Hidden Voice Commands

Nicholas Carlini*, Pratyush Mishra*, Tavish Vaidya**,
Yuankai Zhang**, Micah Sherr**, Clay Shields**,
David Wagner*, Wenchao Zhou**

Berkeley
UNIVERSITY OF CALIFORNIA

* University of California, Berkeley
** Georgetown University


SAMSUNG
12:45
52°


10:08
Lust auf Lunch?
htc


amazon


5:30
SFO - JFK
Departs 8:05pm


NOKIA
htc
10:08
NOKIA
htc


MEINE BIBLIOTHEK
AUS DEM LEBEN EINES TAUGENICHTS
FAUST
SAMSUNG
SAMSUNG

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

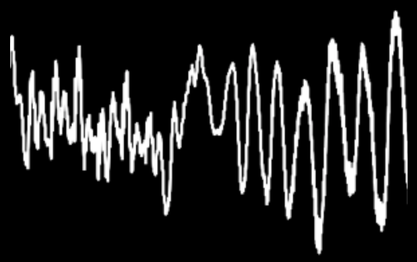
noise to humans, but

speech to devices.

This is an instance of attacks
on Machine Learning

Background

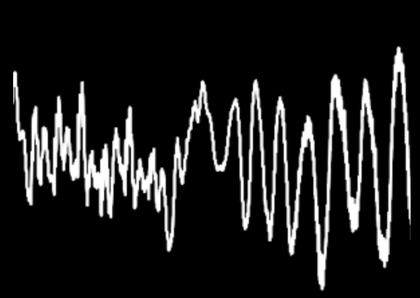
Background



Machine Learning Algorithm

Text

Background

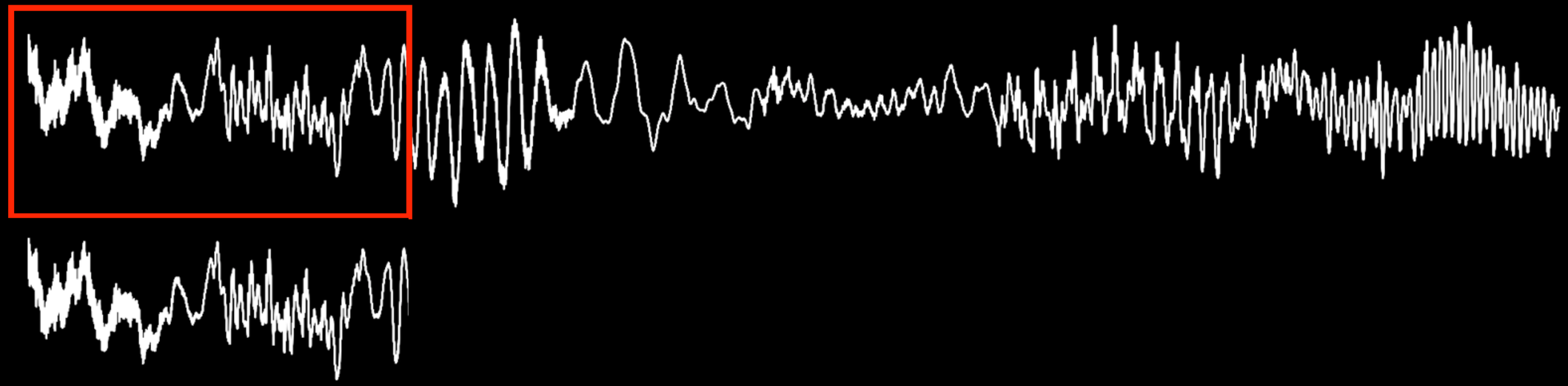


Feature Extraction

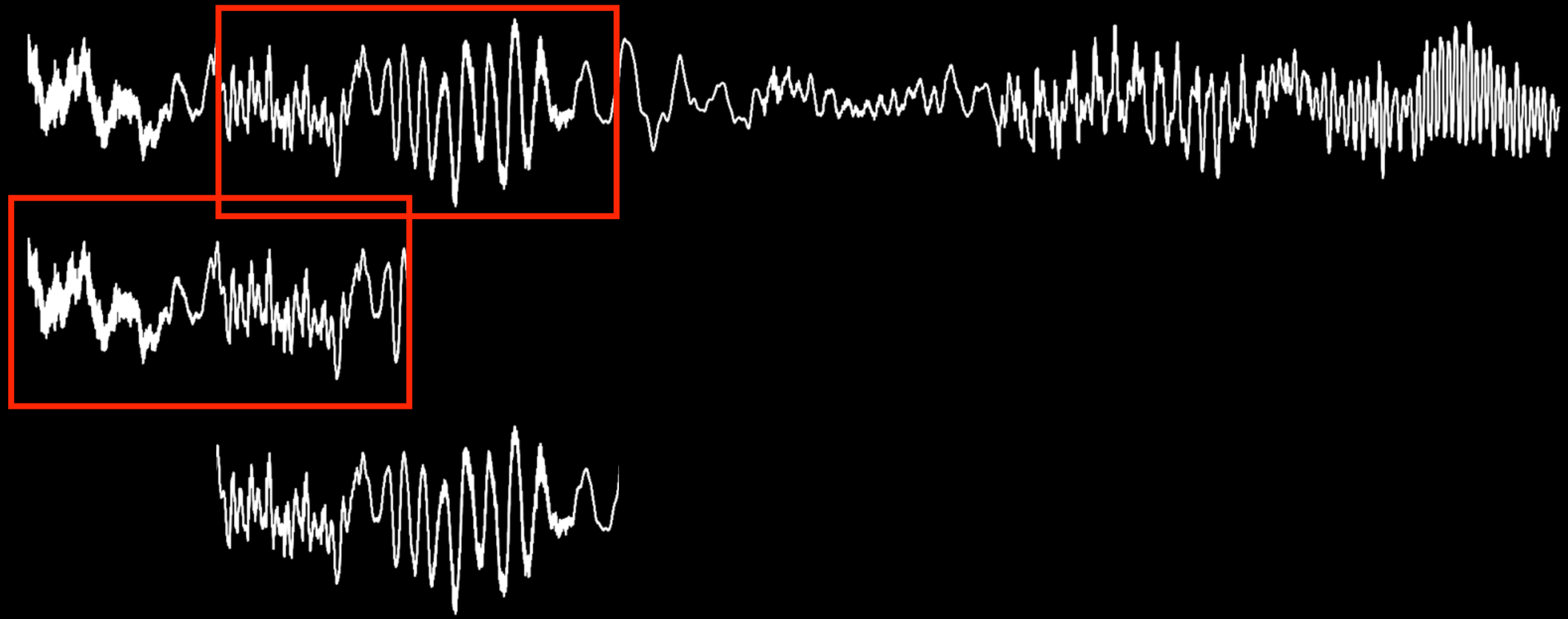
ML Algorithm

Text

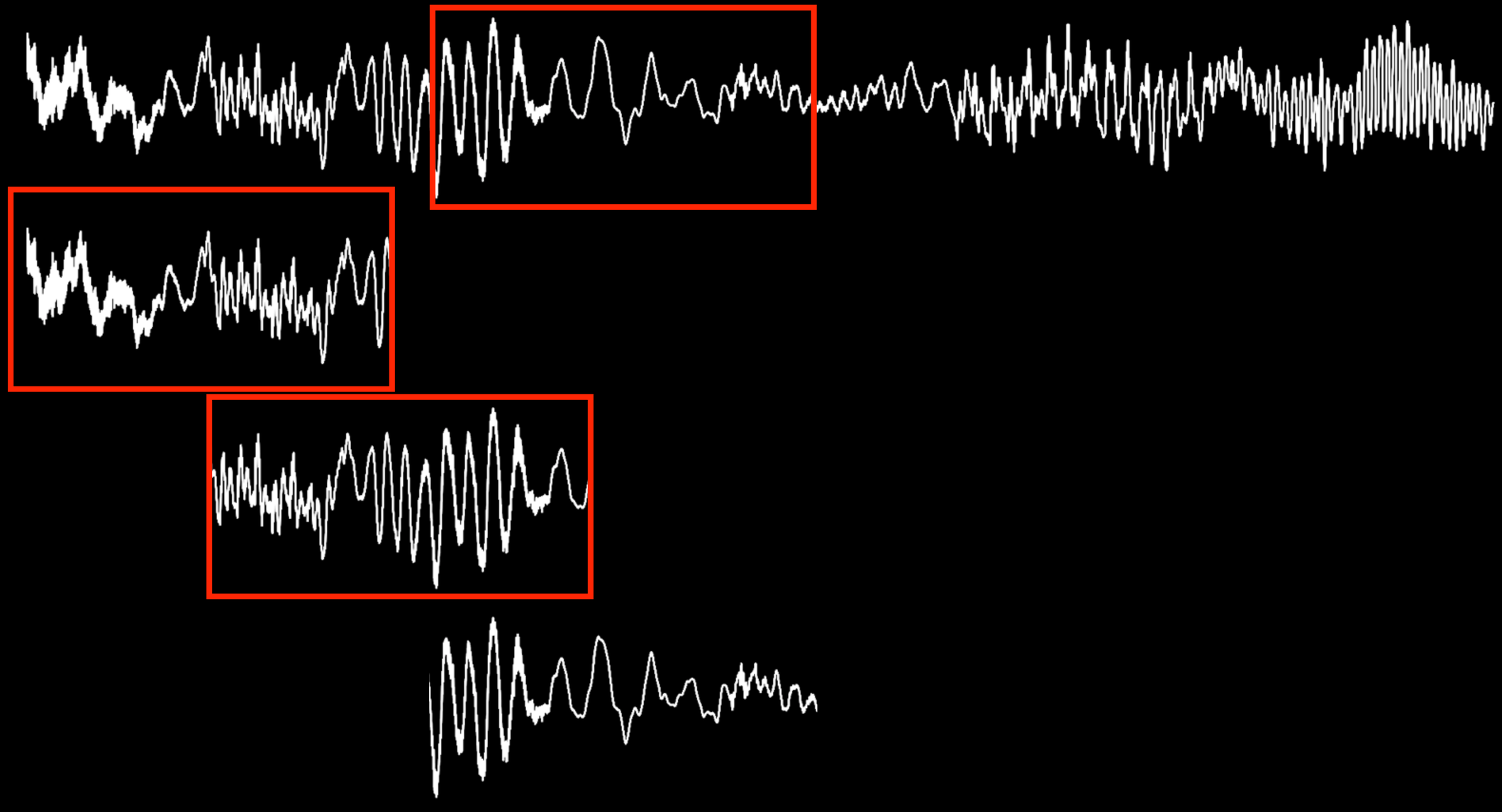
Feature Extraction



