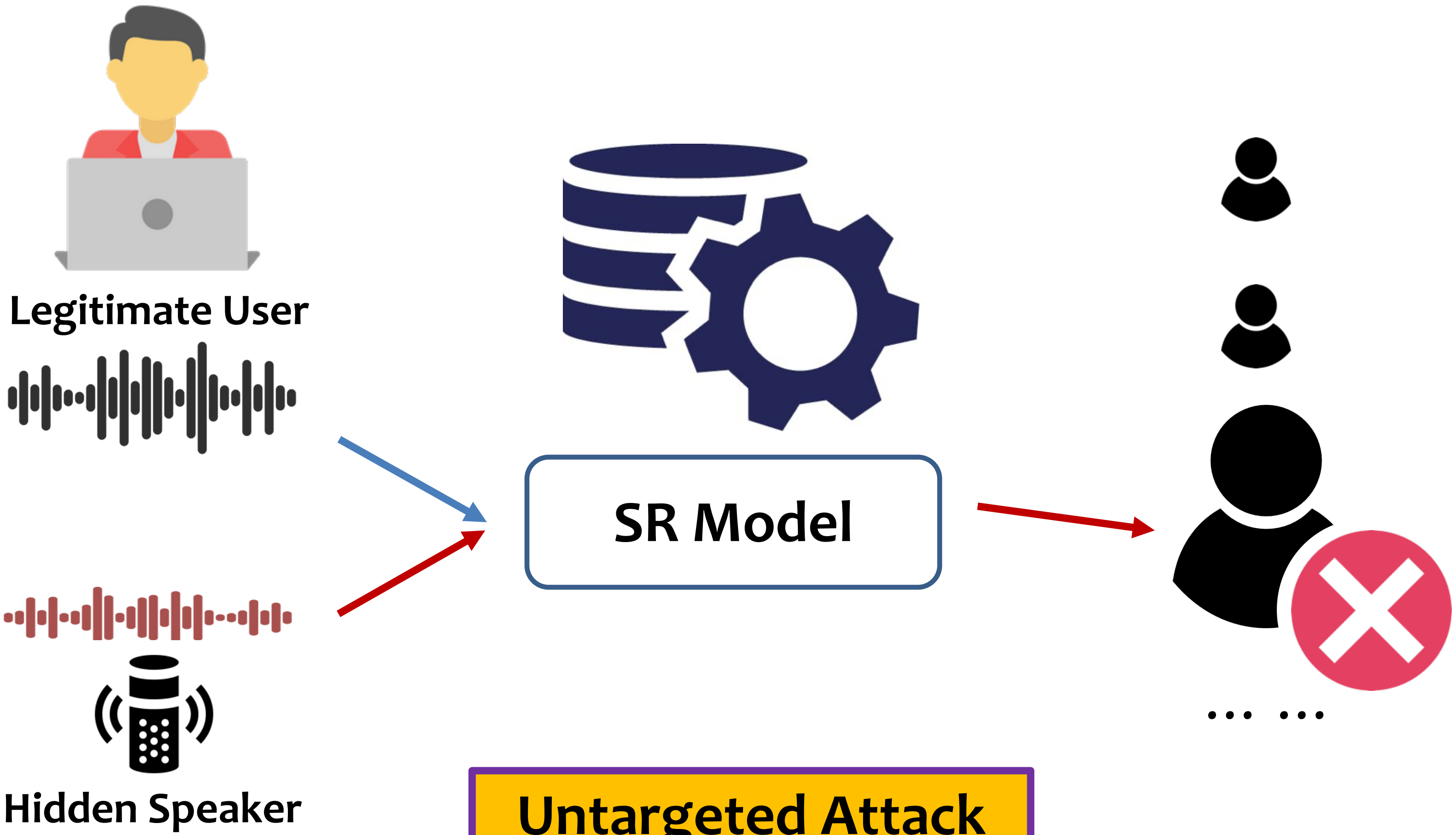
Contribution

- First practical adversarial attack against **multi-class** SR system
- Use the estimated room impulse response to launch **over the air attack**
- Implement gradient-based algorithms to make the attack **unnoticeable**
- Evaluate on a public dataset of 109 English speakers

