CommanderGabble: A Universal Attack Against ASR Systems Leveraging Fast Speech

Zhaohe Zhang, **Edwin Yang**, Song Fang

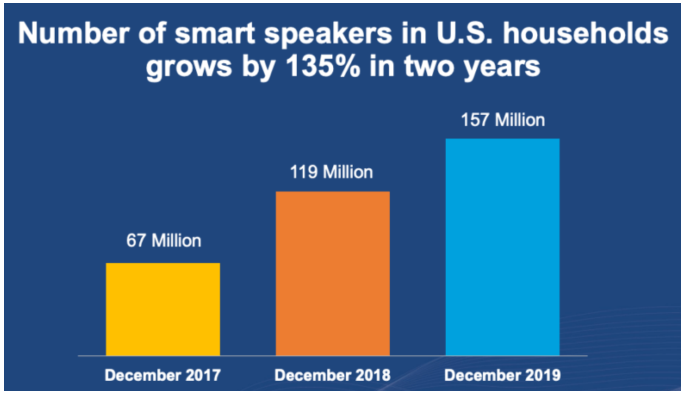
University of Oklahoma

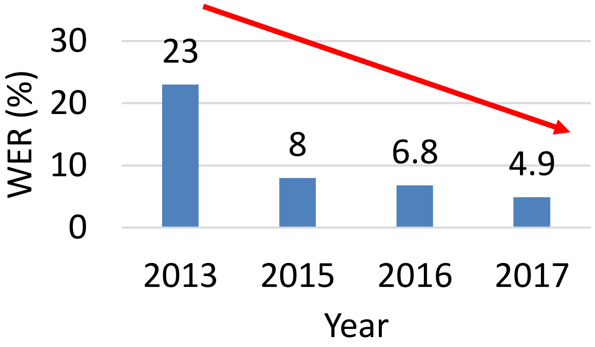
ACSAC 2021

The UNIVERSITY of OKLAHOMA

Background

- Automatic Speech Recognition (ASR) systems are widely available; their accuracy has been greatly improved over time.



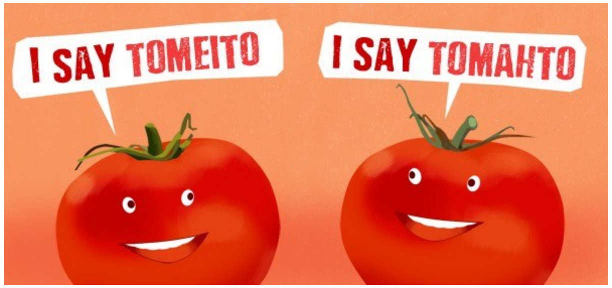


Word error rate for Google's speech recognition

- However, ASR misinterpretations still happen frequently in practice.

SODA!
POP!
SODA!
POP!
SODA!
POP!
SODA!
POP!
SODA!
POP!
SODA!
POP!
SODA!
POP!
COKE.
INTER-REGIONAL GENERIC SOFT-DRINK NAME DEBATE
SHUT
© 2012 - INTERNET-WEBCOMIC.COM

