Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition

Tavish Vaidya, Yuankai Zhang, Micah Sherr and Clay Shields

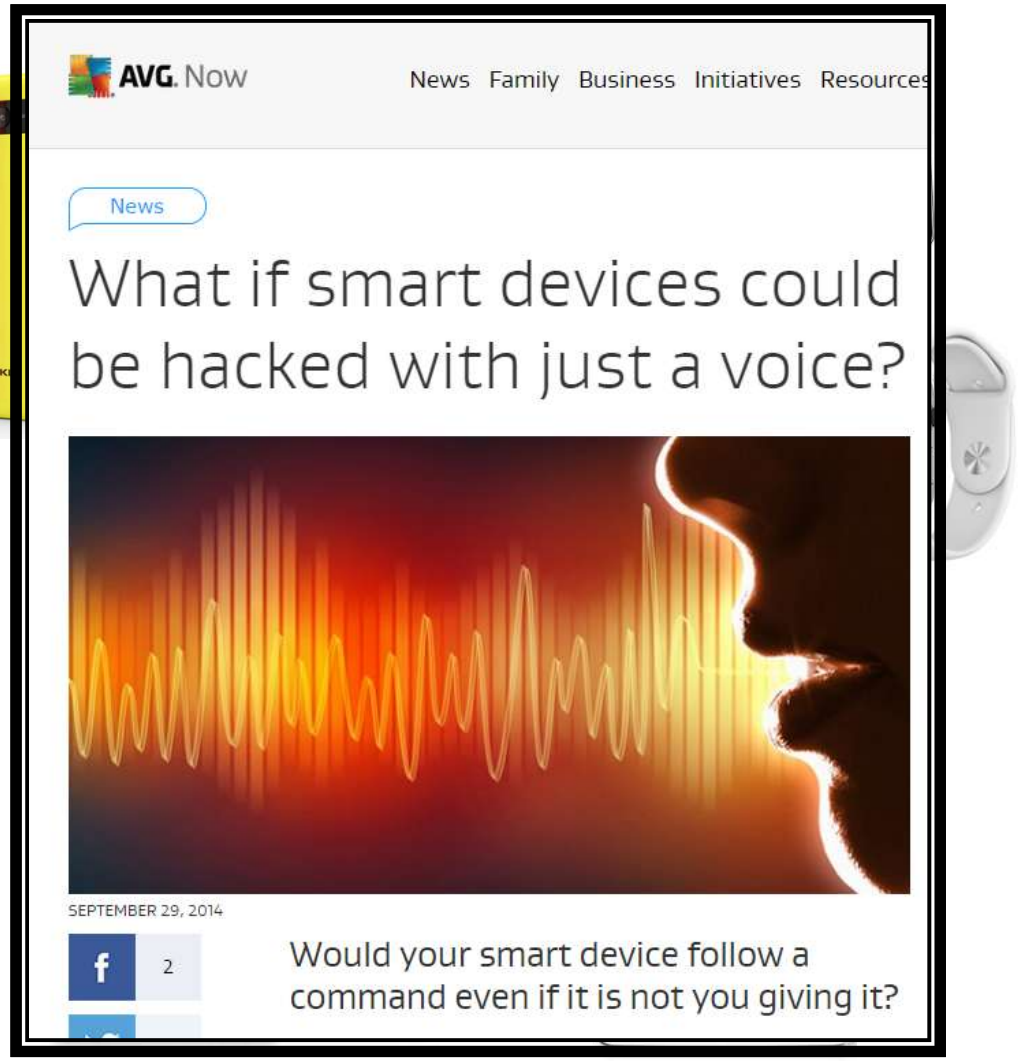
Georgetown University

Presented at **WOOT'15**

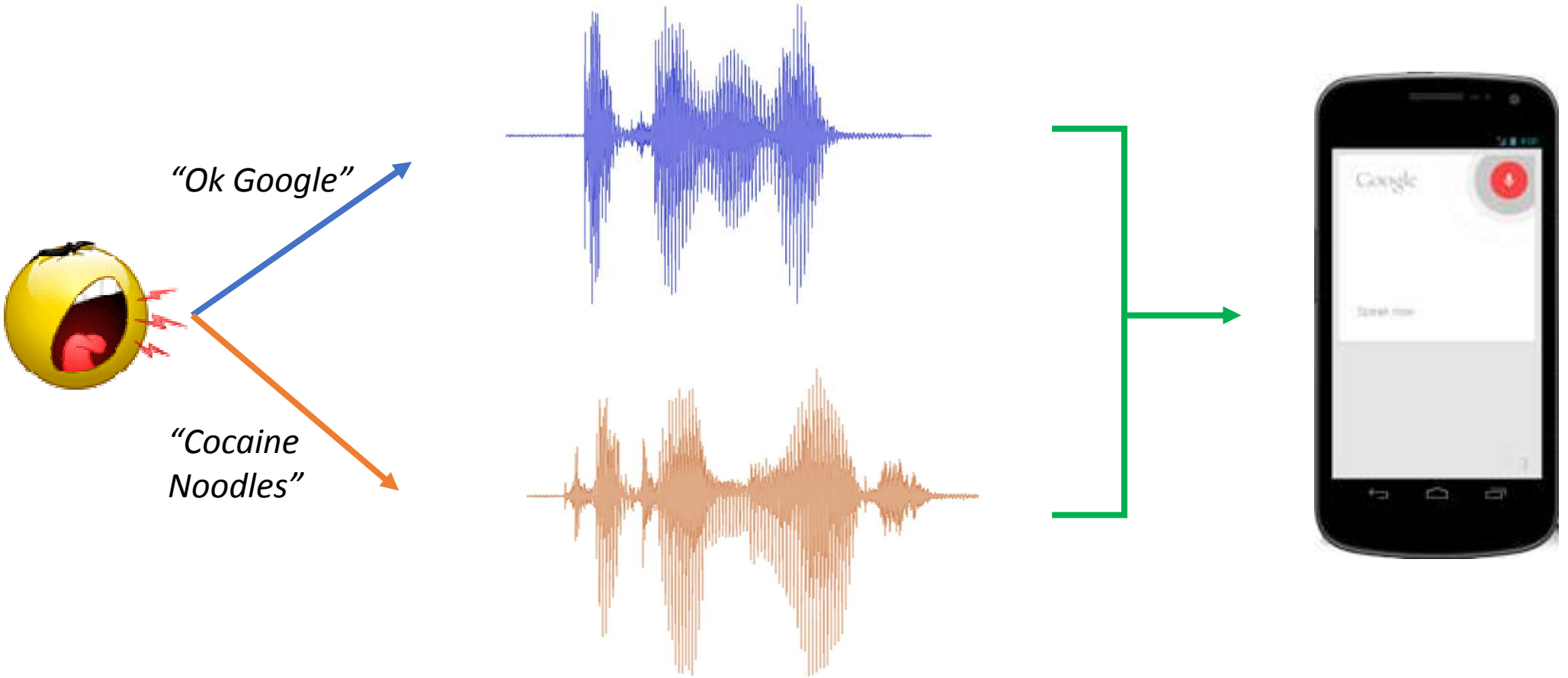
10-11 August, 2015

Voice input is near ubiquitous