Threat Model

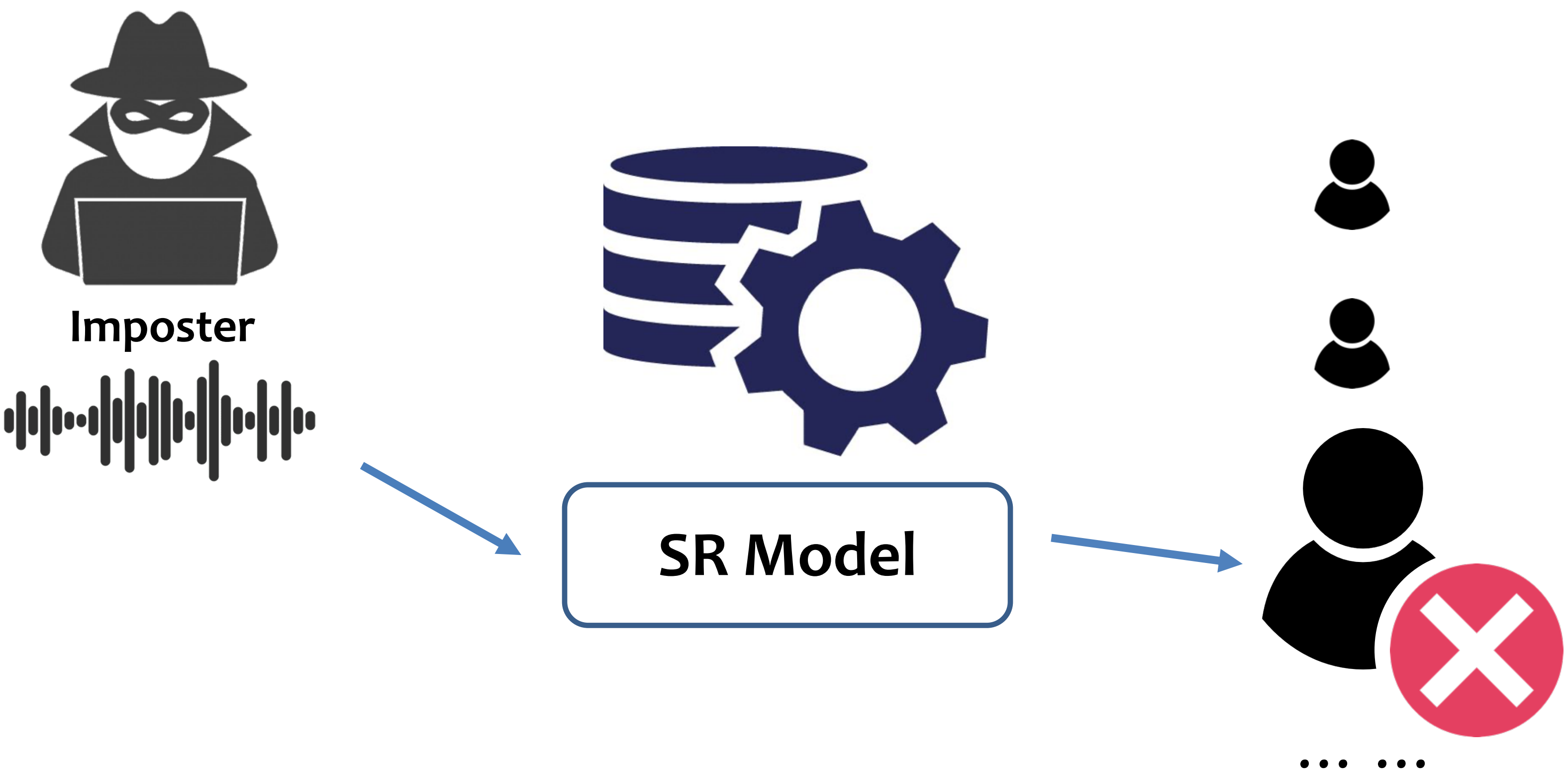


Threat Model

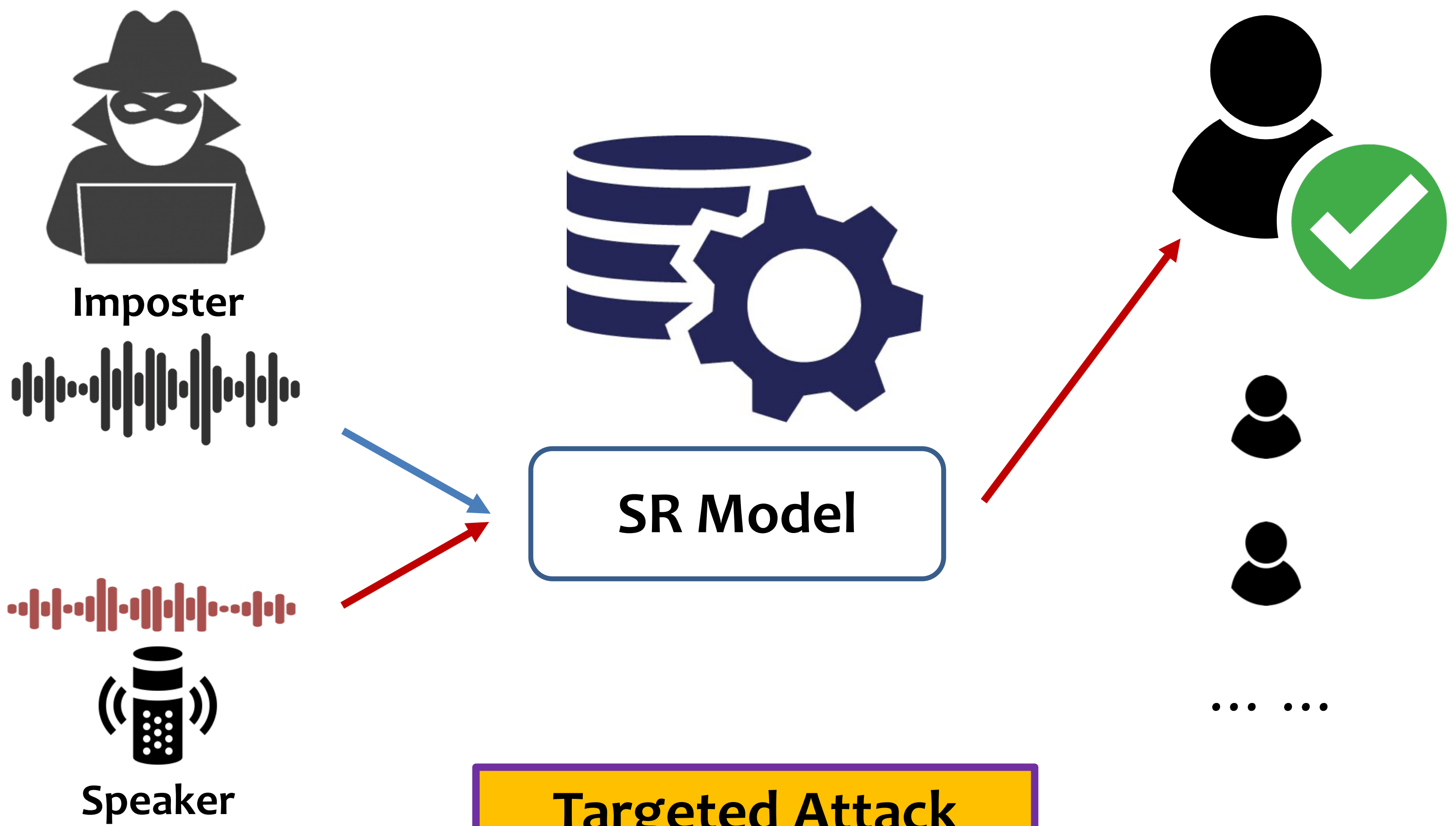


Untargeted Attack

Threat Model



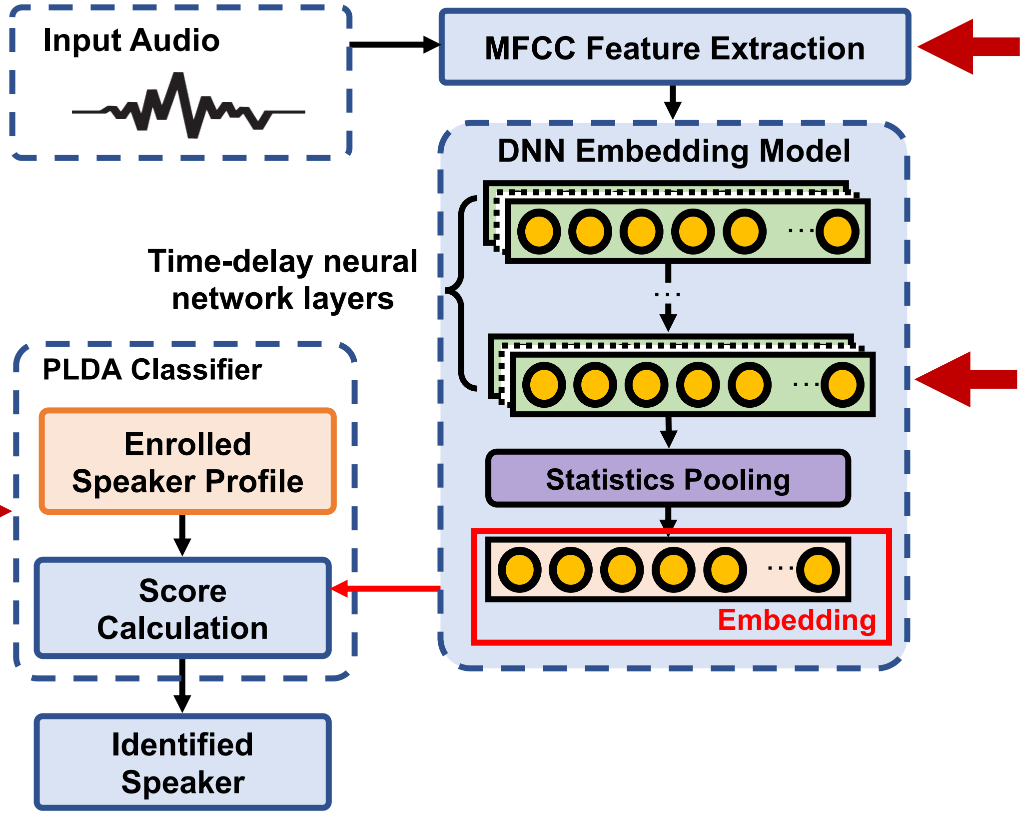
Threat Model



Targeted Attack

Target **Model**

- X-vector [6]
  - The **state-of-the-art *DNN-based multi-class*** speaker recognition model
  - **Components**
    - Mel Frequency Cepstral Coefficients (MFCC)
    - Embedding Model
    - Probabilistic Linear Discriminant Analysis (PLDA)



[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In IEEE ICASSP 2018.

