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

Presenter: Chen Wang

Chen Wang[†], Abhishek Anand [*], Jian Liu[†], Payton Walker[*], Yingying Chen[†], Nitesh Saxena[*]

[†]*WINLAB, Rutgers University, NJ, USA*

** University of Alabama at Birmingham, AL, USA*

RUTGERS UNIVERSITY | NEW BRUNSWICK

ACSAC 2019

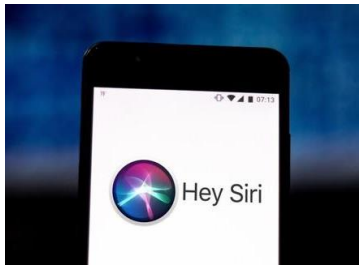
December 9 – 13, 2019

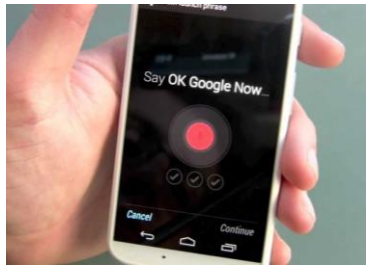
THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Introduction

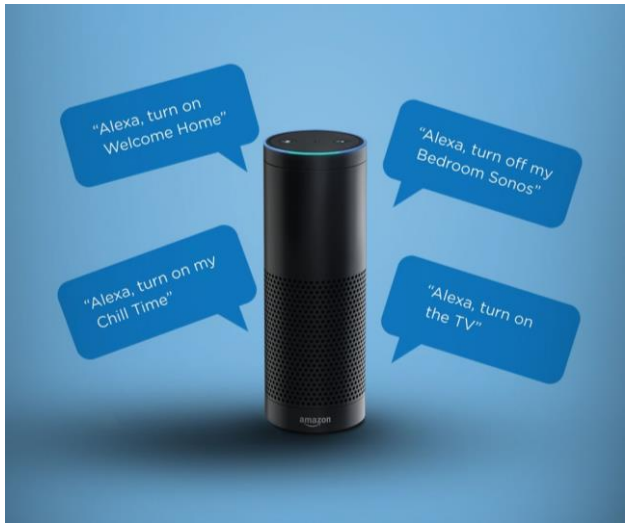
- Widely deployed voice controllable systems (VCS)
  - Convenient way of interaction
  - Integrated into many platforms

