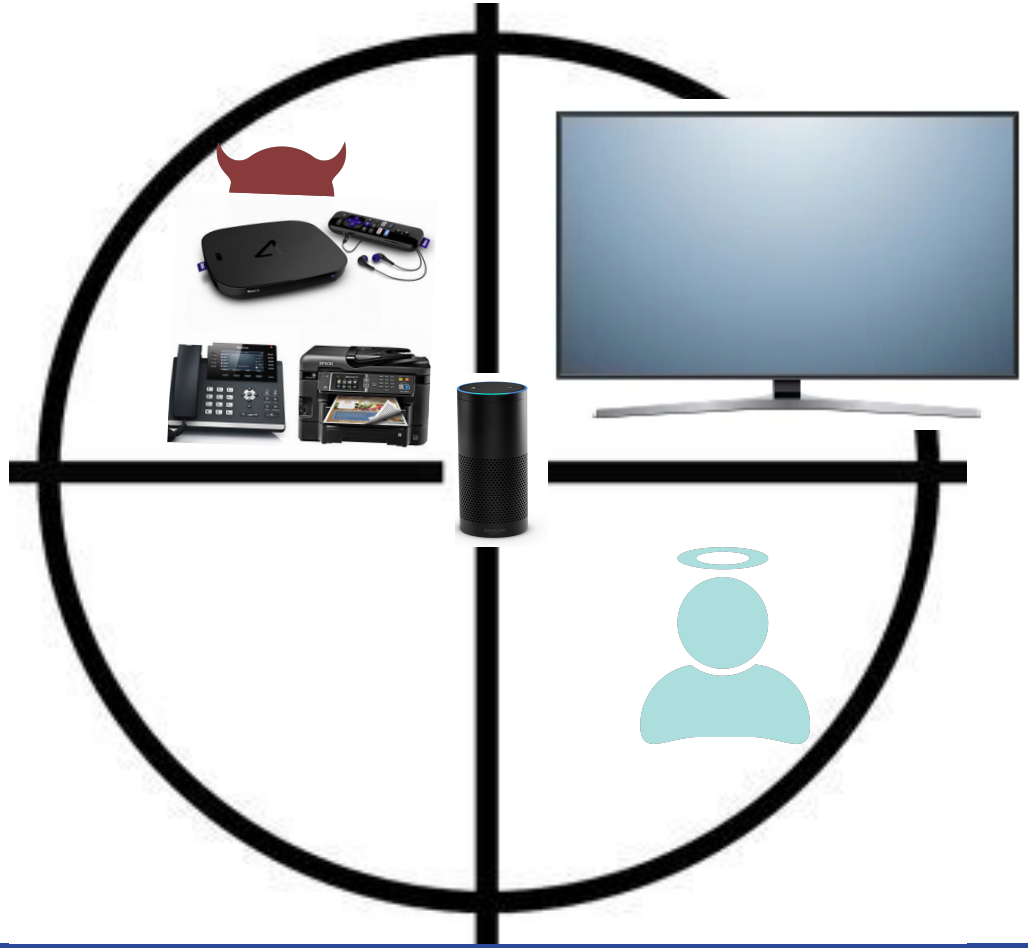
Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems

Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R.B. Butler, and Joseph Wilson

Voice as an Interface

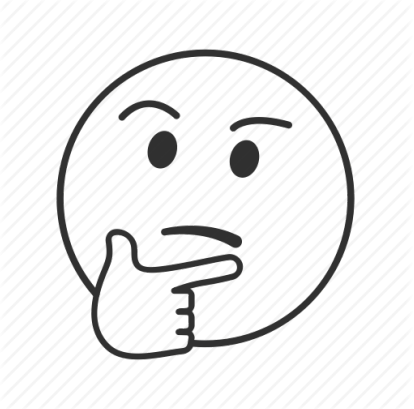
Injecting Commands



Injecting Commands (Demonstration)

You might be thinking...