Problem Formulation

- Threat Model
  - White-box
- Notation

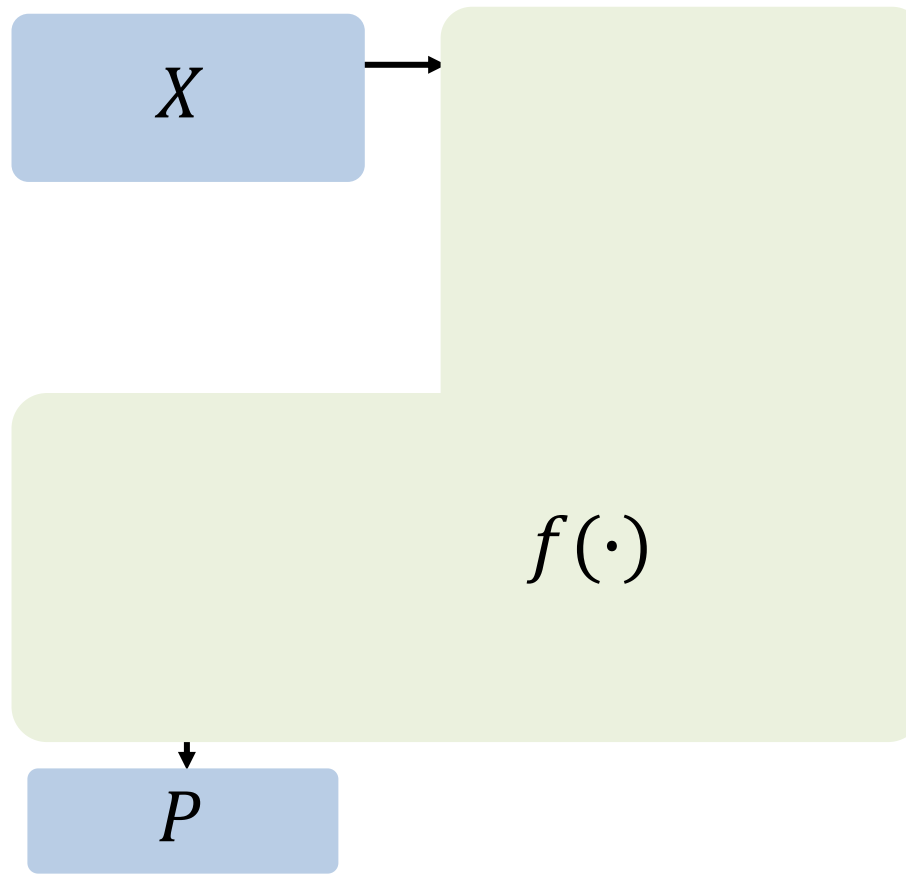
  Input audio – X, original label y

  Embedding model – $f: X \rightarrow P$

  Probability vector – $P = [p_1, \dots, p_i]$