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





Smart appliances

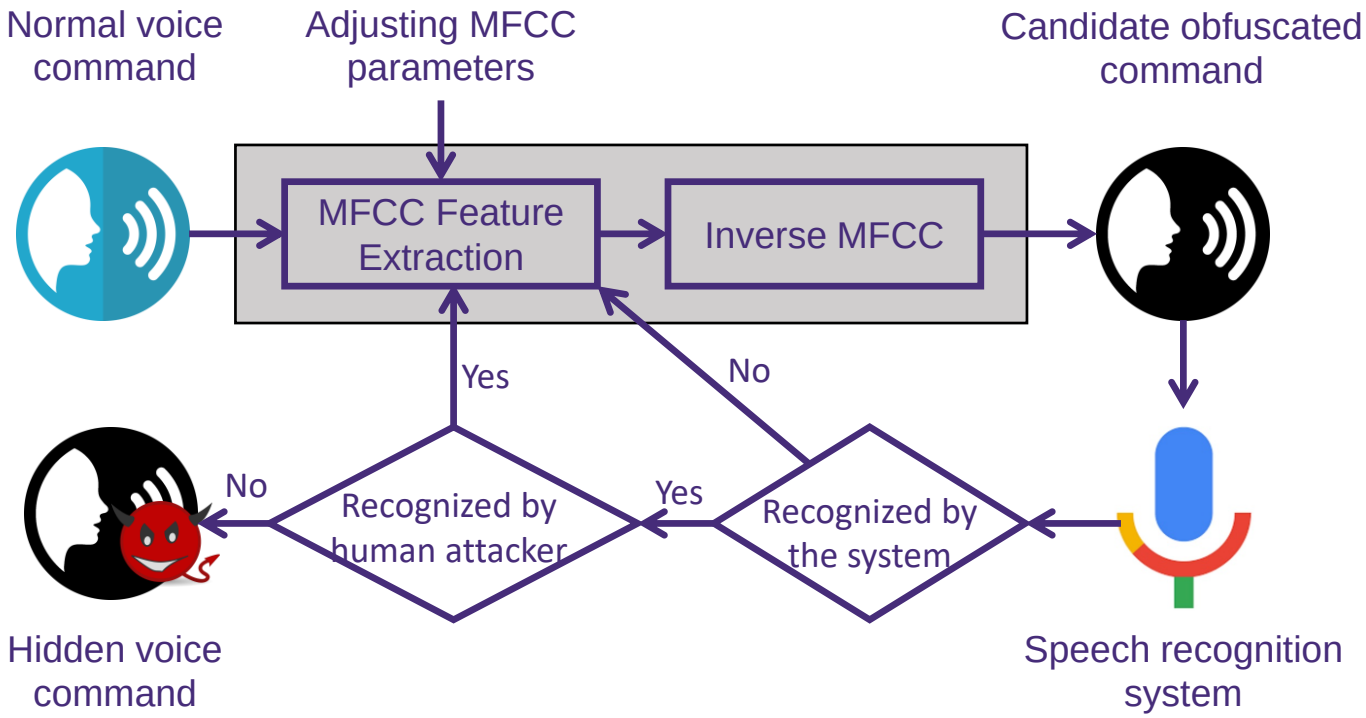


stand-alone assistants

- 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



Related Work

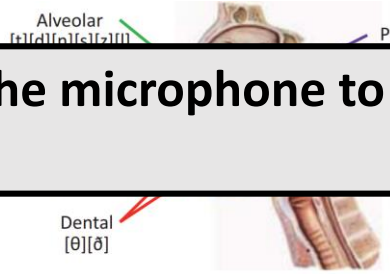
- Defend acoustic attacks based on audio informaiton
  - Voice authentication models

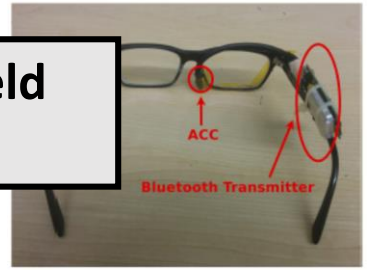
Only relying on speech audio features is vulnerable to hidden voice commands

  - 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**

(e.g., on a wearable)





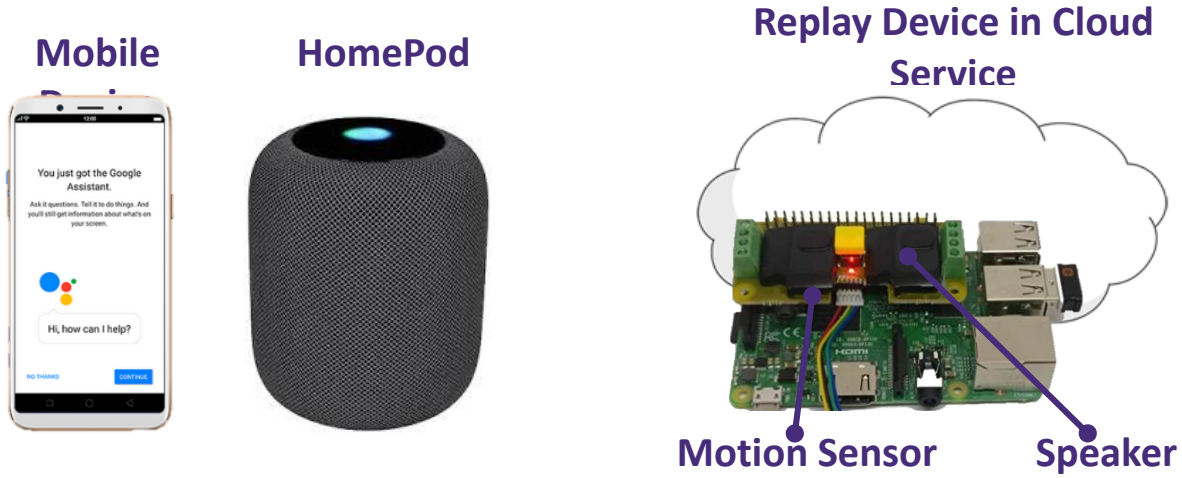
A multi-modality authentication framework is highly desirable to provide enhanced security:

Audio sending modality + vibration sensing modality

Basic Idea

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

- Many VCS devices (e.g., smartphones and voice assistant systems) are already equipped with motion sensors
- 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