- ~~*Benign?*~~
- ~~*Easy to Defend?*~~



Is there a **generic**, **transferrable** way to produce audio that:

- sounds like noise to humans,
- sounds like a valid command to the system?
- works against both **speech** and **speaker** recognition systems
- with **Black-Box** access to target system

Modern Speech Recognition Systems

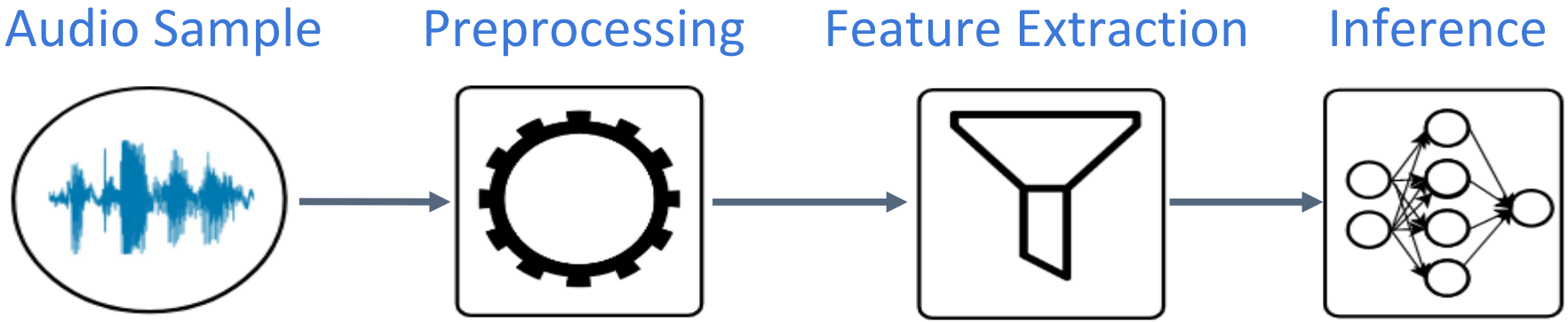
Feature Extraction

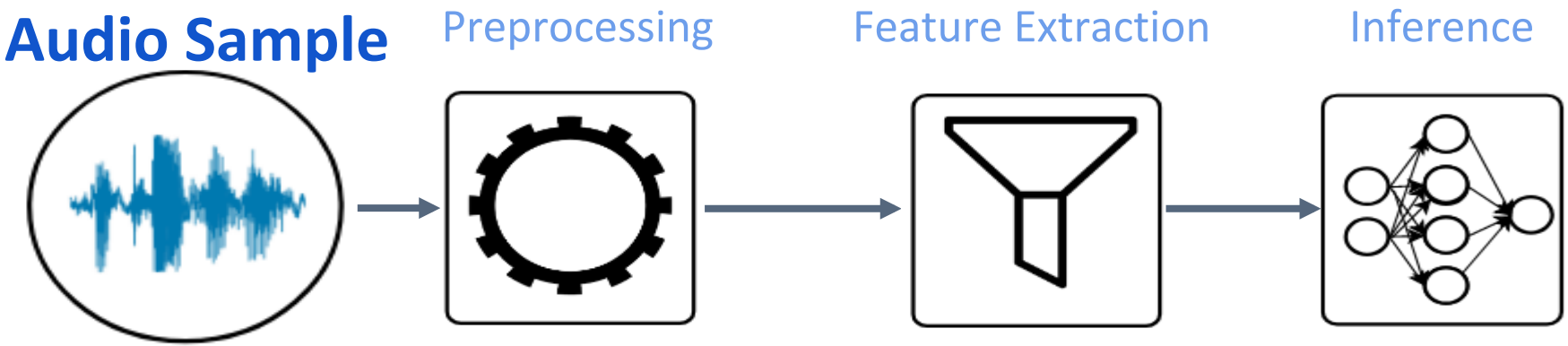
How the Attack Works

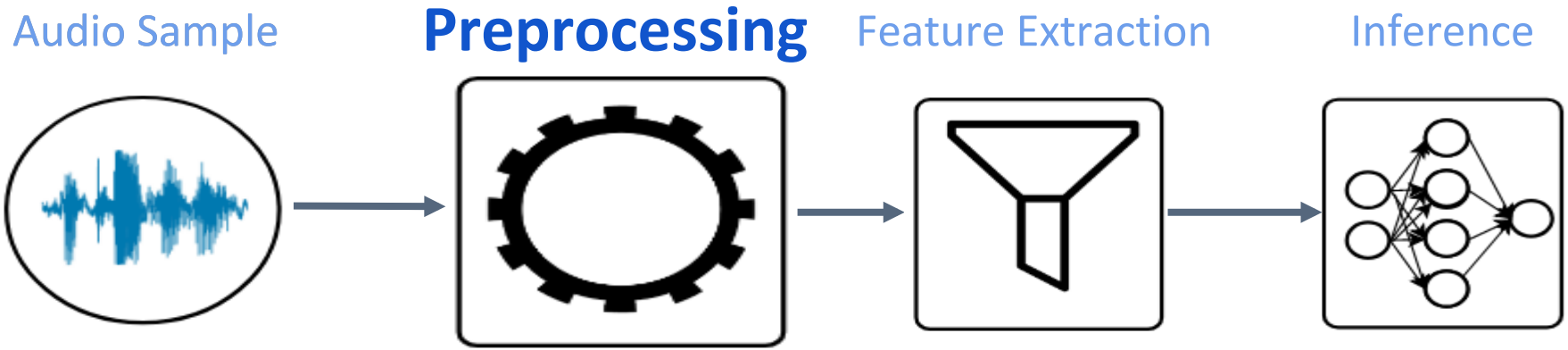
Demo

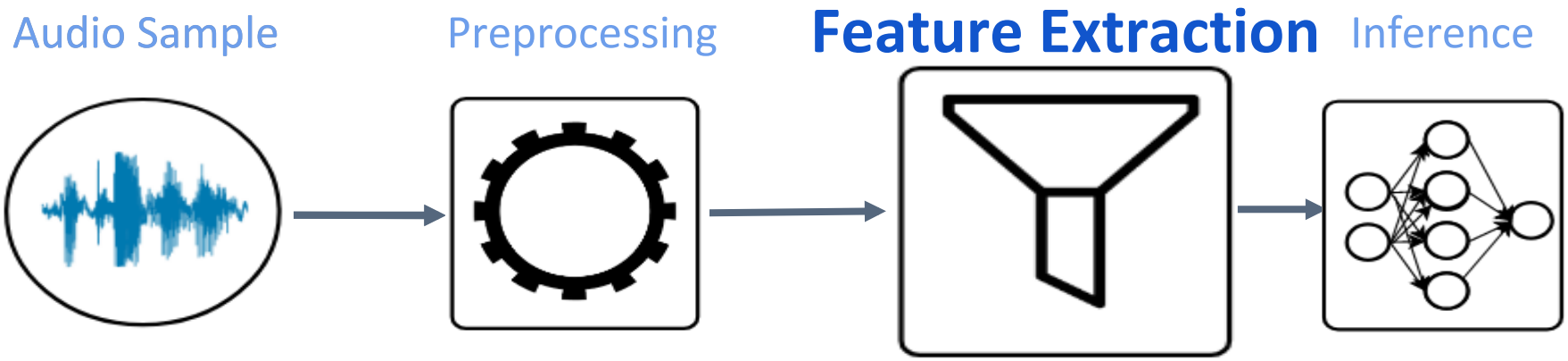