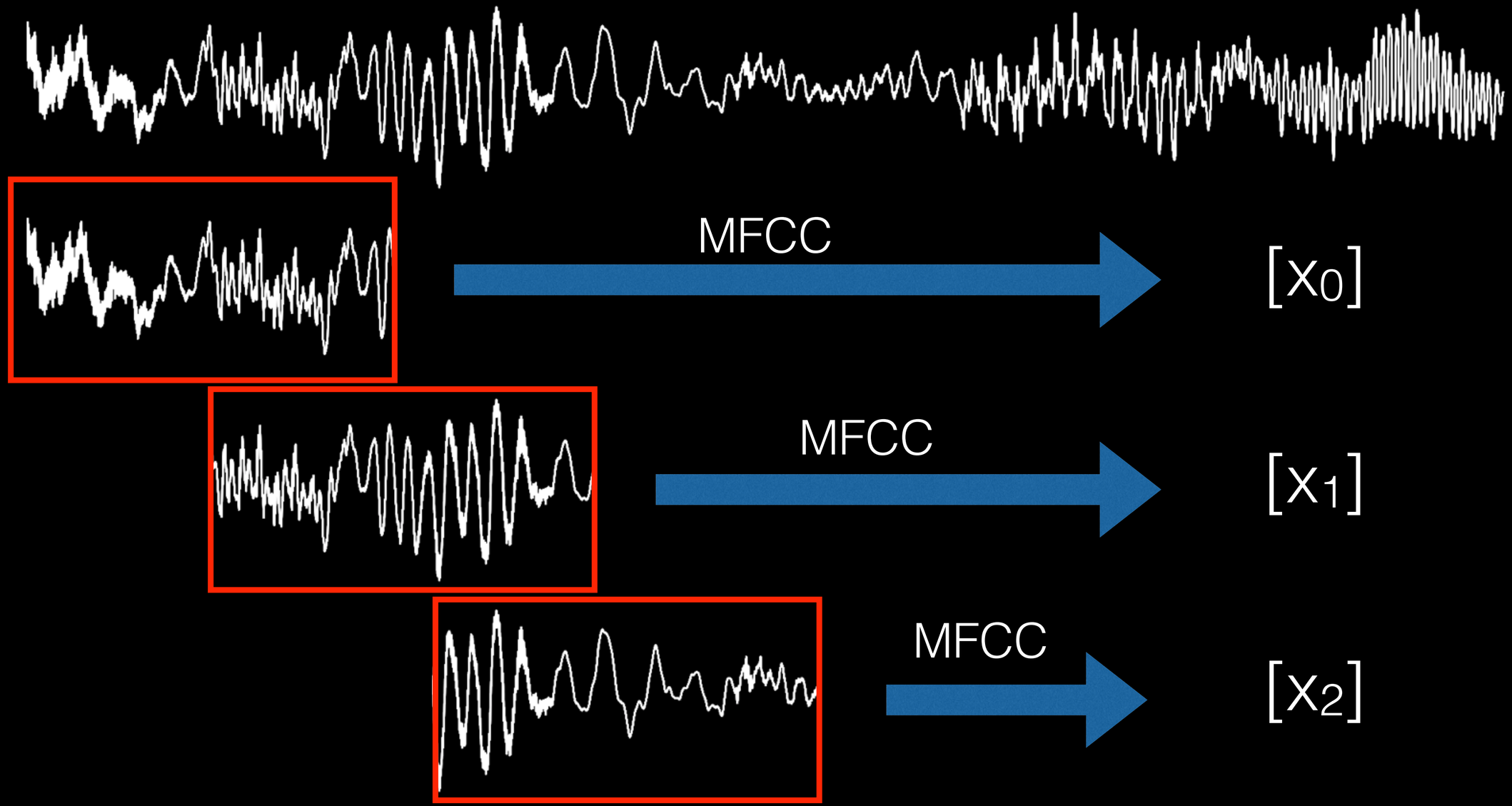
Feature Extraction

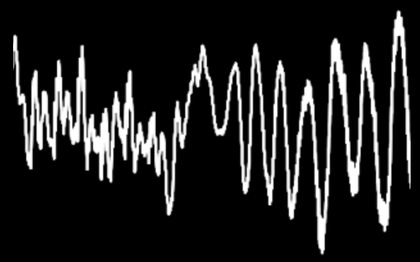


Feature Extraction



Feature Extraction





Feature Extraction

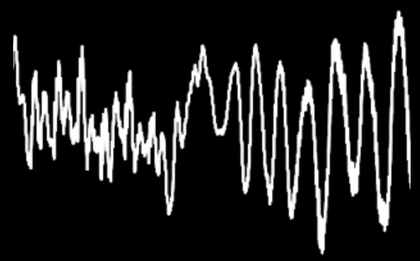
ML Algorithm

Text

First Attack:
White-Box

Assume complete system knowledge
(model, parameters, etc)