- Untargeted Attack
  - Find minimal δ

s.t. $argmax(f(X+\delta)) \neq argmax(y)$
- Targeted Attack
  - Find minimal δ

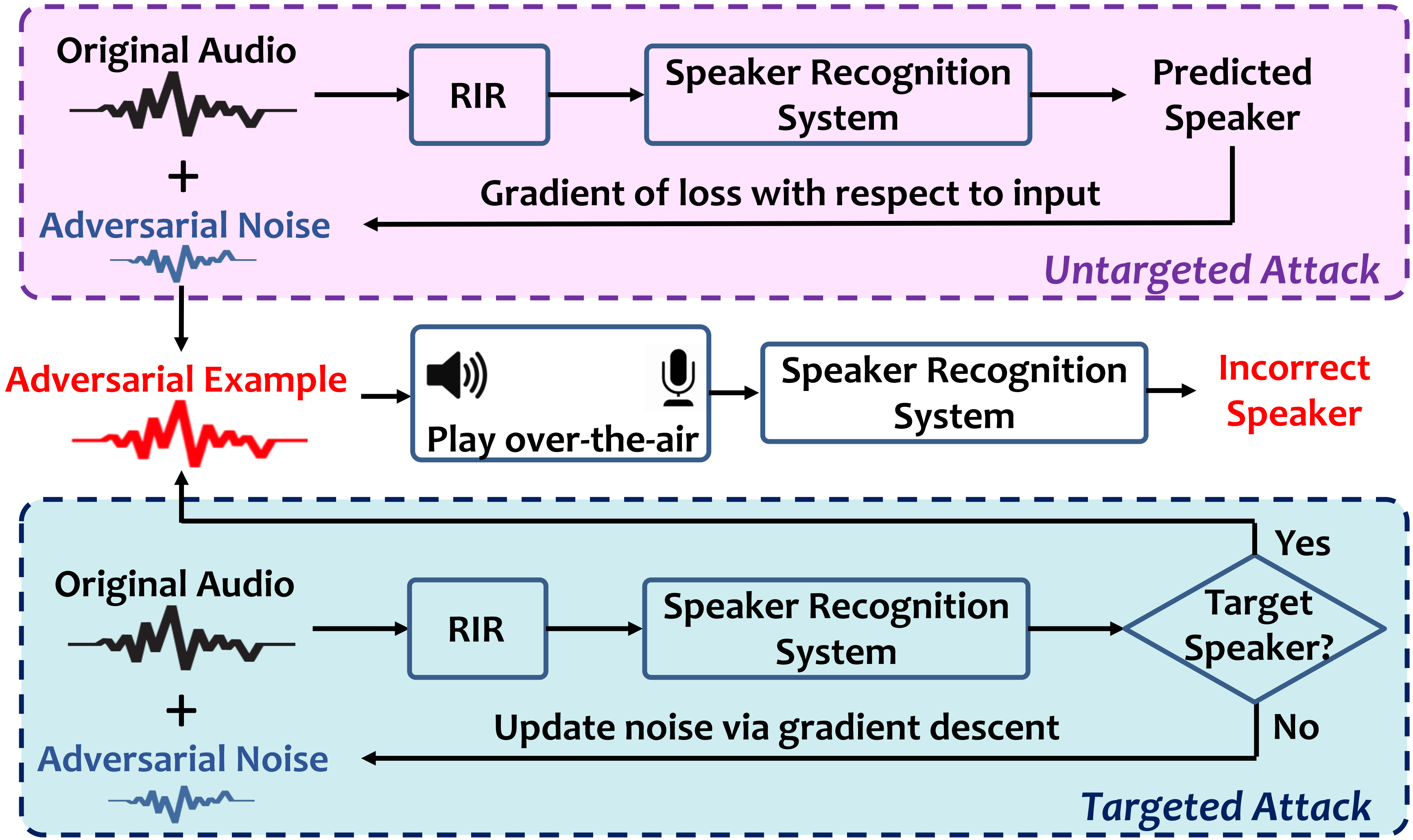
s.t. $argmax(f(X+\delta)) = argmax(y_t)$