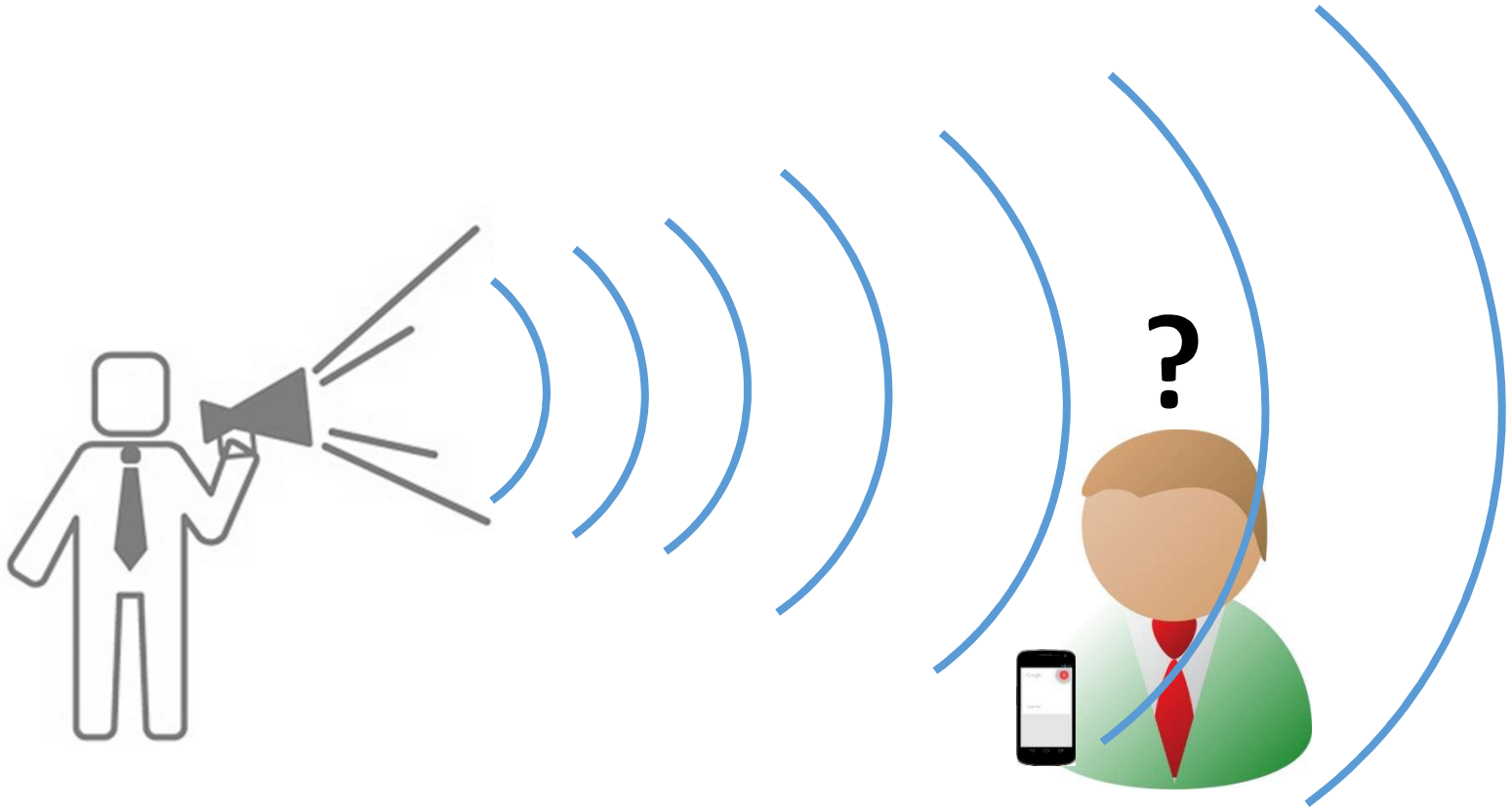
Voice input is near ubiquitous



Can the differences between human and machine understanding of speech lead to attacks?



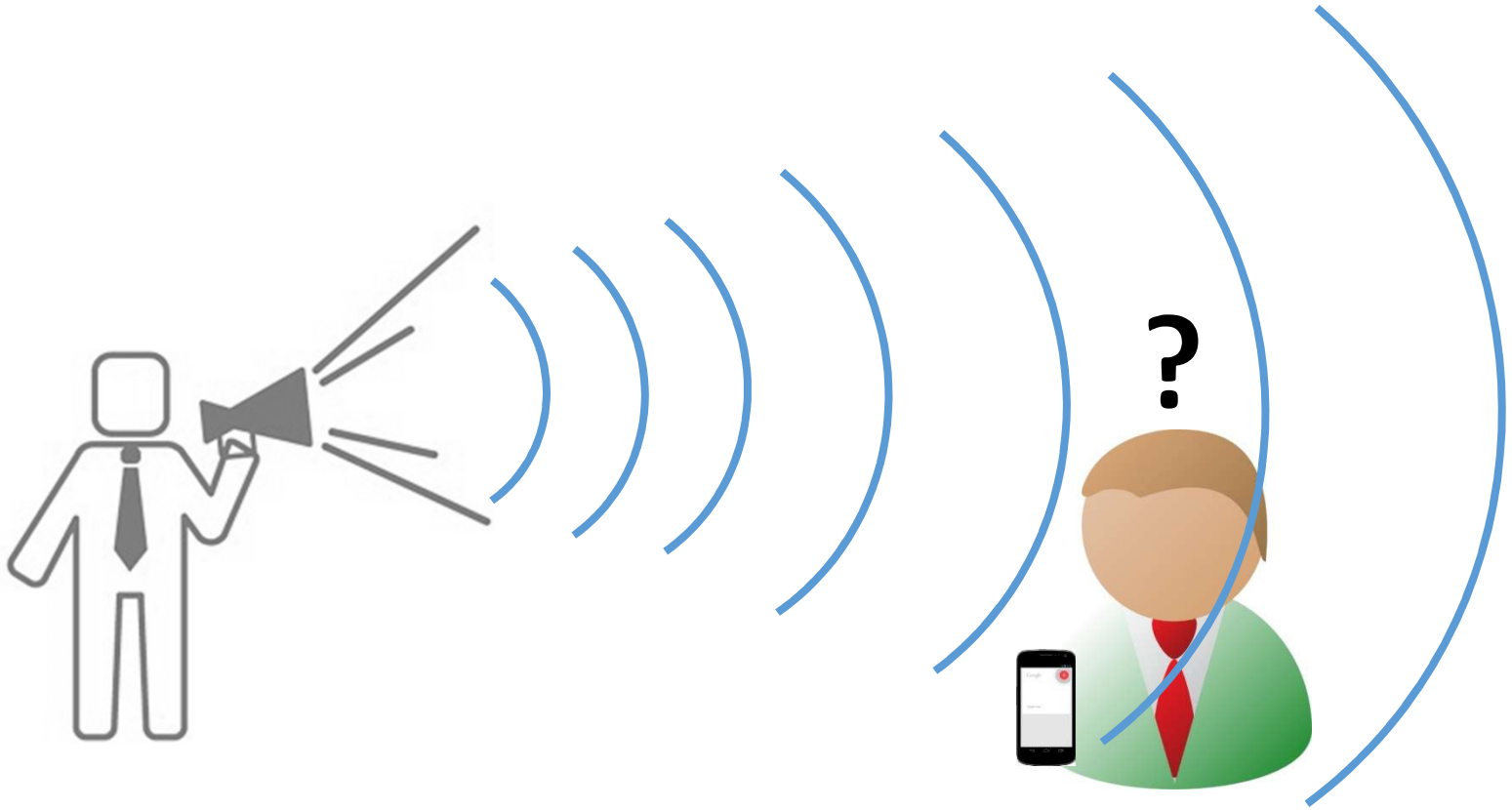
Attack 1



Open malicious webpage

- Serve drive-by-download or malware
- Open up attack surface for further attacks