Takeaway

Modern Speech Recognition Systems

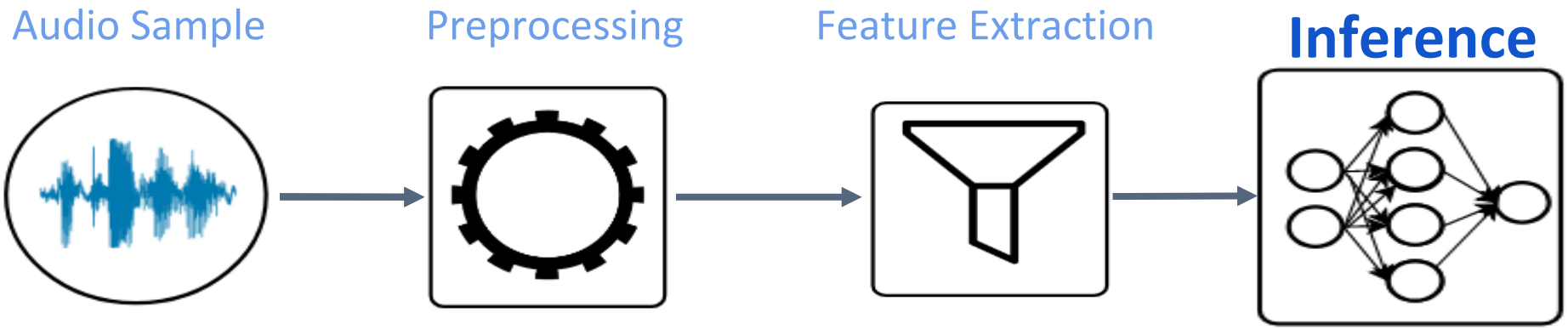



Audio Sample
Preprocessing
Feature Extraction
Inference

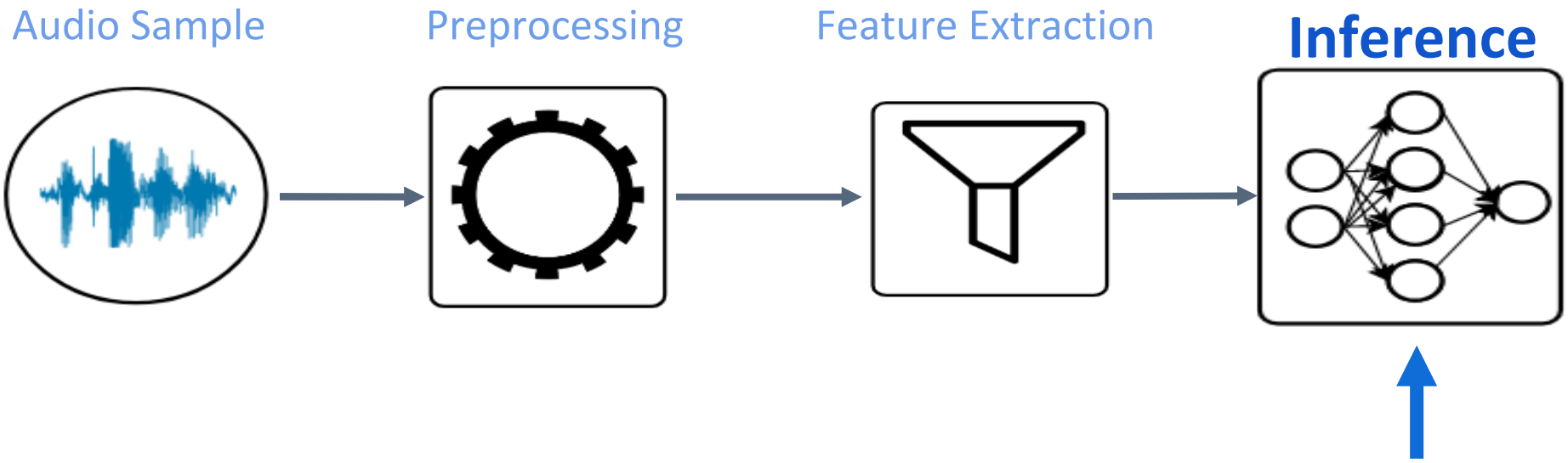

Audio Sample
Preprocessing
Feature Extraction
Inference


Audio Sample
Preprocessing
Feature Extraction
Inference

Modern Speech Recognition Systems



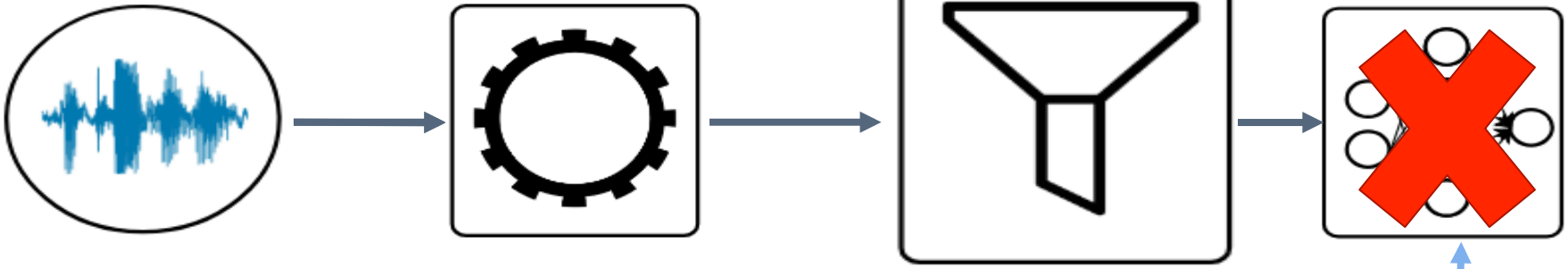
Modern Speech Recognition Systems



*

- N. Carlini and D. Wagner, "*Audio Adversarial Examples: Targeted Attacks on Speech-to-Text,*" IEEE Deep Learning and Security Workshop, 2018
- N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "*Hidden voice commands." in USENIX Security Symposium,* 2016
- X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "*Commandersong: A systematic approach for practical adversarial voice recognition*," in Proceedings of the USENIX Security Symposium, 2018.
- M. Alzantot, B. Balaji, and M. Srivastava, "*Did you hear that? Adversarial examples against automatic speech recognition*," NIPS 2017 Machine Deception Workshop