RUTGERS

WINLAB

Attack Overview



RUTGERS WINLAB

Room Impulse Response Estimation

- Room Impulse Response (RIR) – $h(t)$
  - Model the transfer function between the played audio $x(t)$ and the received audio $y(t)$

$$y(t) = x(t) \otimes h(t)$$

- RIR estimation
  - Play an excitation signal $x_e(t)$

$$x_e(t) = sin\left(\frac{2\pi f_1 T}{ln(\frac{f_2}{f_1})}\left(e^{\frac{t}{T} ln(\frac{f_2}{f_1})} - 1\right)\right)$$

  - Record the response $y_e(t)$
  - Estimate RIR, where $f(t)$ is the time-reversal of $x_e(t)$

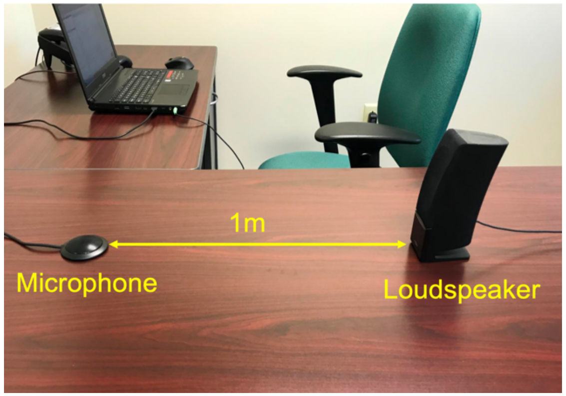
$$h(t) = y_e(t) \otimes f(t)$$

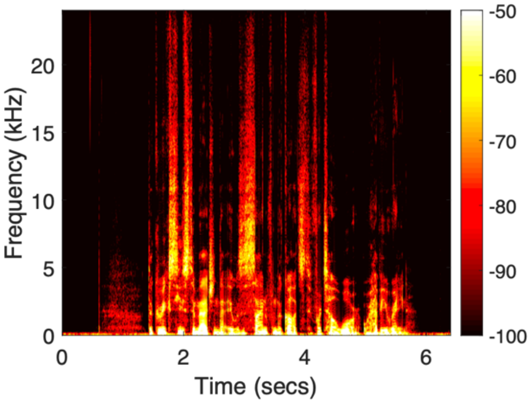
RUTGERS

WINLAB

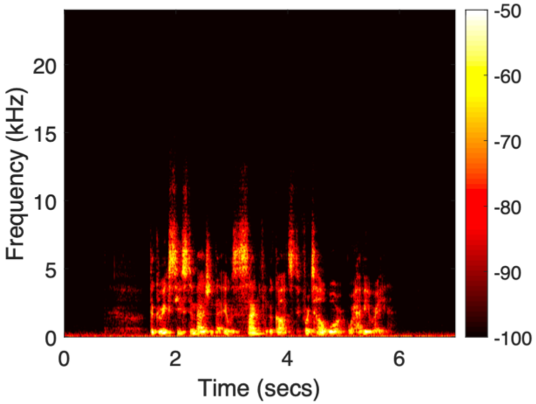
Room Impulse Response Estimation

- Preliminary Experiment
  - $f = 20 - 20kHz$, $\mathrm{T} = 5s$
  - Measured Mean Square Error (MSE)
    - Recorded & Predicted = 0.112
    - Original & Recorded = 0.84

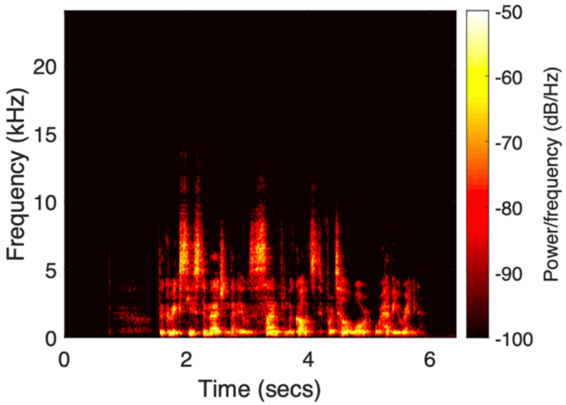




Original Signal



Recorded Signal



Predicted Signal (w/RIR)