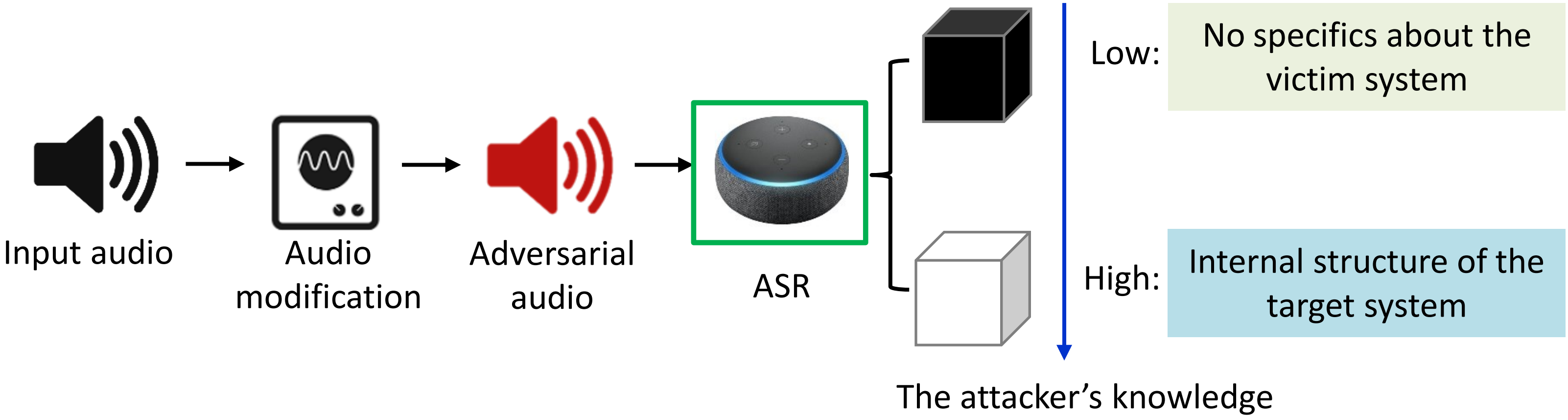
Dialects



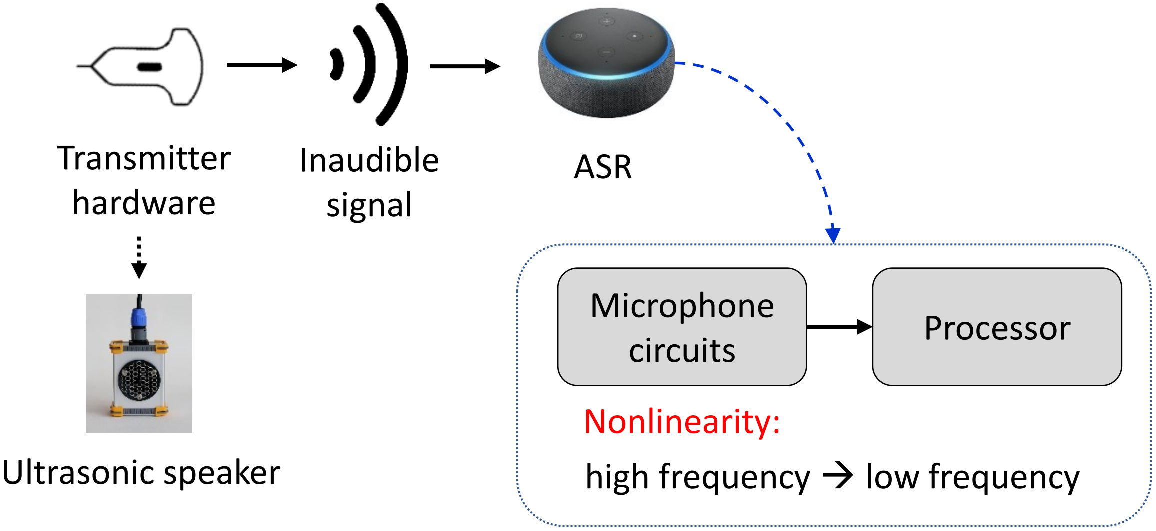
Accents

Existing Attacks on ASRs

- According to the knowledge available for an attacker:

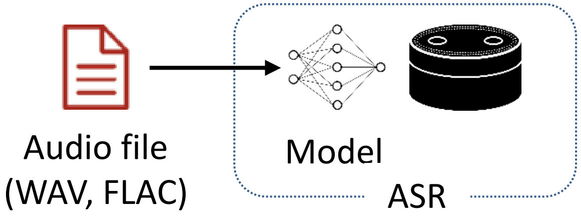


- If specialized hardware is available:



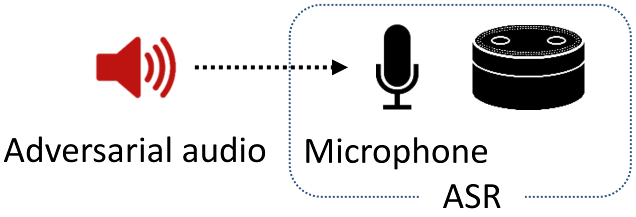
Existing Attacks on ASRs (contd.)

- According to how adversary audio is delivered to ASR:



Over-the-wire

- Audio is directly passed to the target ASR.
- Environmental factors (e.g., noise) have no impact.



Over-the-air

- Audio is played via a speaker towards the target ASR.
- Environmental factors matter.

Phoneme VS. Syllable

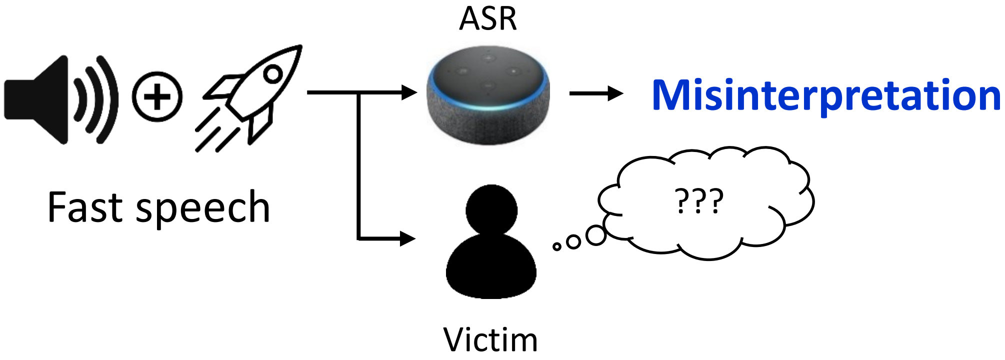
- What are phoneme?
  - ✓ The smallest units of sound which can distinguish two words,
    e.g., /k/ and /b/ → ‘cat’ vs. ‘bat’ => two different words
  - ✓ Classification
    - ➢ Vowel vs. consonant