Recognition

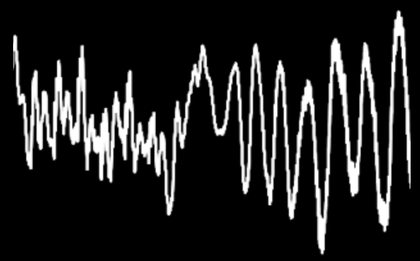


Feature Extraction

ML Algorithm

Text

Attack

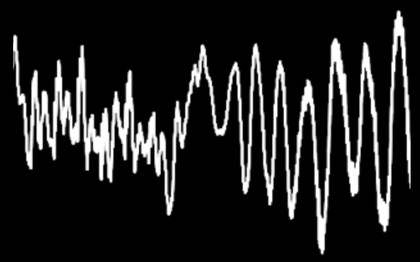


Feature Extraction

ML Algorithm

Text

Attack

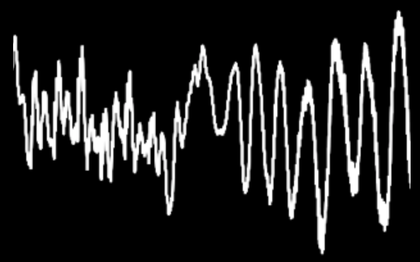


Feature Extraction

ML Algorithm

Text

Attack

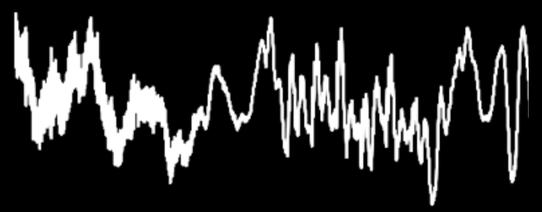


Feature Extraction

ML Algorithm

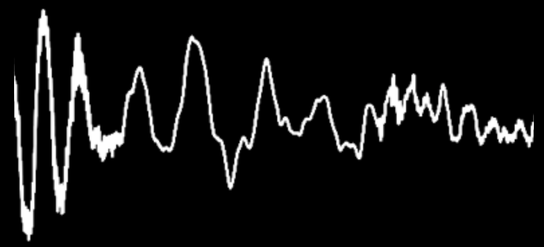
Text

Inverting Feature Extraction



$MFCC^{-1}$