Attack 2



Send text message to particular number

- Monetize the attack using reverse SMS billing or premium SMS service numbers as destination

Attack 3

Enumerate devices in an area (e.g. those belonging to dissidents attending a rally)

Other Attacks

- Denial-of-Service
  - E.g., use public announcement systems to turn on airplane mode
- Sending/forging email
- Sending/forging messages on social media

Attacker Goals

- Execute commands on target device by exploiting its speech recognition system

- Minimize the possibility of alerting the user of the attack
  - Produce *mangled* commands that are understood by the device but not the user

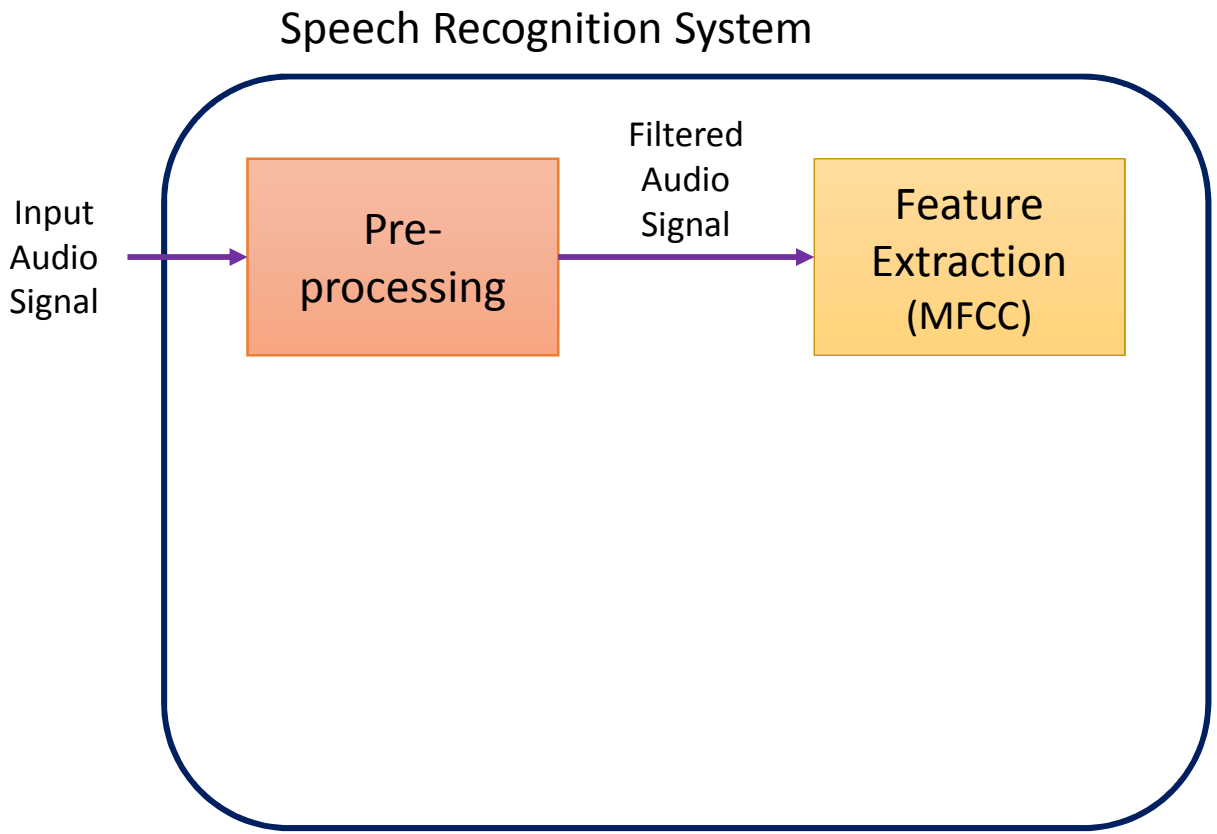
(Non)Assumptions

- Non-assumption: we make no assumption about target speech recognition system
  - Speech recognition model and process are treated as black boxes
  - Attacks are agnostic to particular AI/ML used by target device

- Adversary is able to play audio to target devices
  - E.g., from an elevator speaker, youtube video, LRAD etc.
- Target devices do not apply biometrics or attempt to authenticate users/speakers
- Target devices are always listening to voice input

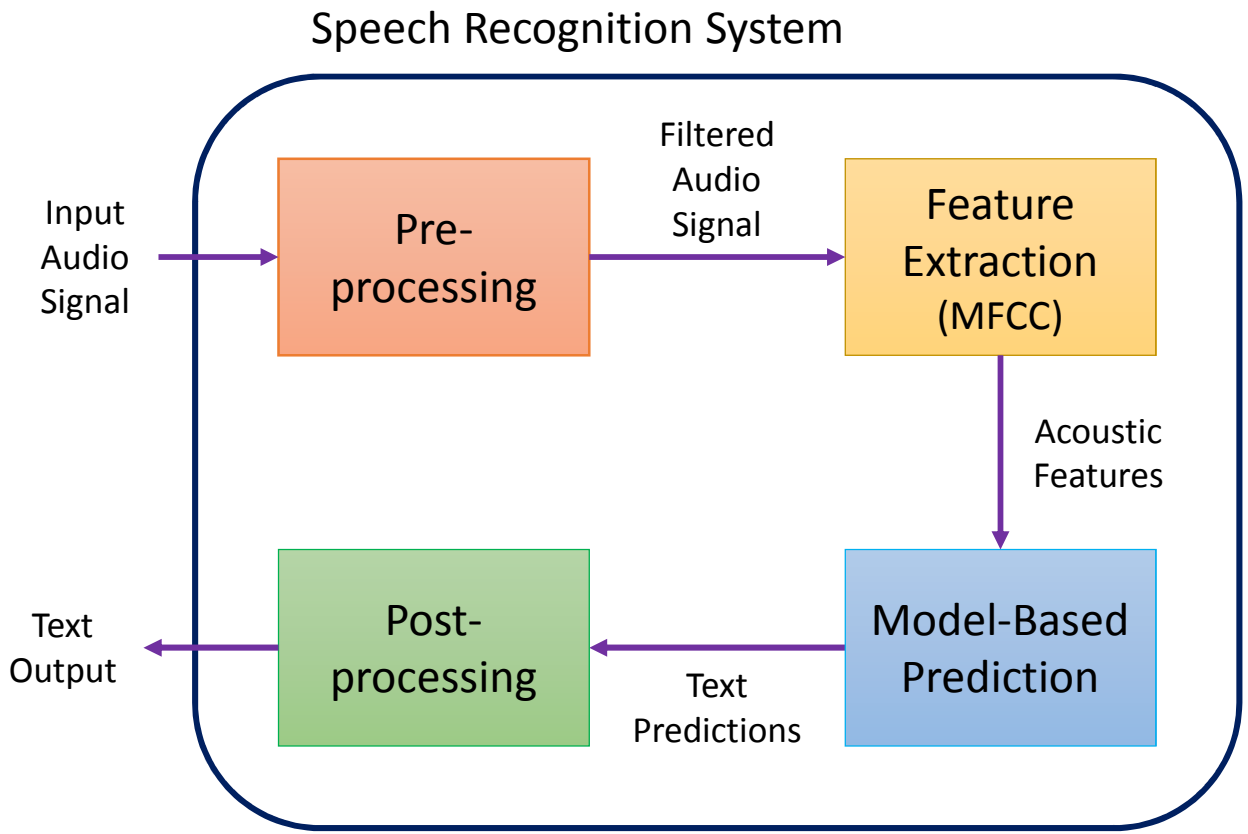
Background: Speech Recognition Overview

- Pre-processing
  - Background noise removal
  - Speech/non-speech segmentation
- Feature Extraction
  - Acoustic features useful for recognizing speech
  - Mel-frequency cepstral coefficients (MFCC) for representing acoustic features