RUTGERS

WINLAB

Adversarial Example Generation

- Untargeted Attack
  - Due to the local linearity of DNN models, a **linear perturbation** is sufficient for untargeted attacks [7]:

$$\begin{cases} X' = X + \delta \\ \delta = \epsilon sign\left(\nabla_X J(X, y)\right) \\ J(X, y) = -y \cdot log(P) \end{cases}$$

  - ***Digital untargeted adversarial example***

$$X' = X + \epsilon sign\Big(\nabla_X\big(- y \cdot log(f(X))\big)\Big)$$

  - ***Practical untargeted adversarial example***

$$X' = X + \epsilon sign\Big(\nabla_X\big(- y \cdot log(f(X \otimes h))\big)\Big)$$

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572 (2014).

Adversarial Example Generation

- Targeted Attack
  - Adversarial example targeting at label y_t can be generated through solving an **optimization problem**:

$$\text{minimize } ||\delta||_2, \text{ s.t. } f(X+\delta) = y_t$$

  - Lagrangian relaxation:

$$\text{minimize } -y_t \cdot log(f(X+\delta)) + c||\delta||_2$$

  - Apply **gradient descent** to find the optimal δ^*
  - ***Digital targeted adversarial example***

$$X' = X + \delta^*$$

  - ***Practical targeted adversarial example***

$$\text{minimize } -y_t \cdot log\Big(f\big((X+\delta) \otimes h\big)\Big) + c||\delta||_2$$

RUTGERS WINLAB

Experimental Methodology

- Dataset
  - CSTR VCTK Corpus
  - Total 44217 utterances spoken by 109 English speakers with various accents, training & testing ratio = 4:1
- Baseline Model
  - 30 dimensional MFCC with frame length of 25 ms
  - Pretrained X-vector model provided in *Kaldi* [8]
- Evaluation Metrics
  - Speaker Recognition Accuracy (%)
  - Attack Success Rate (%)
  - Distortion Metric (dB)

[8] Povey et al., The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Evaluation of Digital Attacks

❑Digital Untargeted Attack

❖ Test set : 8896 audio files

Attack Strength (i.e., ϵ)	No Attack	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Speaker Recognition Accuracy (%)	92.81	84.71	41.33	12.11	2.23	1.37
Attack Success Rate (%)	—	8.73	55.47	86.95	97.60	98.52
Average Distortion (dB)	—	−89.06	−69.15	−49.24	−29.33	−9.41

❑ Digital Targeted Attack

❖ Tested on all original-target speaker combinations (total 109*108 pairs)

Attack Strength (i.e., c)	0.4	0.2	0.1	0.05
Attack Success Rate (%)	77.64	86.05	93.27	96.01
Average Distortion (dB)	−34.22	−32.43	−29.66	−25.94

RUTGERS

WINLAB

Evaluation of Practical Attack

❑Experimental Setup

❖ **Two realistic scenarios**: office & apartment

❖ 10 digital/practical targeted adversarial example tested in each scenario

(a) Office

(b) Apartment

	Playing digital adversarial examples	Playing practical adversarial examples
Office	0%	50%
Apartment	10%	50%

RUTGERS

WINLAB

Audio Samples

- Making Speaker #1 recognized as Speaker #20
  - Original audio
    - Recognized as ***Speaker #1***
  - Practical adversarial audio
    - Misrecognized as ***Speaker #20***
    - *Measured distortion:* $-42.35dB$
  - Genuine speech from ***Speaker #20***

Take-aways

- We demonstrate a practical and systematic adversarial attack against DNN-based speaker recognition systems
- Apply gradient-based algorithms to launch both untargeted and targeted attacks
- Integrate the estimated RIR into the adversarial example generation for a more practical attack
- Conduct extensive experiment in both digital and real-world settings

RUTGERS

WINLAB

Future work: Security Issues on *Voice Recognition Systems at the edge*

- Attacker could control your smart home

Future work: Security Issues on *Augmented Reality* (AR) System

- Attacker could control your 'reality'

Thanks to my collaborators and students