$[x_0]$



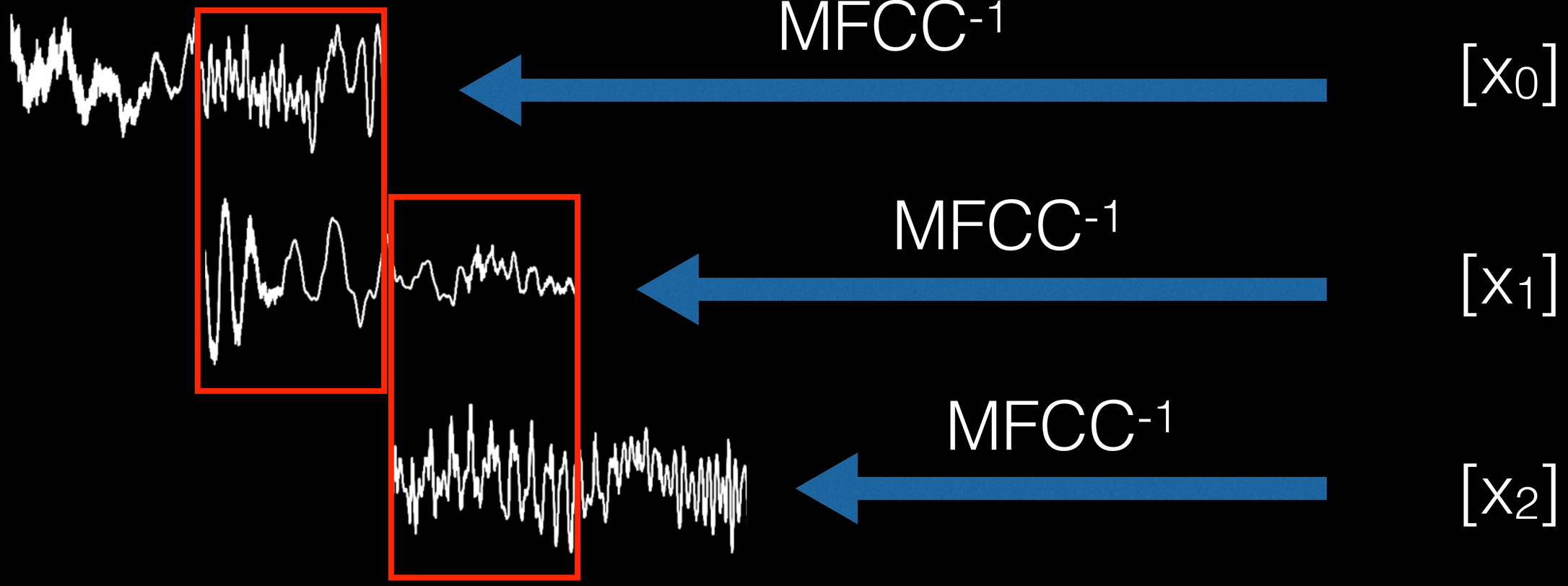
$MFCC^{-1}$

$[x_1]$

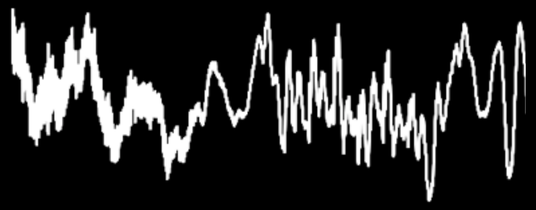
$MFCC^{-1}$

$[x_2]$

Inverting Feature Extraction



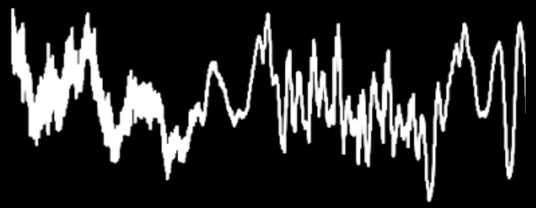
Inverting Feature Extraction



$MFCC^{-1}$

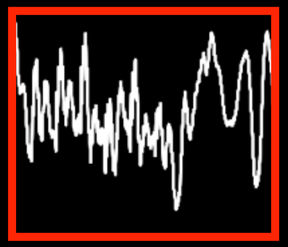
$[x_0]$

Inverting Feature Extraction



$MFCC^{-1}$

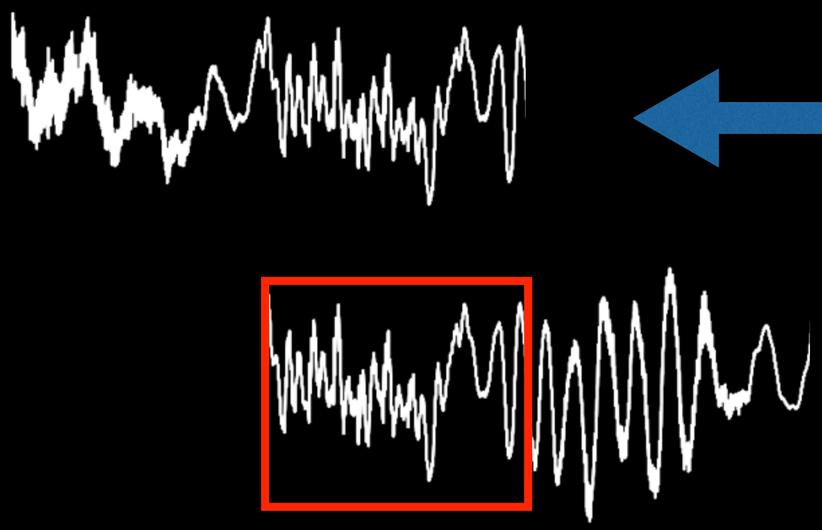
$[x_0]$



$MFCC^{-1}$

$[x_1]$

Inverting Feature Extraction