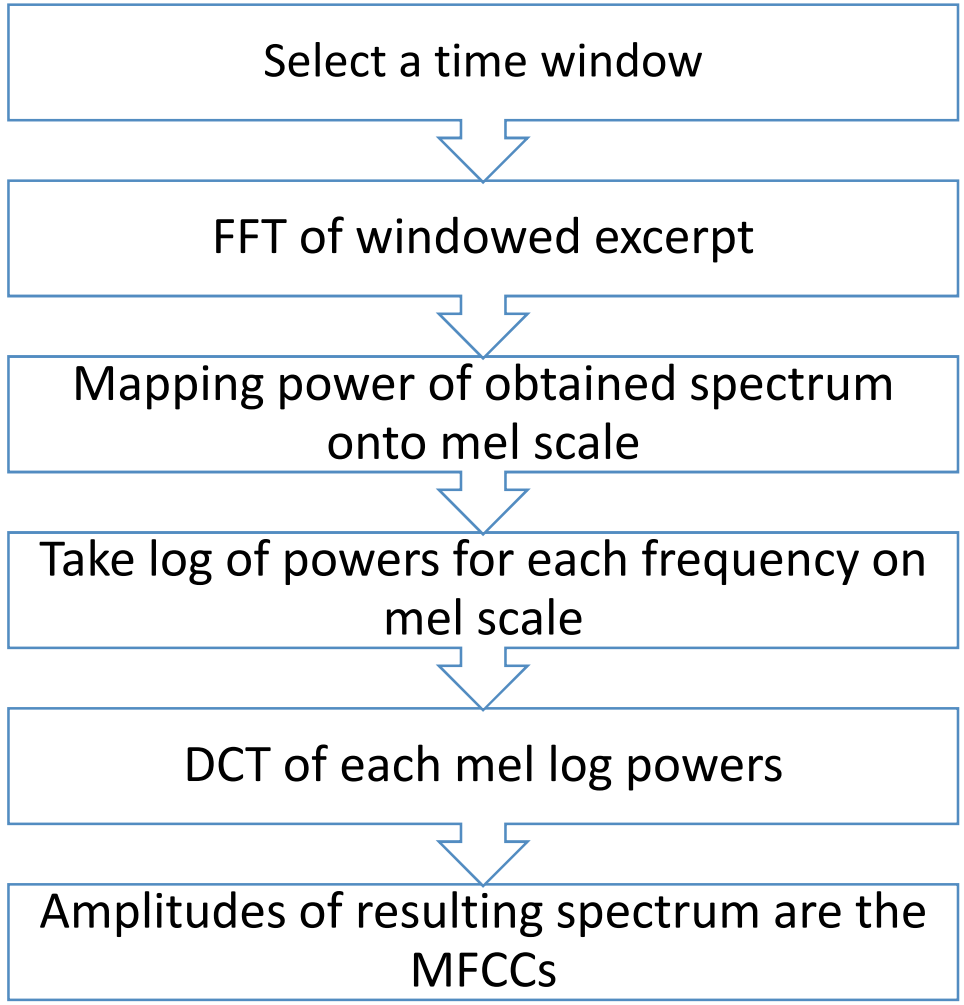
Background: Speech Recognition Overview

- Model Based Prediction
  - Extracted acoustic features of input signal matched against existing models
  - Models typically constructed using statistical approaches
- Post-processing
  - Optionally, rank generated predictions using additional information
  - E.g., enforcing grammar rules, subject matter, locality of words, etc.

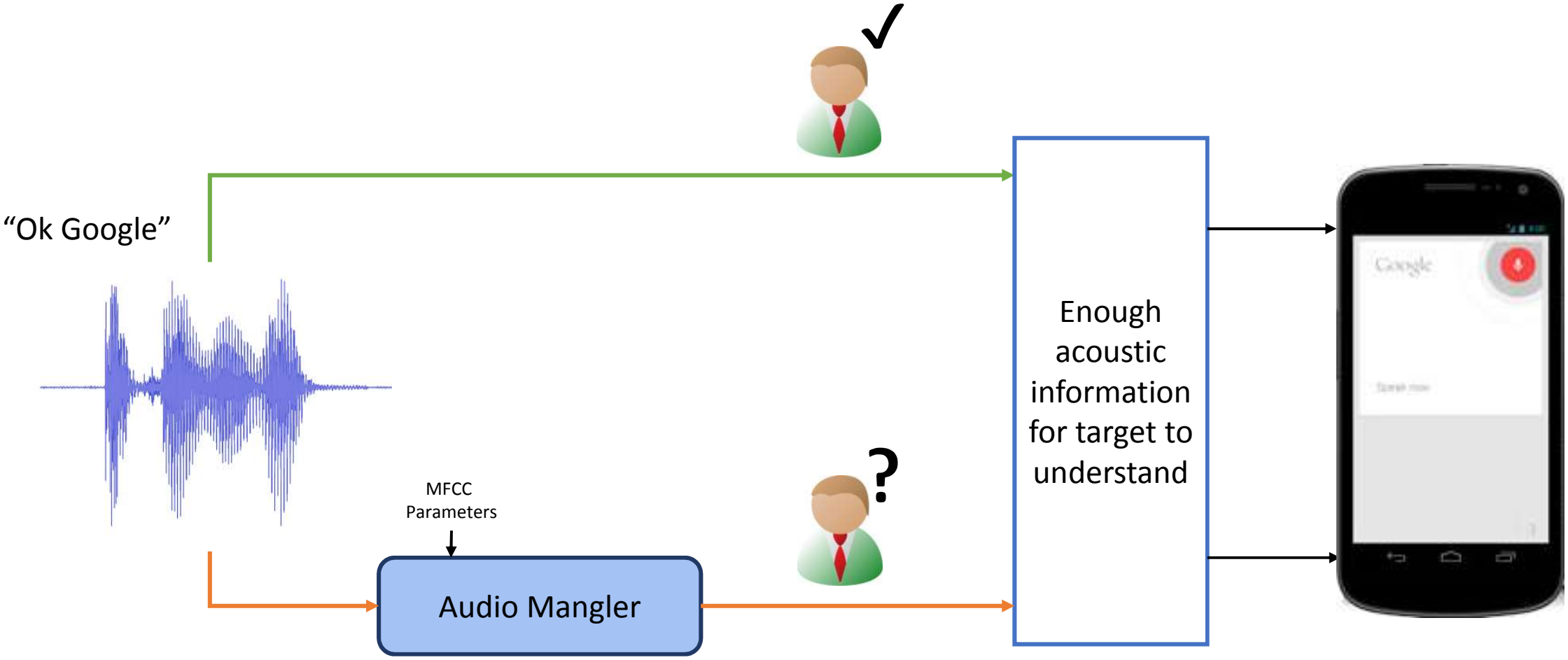


Mel-Frequency Cepstral Coefficients (MFCCs)