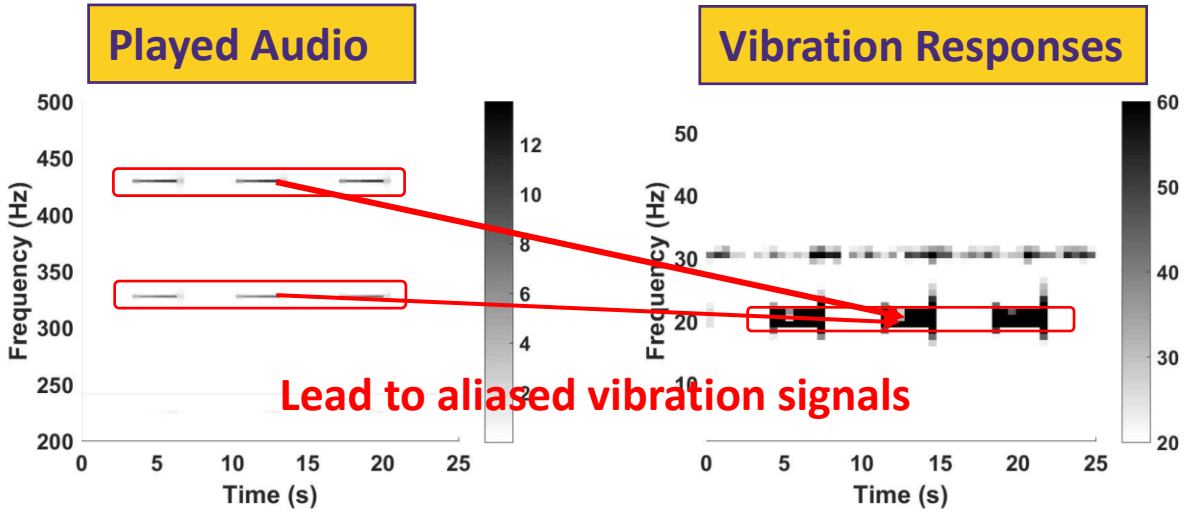


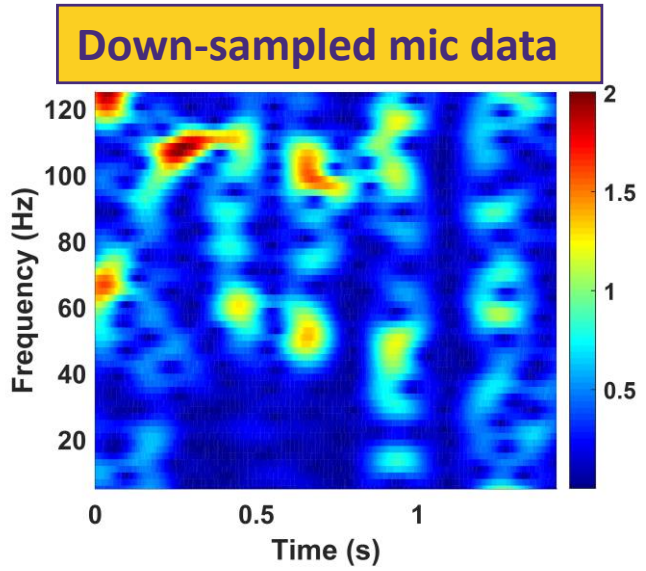
Front-end playback **Back-end playback**

Capturing Voice Using Motion Sensors

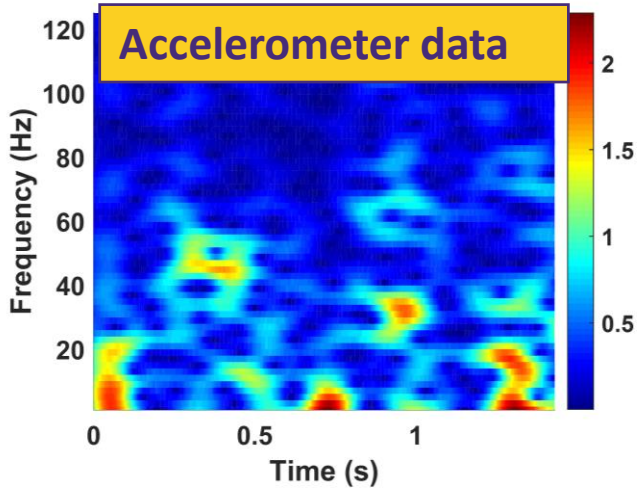
- Shared surface between loudspeaker and microphone
- Low sampling rate motion sensors (e.g., < 200Hz)
- Nonlinear vibration responses
- Distinct vibration domain

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





"show facebook.com"



Why Vibration?

