Hidden Voice Commands


* University of California, Berkeley
** Georgetown University
Voice channel opens up new possibilities for attack
Today:

"Okay google, text [premium SMS number]"
In the future?

"Okay google, pay John $100"
We make voice commands *stealthy*. 
We produce audio which is **noise** to humans, but **speech** to devices.
This is an instance of attacks on Machine Learning
Background
Background

Feature Extraction ➔ ML Algorithm ➔ Text
Feature Extraction
Feature Extraction
Feature Extraction

MFCC

\[ [X_0] \]

MFCC

\[ [X_1] \]

MFCC

\[ [X_2] \]
Feature Extraction → ML Algorithm → Text
First Attack: White-Box

Assume complete system knowledge (model, parameters, etc)
Recognition

Feature Extraction → ML Algorithm → Text
Attack

Feature Extraction

ML Algorithm

Text
Attack

Feature Extraction → ML Algorithm → Text
Attack

Feature Extraction → ML Algorithm → Text
Inverting Feature Extraction

MFCC\(^{-1}\) \[X_0\]

MFCC\(^{-1}\) \[X_1\]

MFCC\(^{-1}\) \[X_2\]
Inverting Feature Extraction

MFCC$^{-1}$

$[X_0]$

MFCC$^{-1}$

$[X_1]$

MFCC$^{-1}$

$[X_2]$
Inverting Feature Extraction

$[X_0]$
Inverting Feature Extraction

MFCC$^{-1}$

$[x_0]$

MFCC$^{-1}$

$[x_1]$
Inverting Feature Extraction

\[ \text{MFCC}^{-1} \]

\[ [x_0] \]

\[ [x_1] \]
Inverting Feature Extraction

MFCC^{-1} \rightarrow [X_0] \rightarrow \text{MFCC^{-1}} \rightarrow [X_1] \rightarrow \text{MFCC^{-1}} \rightarrow [X_2]
Inverting Feature Extraction

MFCC

$[x_0]$  $[x_1]$  $[x_2]$
Actually not that easy
Playing attacks over-the-air

1. Create a model of the physical channel
2. Use model to predict effect of over-the-air
3. Validate model by playing potential obfuscated commands during generation
Demo
Okay Google, take a picture
Demo
Okay Google, text 12345
Demo
Okay Google, browse to evil.com
Not Over-The-Air Demo
Okay Google, browse to evil.com
Limitations

No background noise, in an echo-free room.

Assumes complete knowledge of model.
Can we make this attack practical?

Can we remove the white-box assumption?
Yes.

... but at the expense of attack quality.
Black-Box Attack

Audio Obfuscater → Speech Recognition → Text
Evaluation
Demo
White-Box

- Attack on open system
- Commands heavily obfuscated
- Works when played over-the-air
- Doesn't tolerate background noise

Black-Box

- Practical real-world attack
- Somewhat possible to recognize
- Works when played over-the-air
- Background noise and echo okay
Defenses?

- Notify the user that an action was taken.
- Challenge the user to perform an action.
- Detect and prevent the malicious commands.
Detect and Prevent

Successfully trained simple machine learning classifier: learn the difference between attack commands and actual commands
Conclusion

Voice: new paradigm for human-device interaction. This brings many new risks.

Something here on our hidden attacks.

The impact of these attacks will increase.

Future work is needed to construct defenses.

http://hiddenvoicecommands.com/