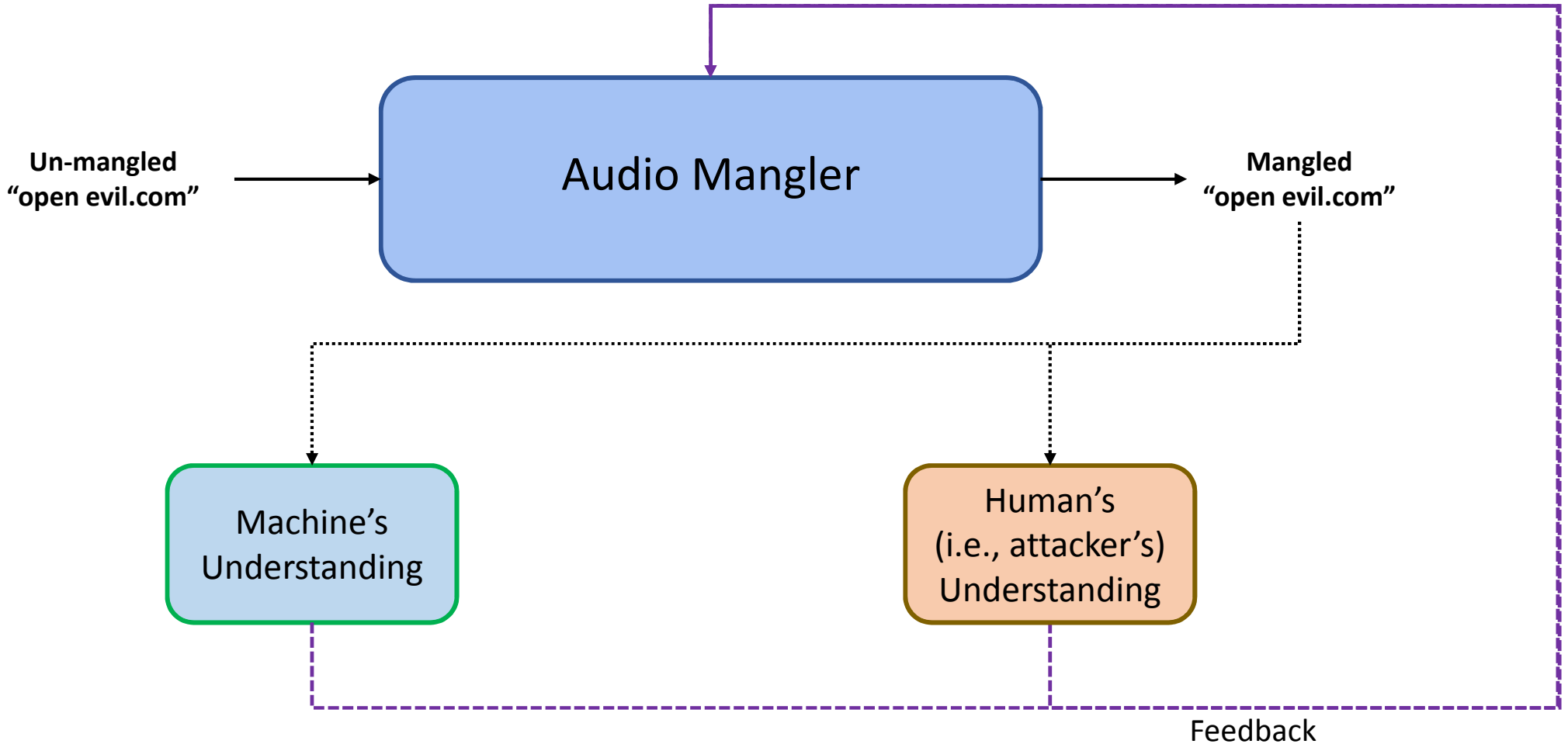
- Cepstral coefficients represent acoustic features in audio signal
- MFCC closely approximates human response to auditory sensation
- Allows for better representation of sound