- Existing speech/voice recognition methods are based audio domain voice vocal features
- Hidden voice commands are designed to duplicate these audio domain features by iteratively modify a voice command
- Audio-induced surface vibrations
  - An additional sensing domain, distinct to audio
  - Hard to be forged from audio signals in software
  - Similar audio features result in distinct vibration features
  - Resulting vibration responses are device-dependent (device physical vibrations, motion sensors)

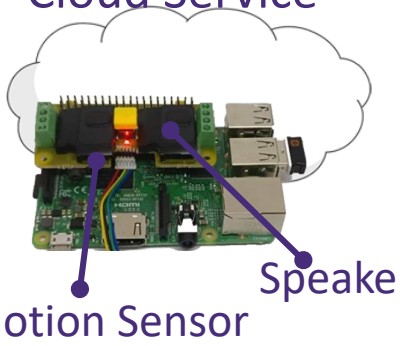
The vibration domain approach can work in conjunction with the audio domain approach to more effectively detect the hidden voice commands.

System Overview

Mobile Device or HomePod

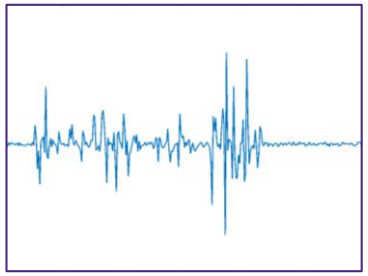
Frontend Playback

Replay Device in Cloud Service



Backend Playback

Accelerometer Readings



Data Calibration

- Vibration Noise Removal
- Voice Command Segmentation

Vibration Feature Derivation

- Time/Frequency Domain Statistical Features
- Acoustic Features (MFCC, Chroma Vector)

Vibration Feature Selection

- Feature Normalization
- Statistical Analysis based Selection

Hidden Voice Command Detection

Supervised Learning-based Classifier

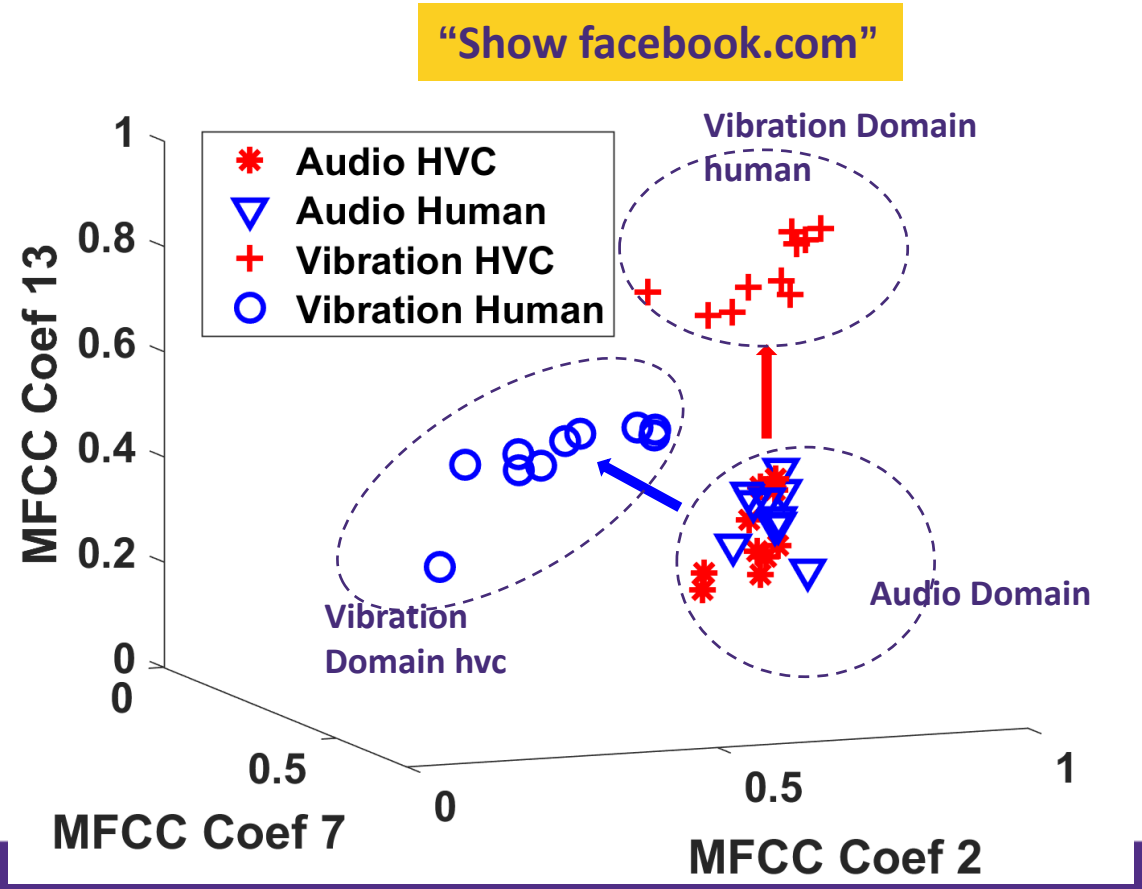
- Simple Logistic
- Random Tree
- Random Forest
- SMO