Modern Speech Recognition Systems



Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

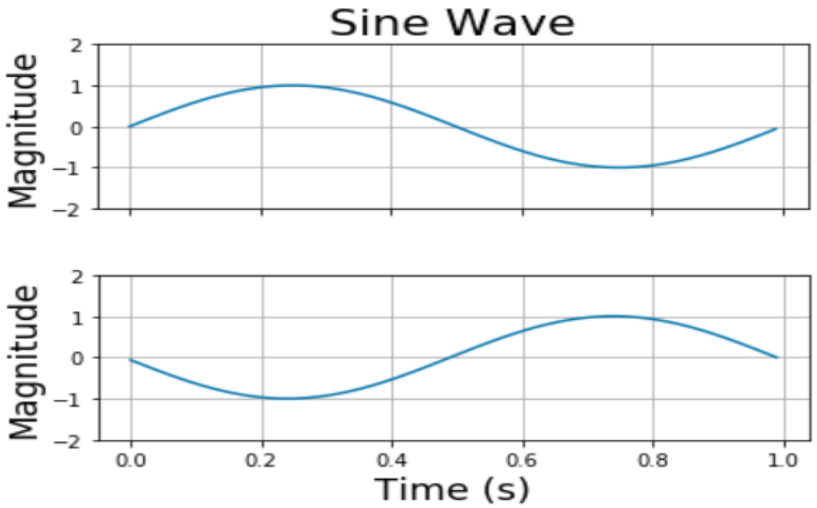
Demo

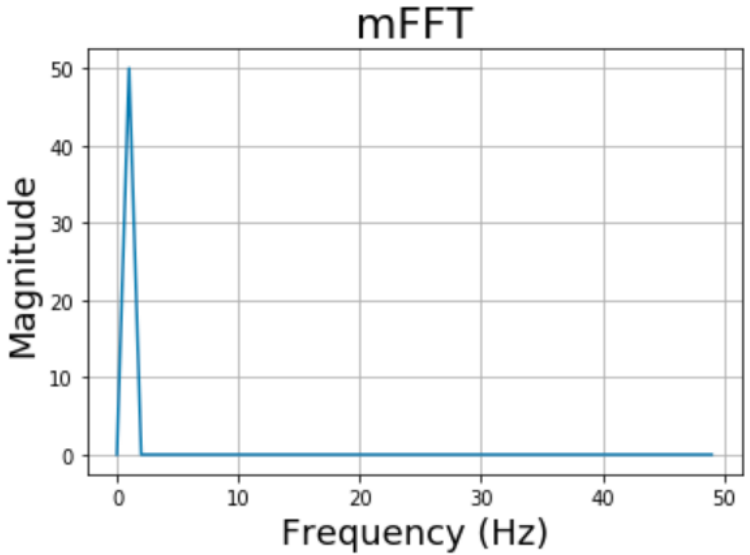
Takeaway

Feature Extraction

- Designed to **approximate** the **human ear**
- **Retains** the most important **features**
- Magnitude Fast Fourier Transform **(mFFT)**

- Converts **time domain** to **frequency domain**
- **Multiple Inputs** can have **same output**
- mFFT is **lossy**