Attack Overview

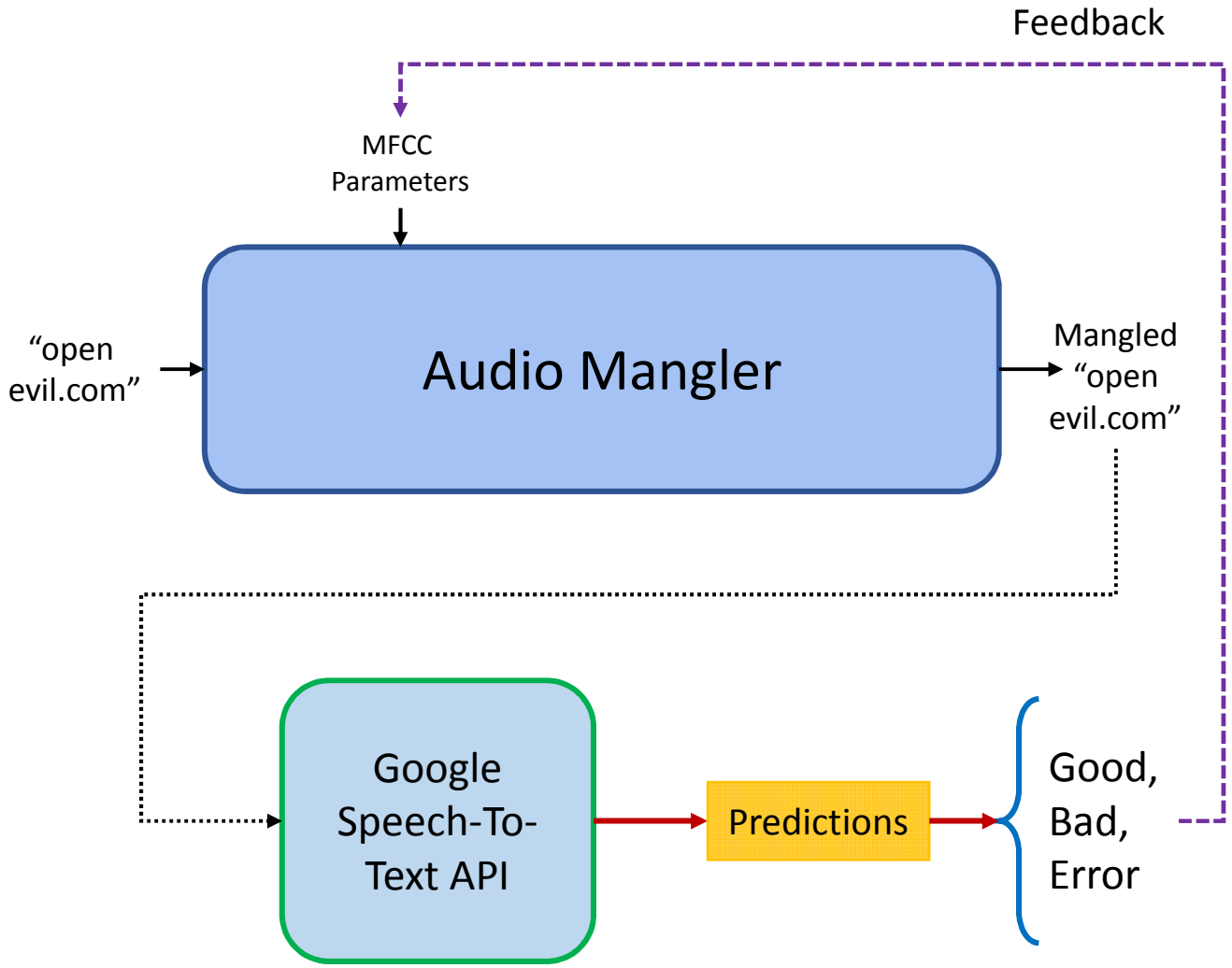


Generating attack commands



MFCC Tuning

- MFCC computation has various parameters
- We modify 4 independent parameters:
  1. `wintime`
  2. `hoptime`
  3. `numcep`
  4. `nbands`
- Experimentally observed the effect of changing each parameter
- Perceived quality of mangled audio varies with different parameter values
- Used Google's Speech-to-Text Speech Recognition API to narrow down parameters

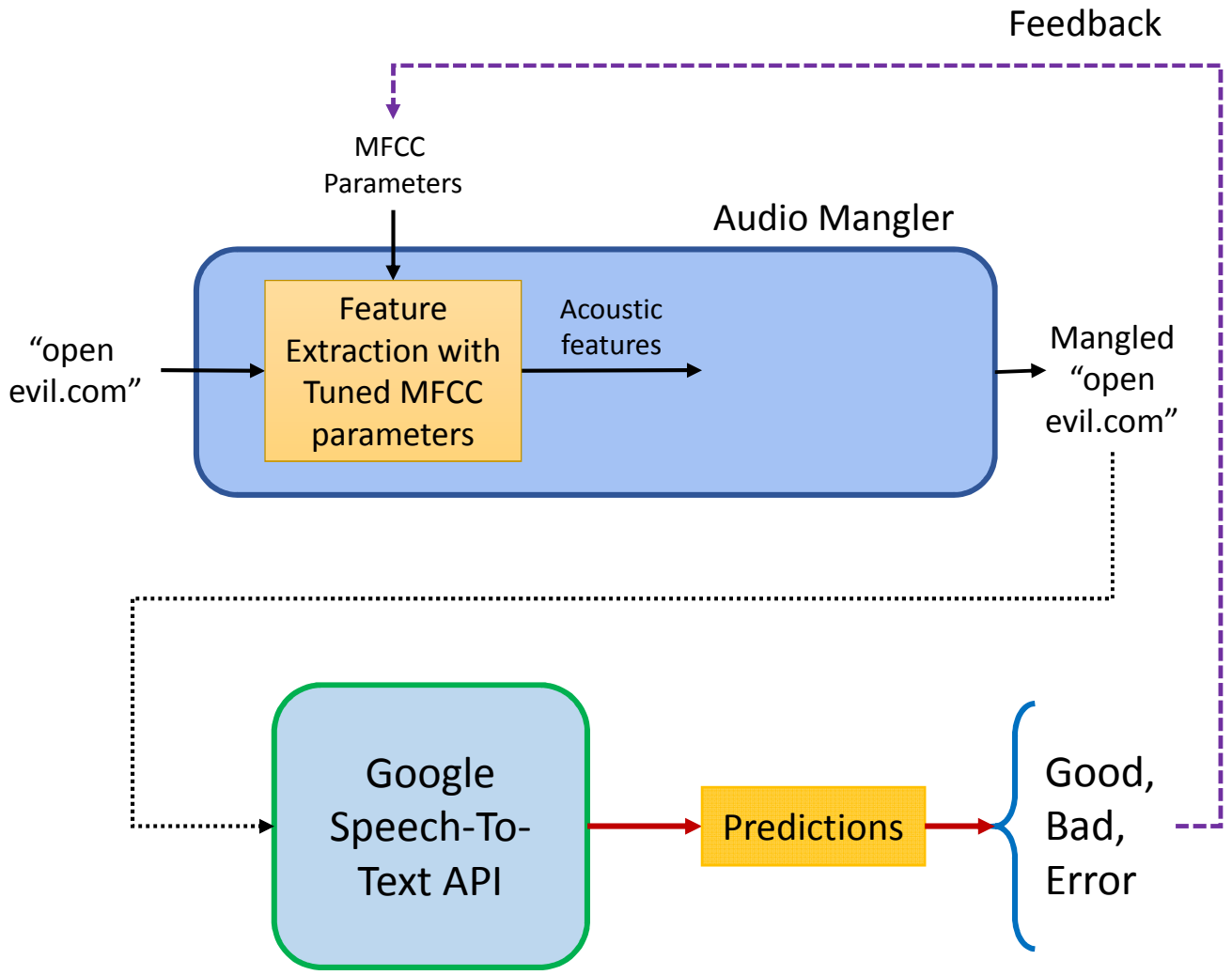


Feature Extraction with Tuned MFCC Parameters

- Tuned parameters are used for computing MFCC

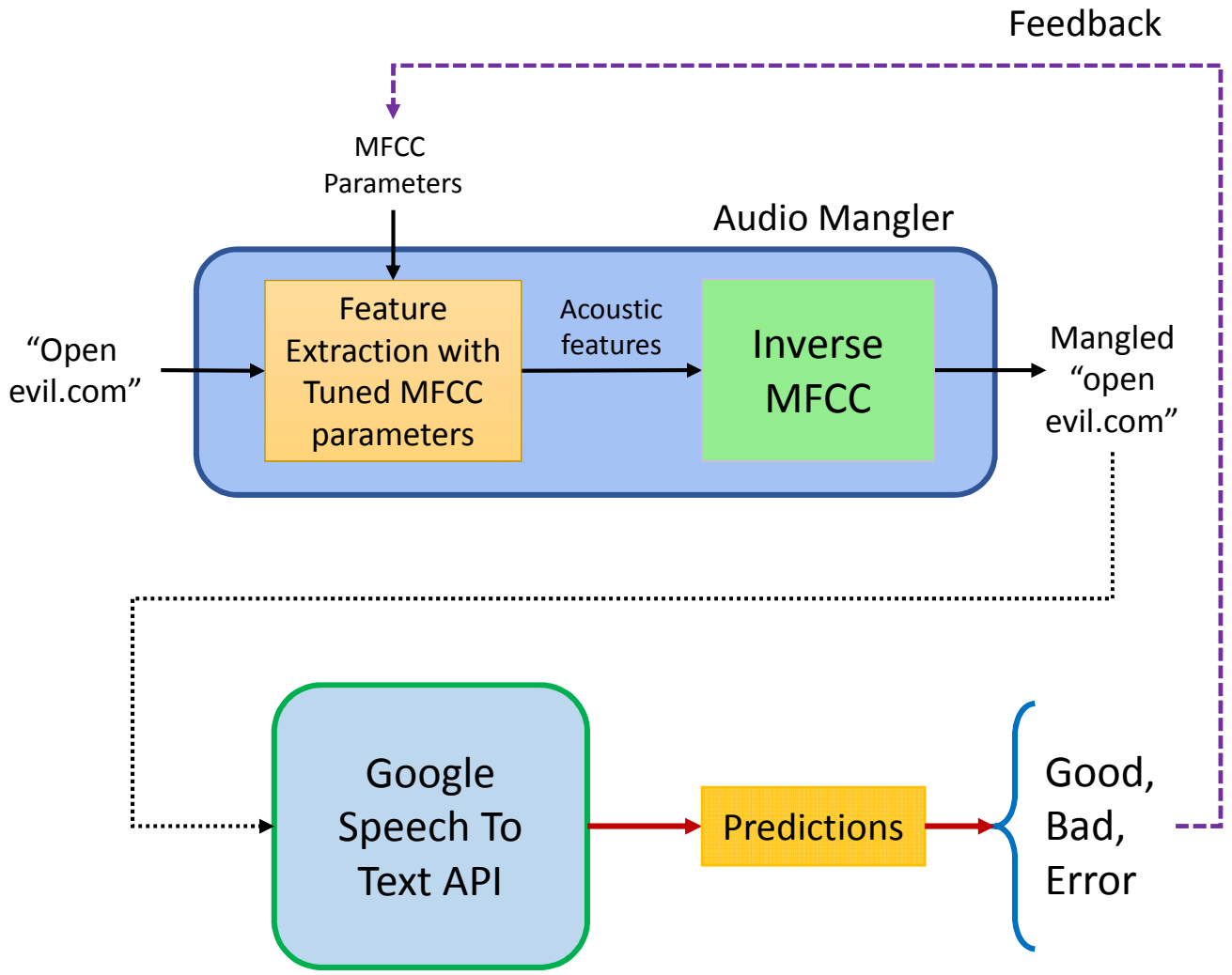
- MFCC computation is *lossy*
  - Signal is considered statistically constant over a small time window
  - Energy level of closely spaced frequencies are aggregated in various frequency regions on mel frequency scale
  - *MFCCs do not retain all information about the original input*