Unsupervised Learning-based Classifier

- K-means
- K-medoid

Vibration Feature Derivation

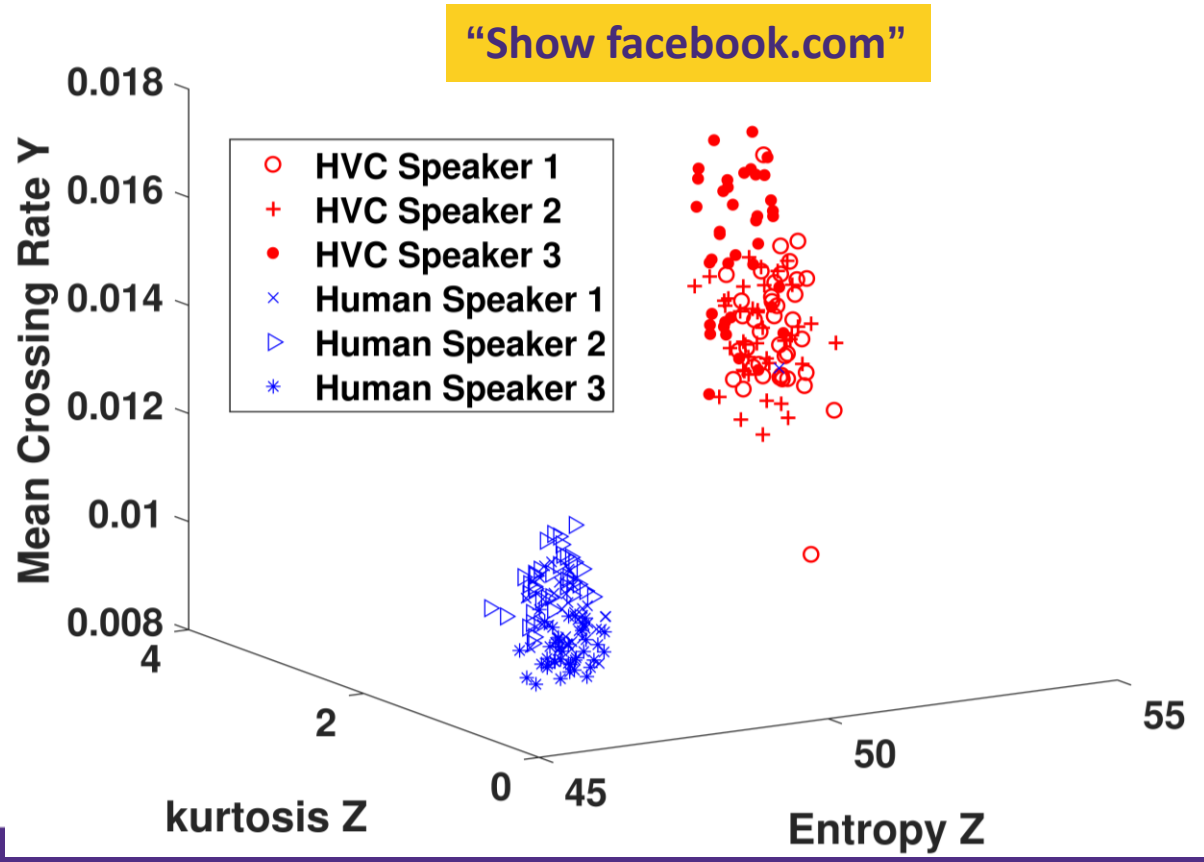
- Unique and hard to forge vibration features
  - Statistical features in time and frequency domains
  - Deriving Acoustic Features from Motion Sensor Data
    - MFCC
    - Chrome vectors
- Nonlinear relationship between audio features and vibration features