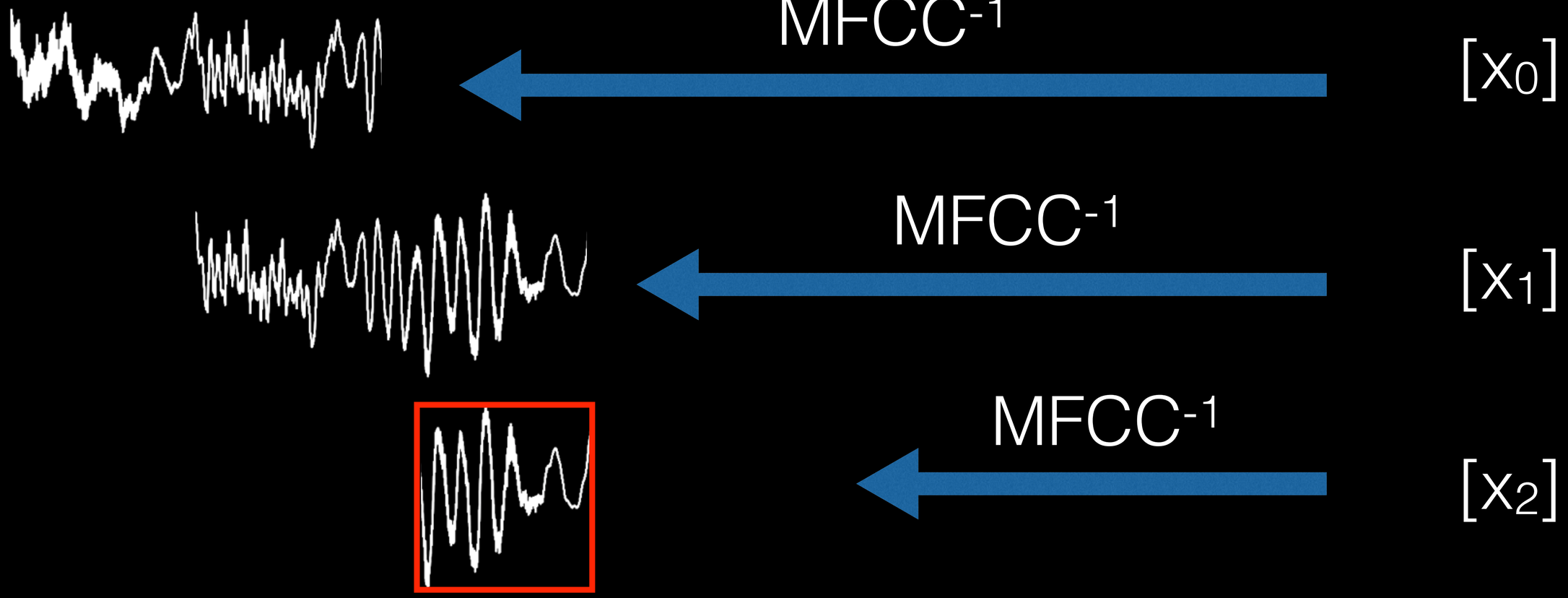
$MFCC^{-1}$

$[x_0]$

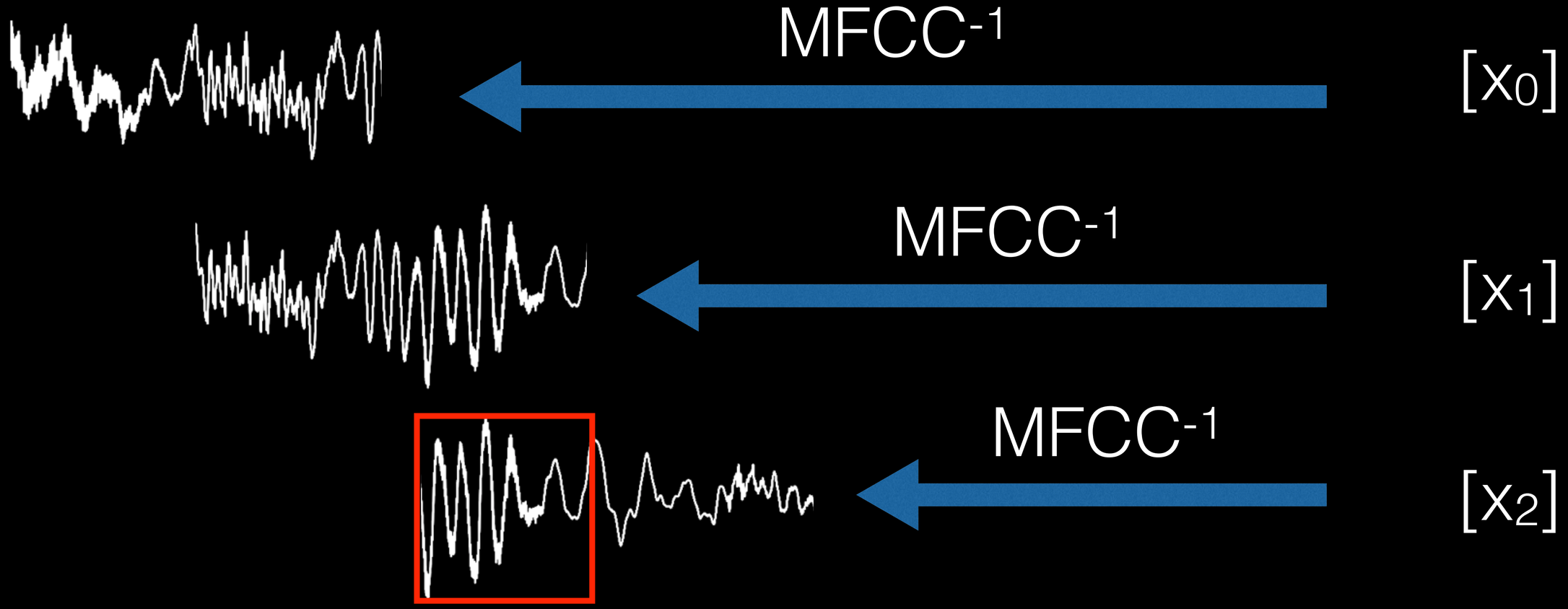
$MFCC^{-1}$

$[x_1]$

Inverting Feature Extraction



Inverting Feature Extraction



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

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.

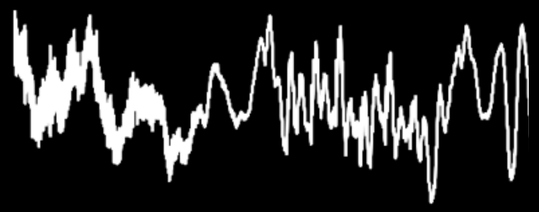
Can we make this attack practical?

Can we remove the white-box assumption?

Yes.

... but at the expense of attack quality.

Black-Box Attack

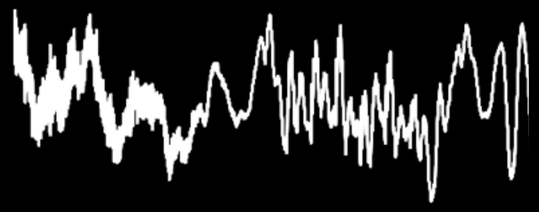


Audio Obfuscater

Speech Recognition

Text

Black-Box Attack



MFCC

$MFCC^{-1}$

Speech Recognition

Text

Evaluation

Demo

SAMSUNG

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.
[combine]This brings many new risks.

Something here on on our hidden attacks.

The impact of these attacks will increase.

Future work is needed to construct defenses.

http://hiddenvoicecommands.com/