- Tuned MFCC parameters are intended to further increase this loss



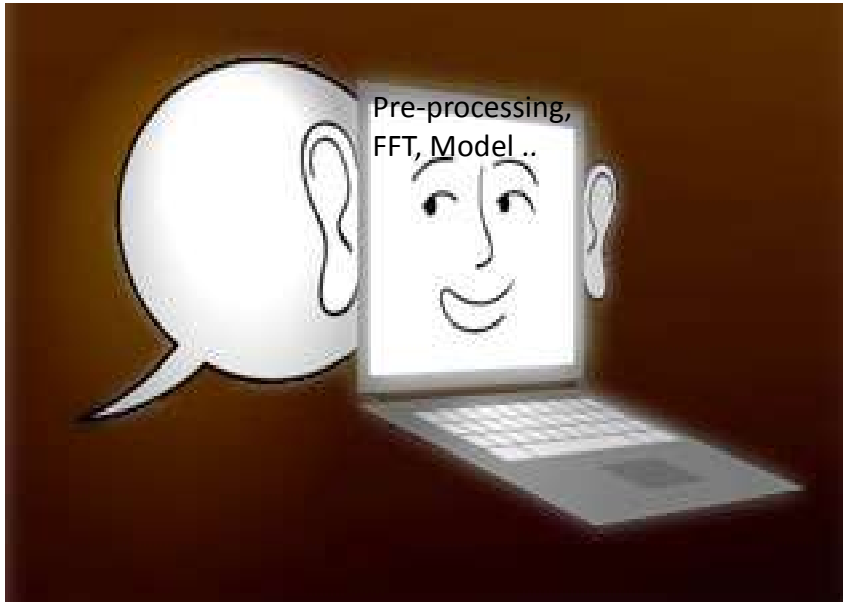
Inverse MFCC Computation

- Extracted audio features converted back to audio signal
- MFCC computation steps are reversed
- White noise added to (re)construct mangled audio command



Mangled commands are crafted to contain acoustic information for a targeted speech recognition system to work, but the human brain doesn't work the same way as machine speech recognition systems!

http://www.ucsf.edu/news/2014/01/111506/ucsf-team-reveals-how-brain-recognizes-speech-sounds



Evaluation

Goal: Determine that mangled commands…

1) …activate functionality on phone (comprehension by machine); and
2) …are difficult for humans to interpret (non-comprehension by human listeners)

Consider 4 types of commands:

- activating the voice command input (i.e., “OK Google”)
- calling a number
- sending a text message to a number
- opening a website (tested against two websites)

Comprehension by Machine

Experimental setup

- Tested the audio commands against Google Now
- Samsung Galaxy S4 smartphone with Android version 4.4.2
- Commands were played via speakers placed ~30 cm from phone

Baseline

(un-mangled commands)

- Un-mangled versions of all commands were played
- All candidates successfully activated functionality on the device

Attack

(mangled commands)

- 500 potential candidates filtered using Google's STT
- 105 candidates manually chosen by 2 authors
- All selected attack candidates successfully activated functionality on the device

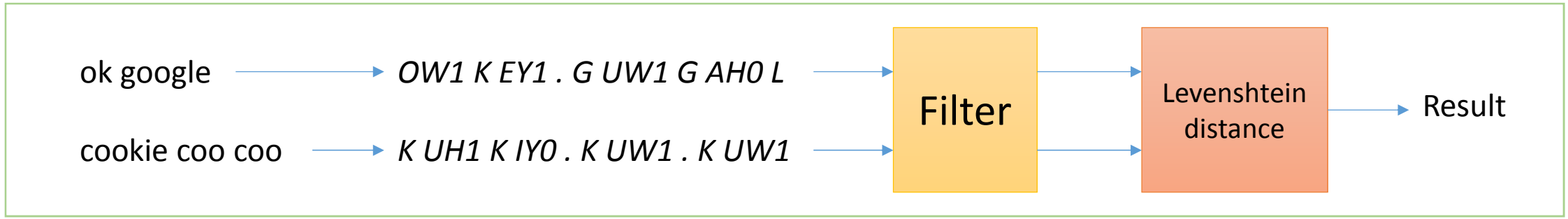
Non-comprehension by Human Listeners

Experimental setup

- Amazon Mechanical Turk user study
- Task: Evaluators given 4 unique audio commands to transcribe
  - Asked to provide their best guess
  - Given bonus ($$$) for correct transcriptions
  - Audio samples included both mangled and un-mangled commands
- Conservative test: evaluators could replay audio, listen under ideal conditions, etc.