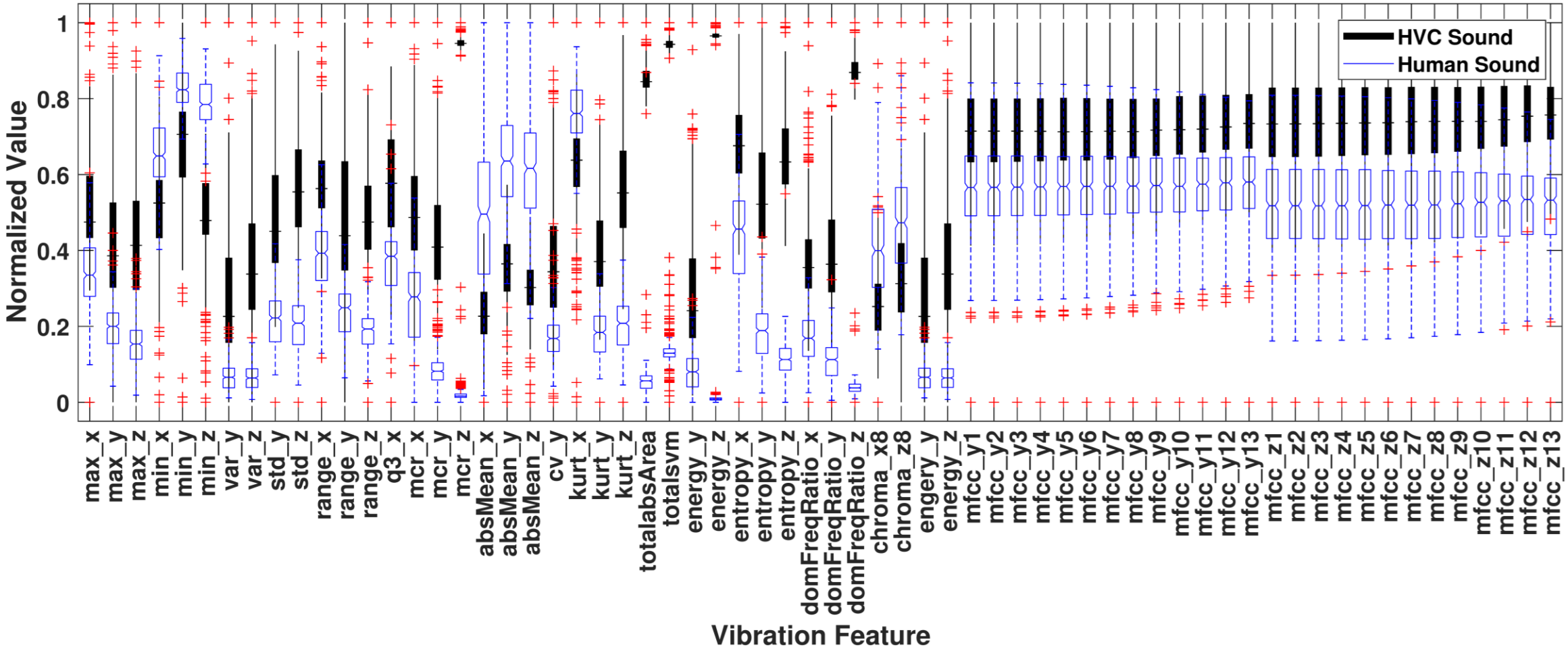
Vibration Feature Derivation

- 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

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

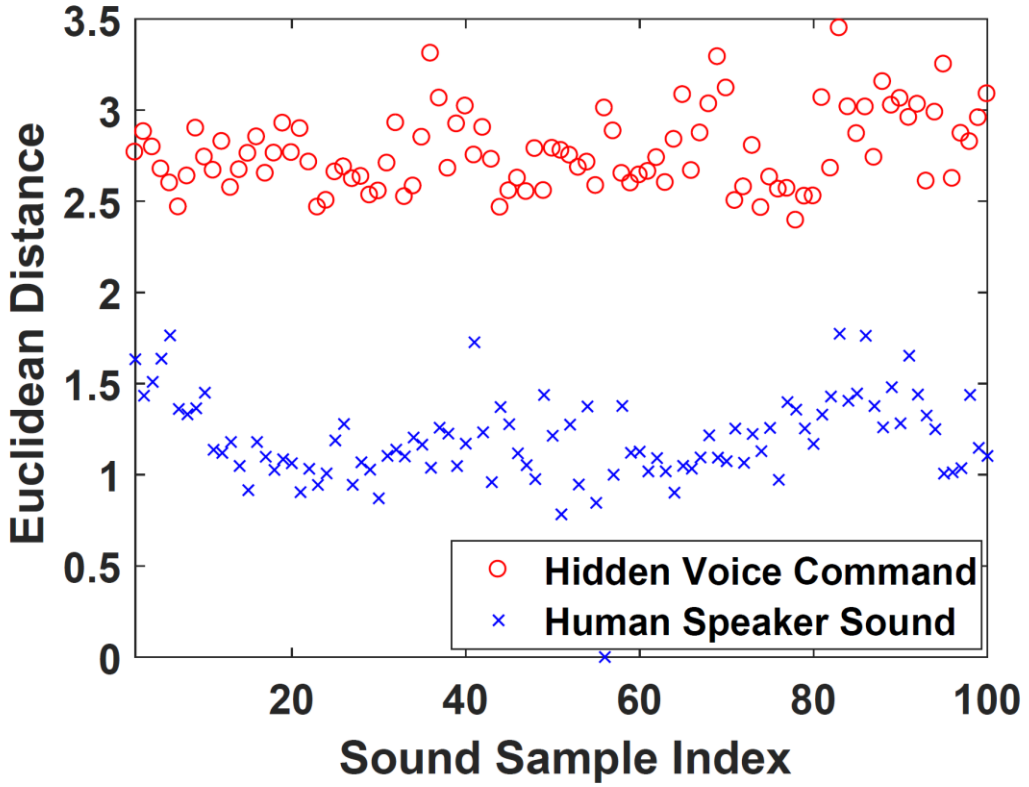


Feature Selection Based on Statistical Analysis



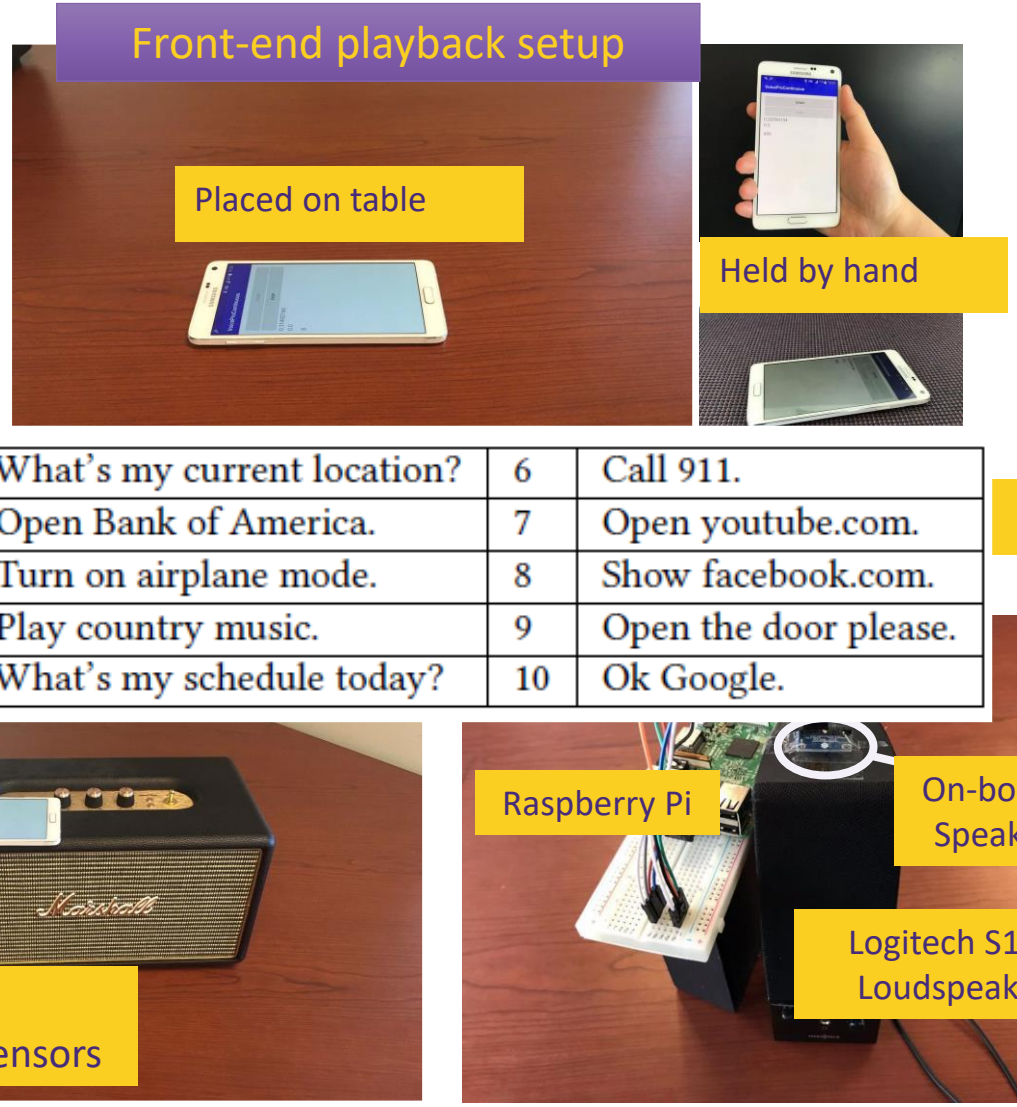
Hidden Voice Command Detection

- 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