Modern Speech Recognition Systems

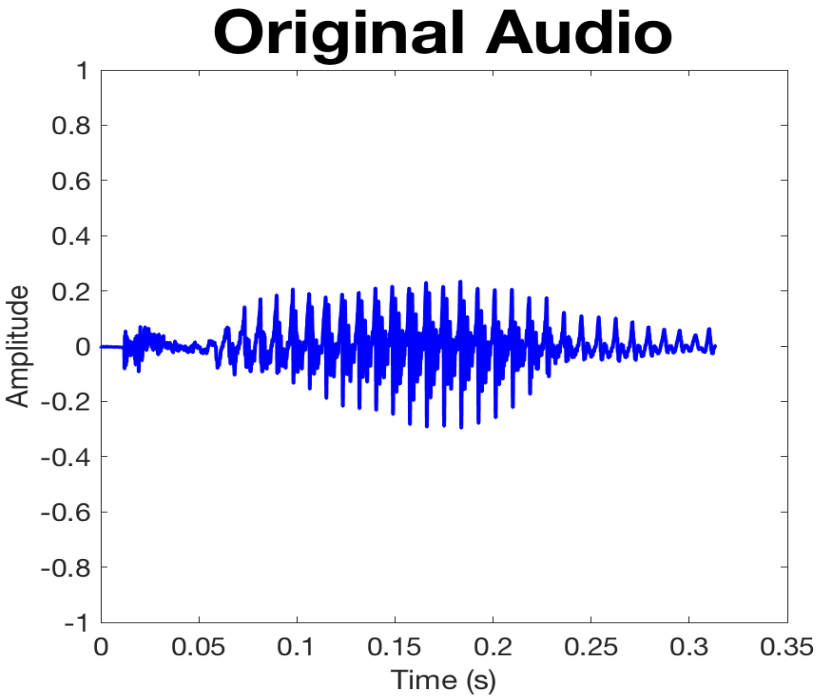
Feature Extraction

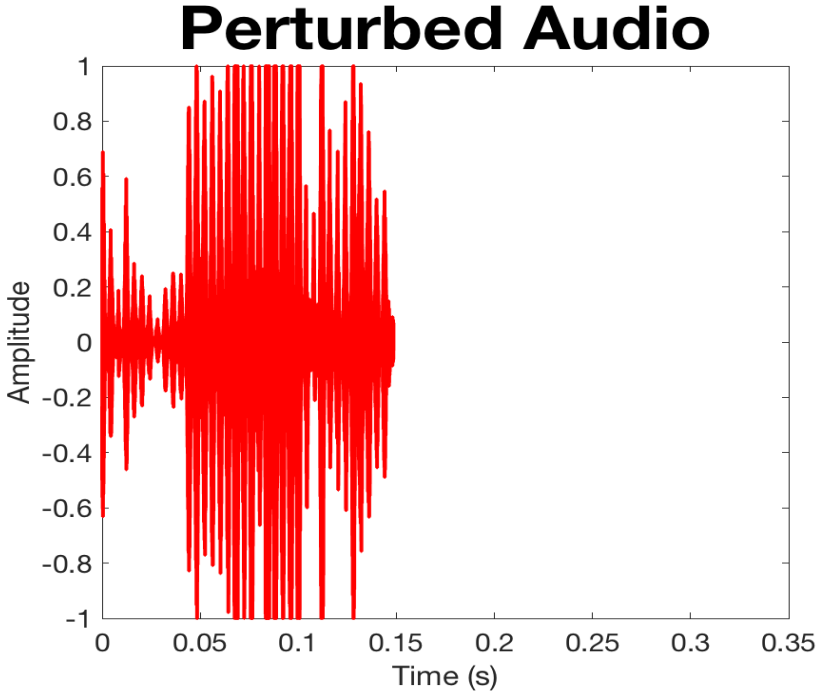
How the Attack Works

Demo

Takeaway

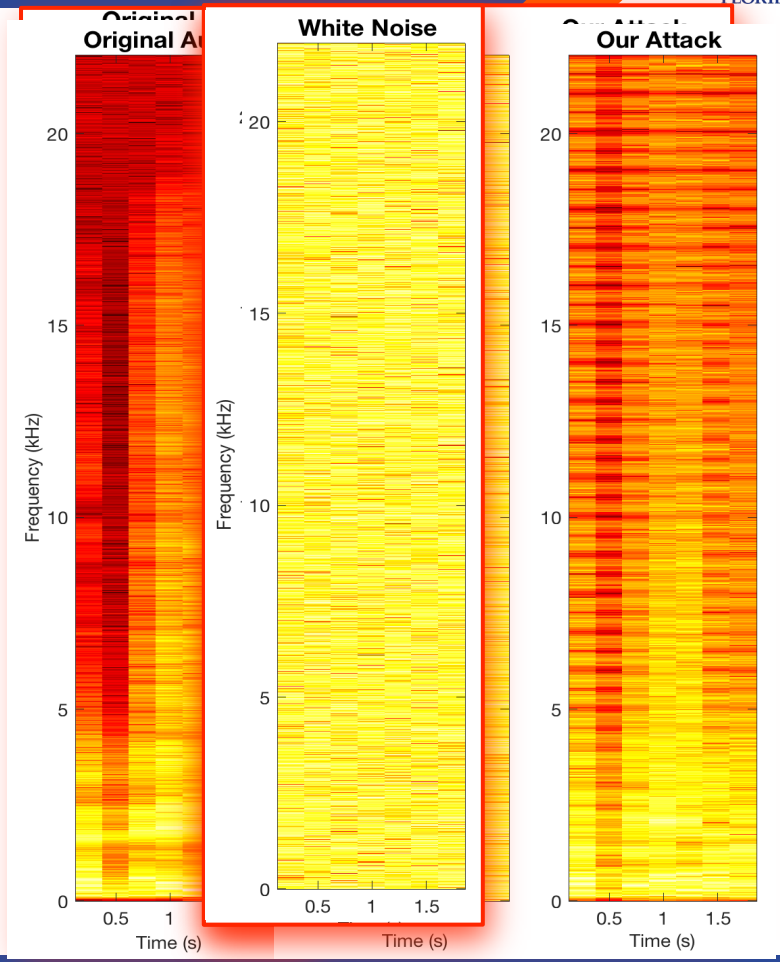
- Grouped into **4 types**





How the Attack Works (Psychoacoustics)

- Intelligibility **hard** to **measure**
- Fundamentals of **psychoacoustics**
- **Spread energy** across **spectrum**



How the Attack Works (Evaluation)

	Speech	Speaker
Task	**Noise -> text**	**Noise -> user**
Data	**> 20,000** successful attack samples	**22** speakers
Queries	**<10 queries to model (a few seconds!)**	
Models	KALDI, IBM, HOUNDIFY, intel, CMU Sphinx, bing, wit.ai	Microsoft Azure **x2**

Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway

Demo! (1/2)

More at:

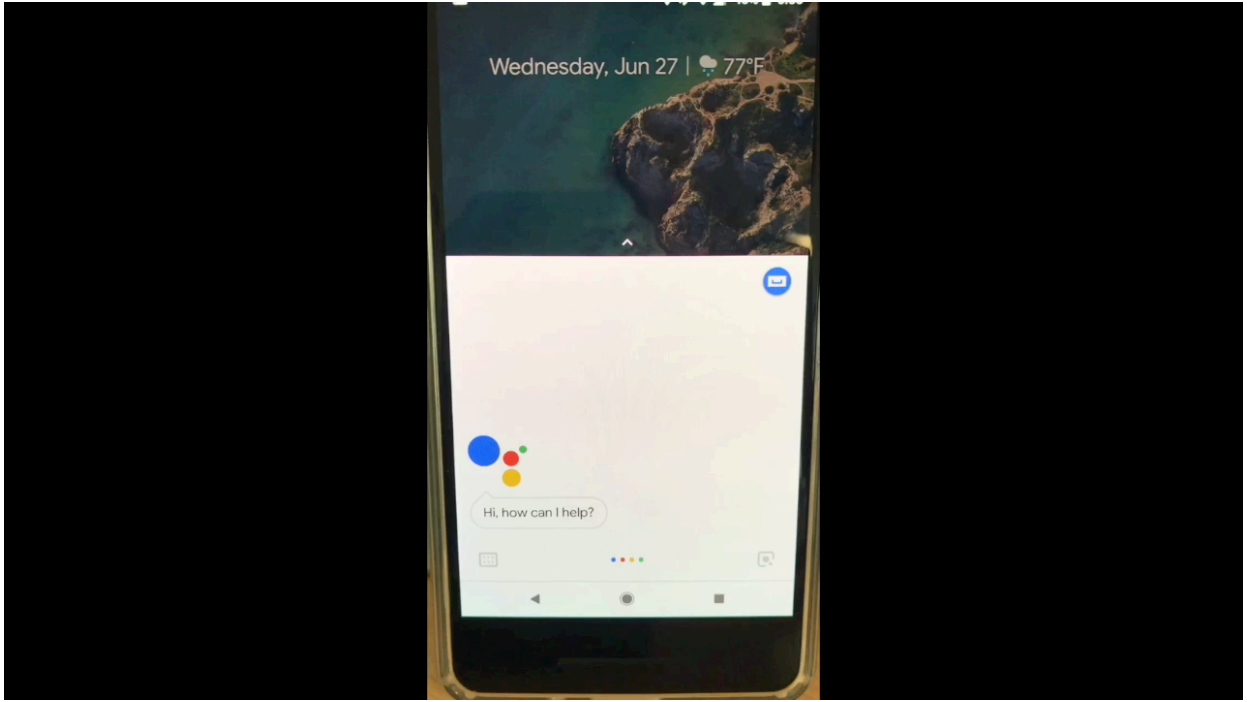
https://sites.google.com/view/practicalhiddenvoice/home

Demo! (2/2)

More at:

https://sites.google.com/view/practicalhiddenvoice/home

Demo! (2/2)



More at:

https://sites.google.com/view/practicalhiddenvoice/home

Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway

Takeaway

- Simple, efficient audio transformations yield “**noise**” that is **understood** as commands **by speech systems**
- The **model** is **irrelevant**
- **All systems** we tested are **vulnerable**
- **Achieve** the same **goals** as traditional **Adversarial ML**

Project webpage:
sites.google.com/view/practicalhiddenvoice/home

hadi10102@ufl.edu
hadiabdullah.github.io
hadiabdullah1

- Easy to get around
- Artificially generate a target's speech
  - LyreBird
- Capture and stitch together target's speech

Defenses

- Must be implemented at or before feature extraction
- Adversarial Training?
- Voice Activity Detection?
- Environmental Noise
- Liveness Detection
  - Blue et al.*

*

L. Blue, L. Vargas, and P. Traynor, "*Hello, is it me you're looking for? Differentiating between human and electronic speakers for voice interface security,*" in 11th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 2018.