Evaluation Metric

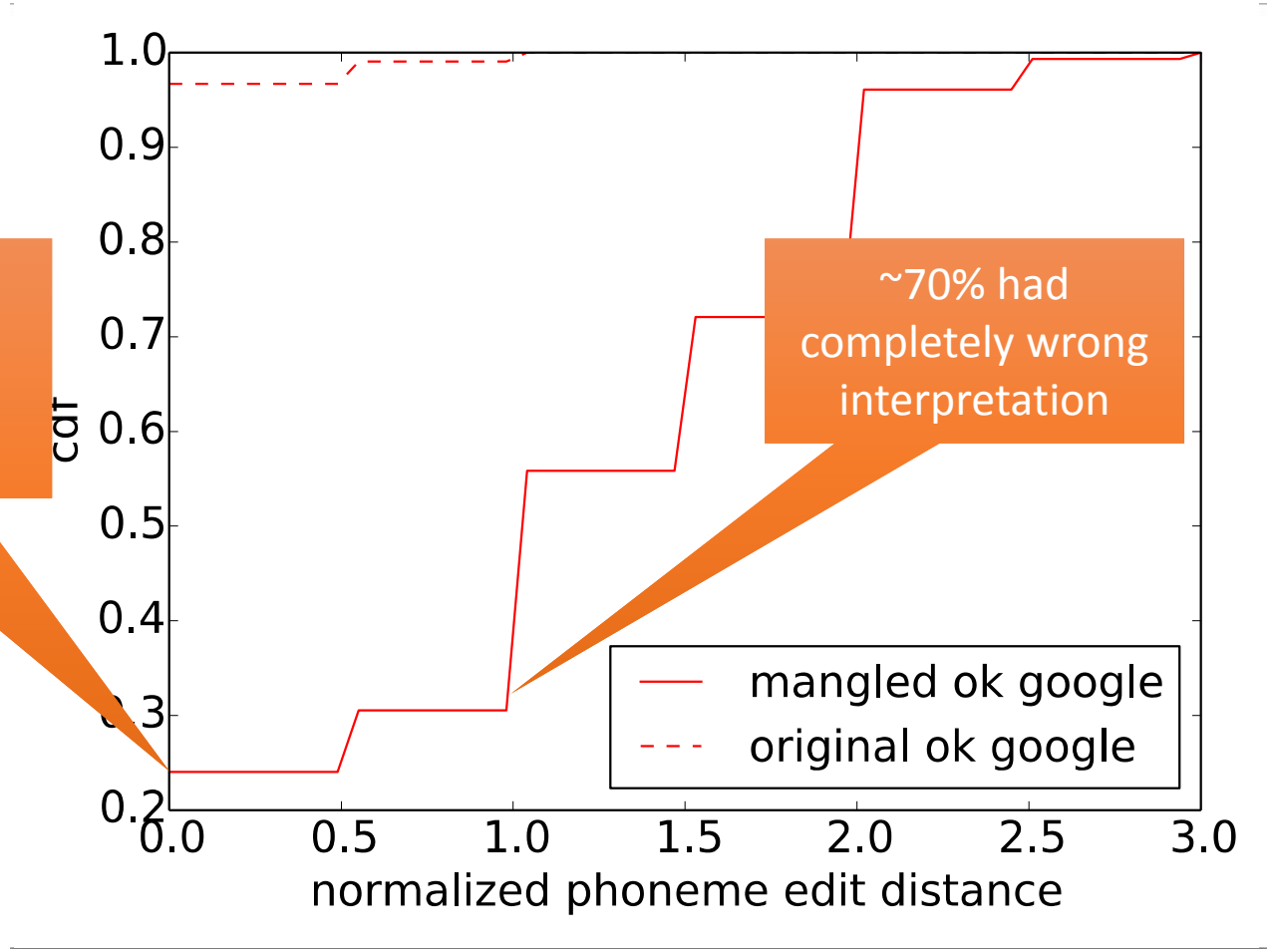
- Levenshtein edit distance (of phonemes) between correct and human-provided transcriptions
- Normalized w.r.t. length of correct transcription



Human Understanding of Mangled Commands

Example transcripts:
- cookie coo coo
- Oh ee oh ah ah
- ha he ho ha
- Seek approval
- puchee poo poo

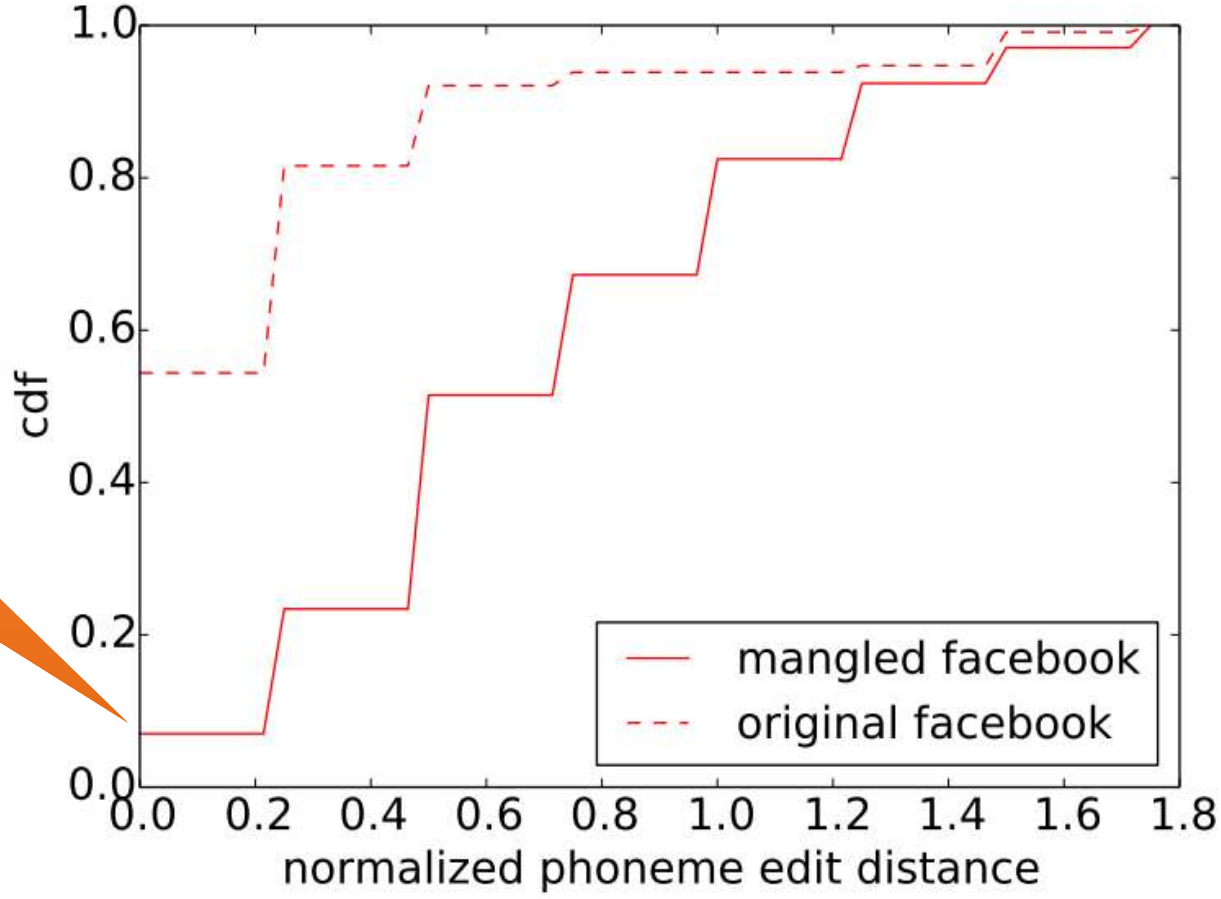
Very few understood the audio correctly!



~70% had completely wrong interpretation

Human Understanding of Mangled Commands

Less than 10% interpreted mangled audio correctly



Example transcripts:

- sh facebook got it
- how do you spell .com
- Small hairs button dot car
- for place spectrum.com
- essa tres quatro dot come

Summary

- Voice command input systems are ubiquitous, but lack security
- There exists a gap in the ability of humans and machines to understand audio signals
- We examined the possibility of exploiting this gap on voice command inputs
- Preliminary results show that this gap can be exploited

Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition

Tavish Vaidya, Yuankai Zhang, Micah Sherr and Clay Shields

Georgetown University