Supervised-learning

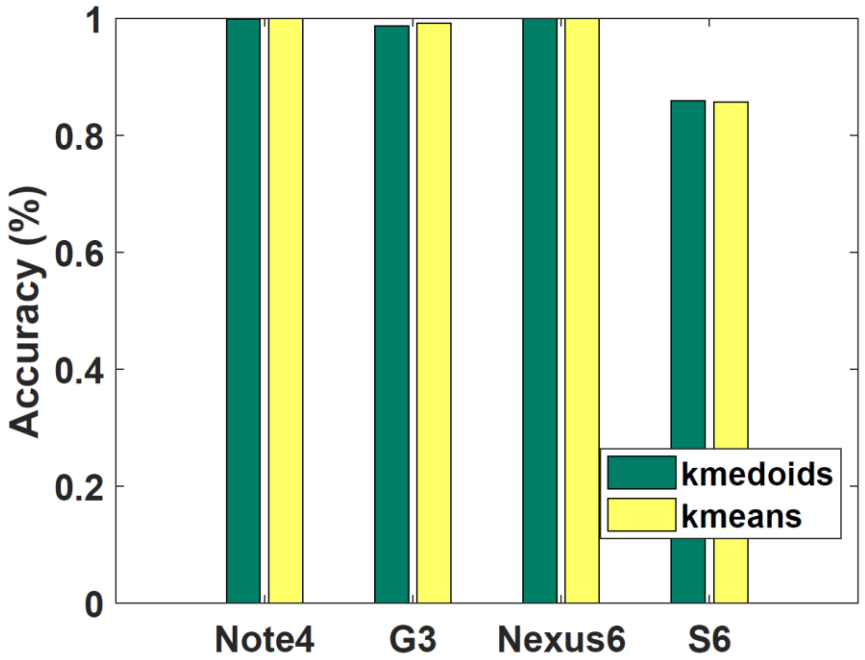
Front-end playback setup

	Note 4	G3	Nexus 6	S6
SimpleLogistic	100%	99.8%	100%	88.3%
SMO	100%	99.9%	99.9%	85.4%
Random Forest	100%	99.5%	100%	93.1%
Random Tree	99.9%	98.1%	100%	87.4%

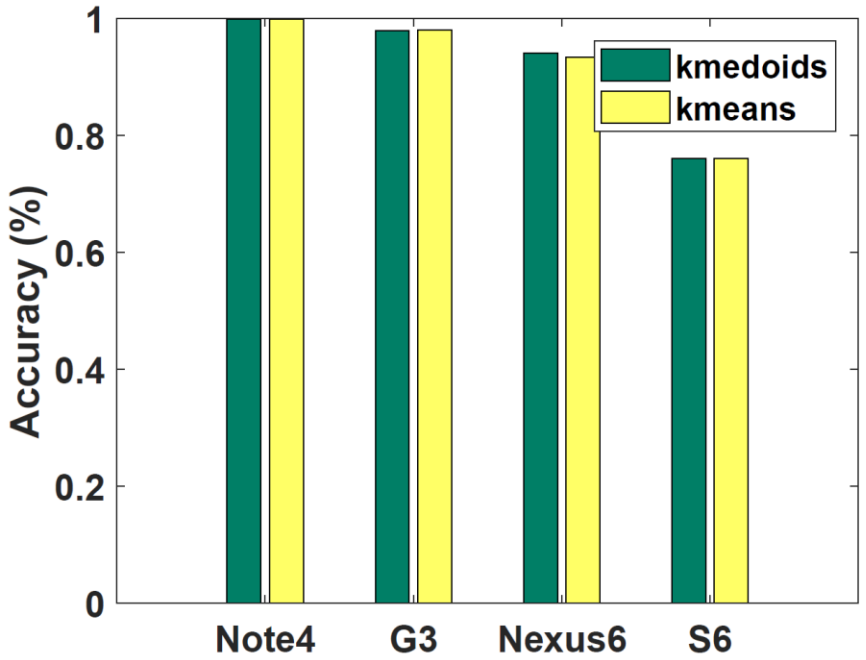
Back-end playback setup

	Note 4	G3	Nexus 6	S6
SimpleLogistic	99.9%	99.8%	99.8%	95.3%
SMO	99.9%	99.9%	99.3%	95.0%
Random Forest	99.9%	100%	99.8%	95.3%
Random Tree	99.9%	98.9%	97.9%	89.7%

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%

Conclusion

- Show 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 (i.e., frontend playback and backend playback)
- Extensive experiments show that the hidden voice commands can be detected based on their speech features in the vibration domain with high accuracy

Thank you