- What is a syllable?
  - ✓ A single, unbroken sound within a spoken or written word,
    e.g., ‘cat’ vs. ‘water’ => 1 syllable vs. 2 syllables

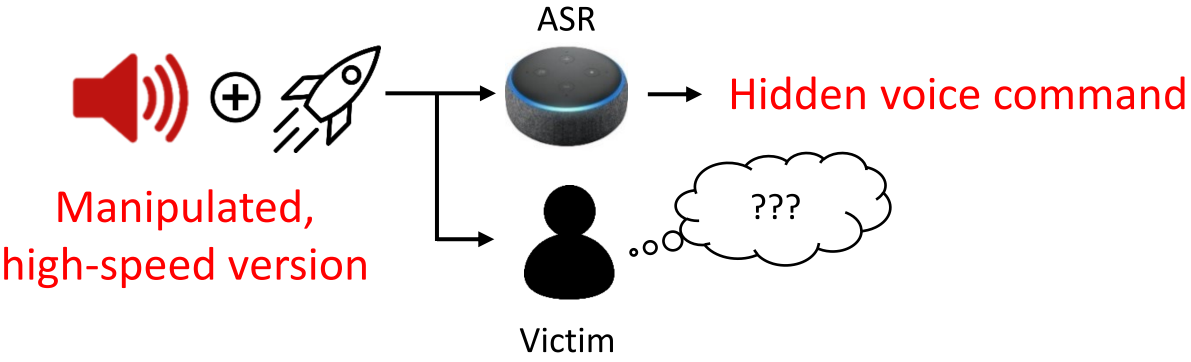
Syllable Structure	Example
V	I
CV	me, see
VC	up, in
CVC	cat, map
CCV	try, sly
CCVC	slip

Motivation

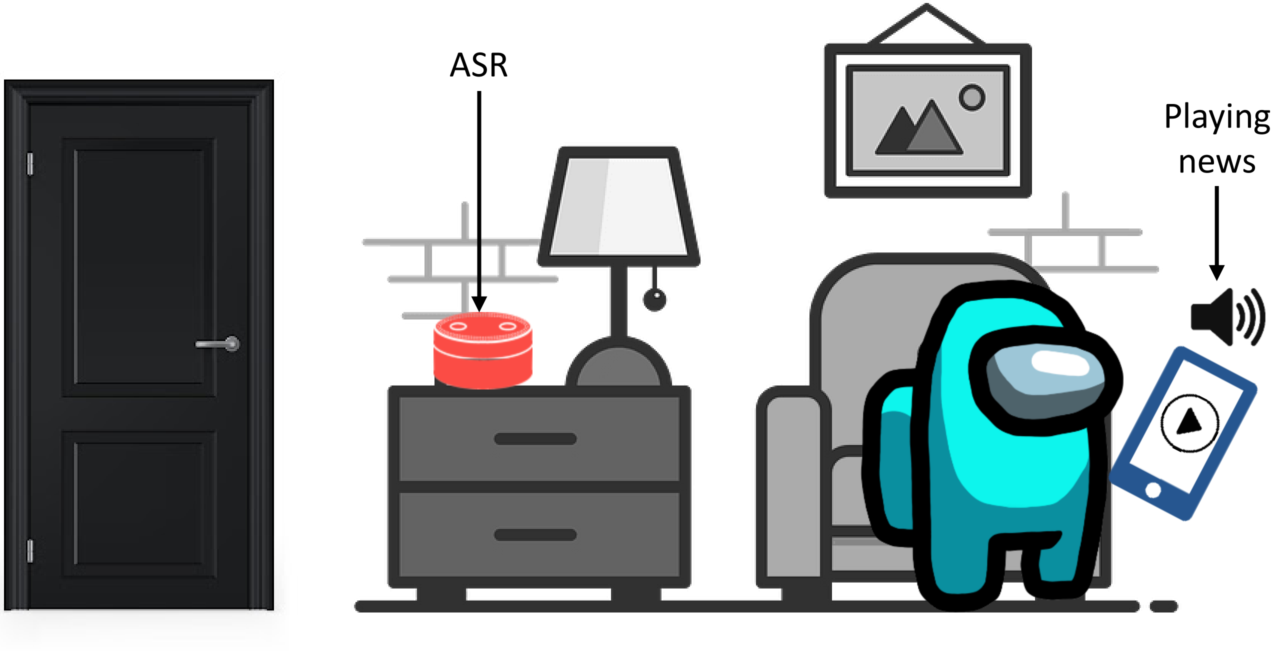
❖ Impact of fast speech



What if we carefully manipulating the phonetic structure of a target voice command?



Attack Scenario



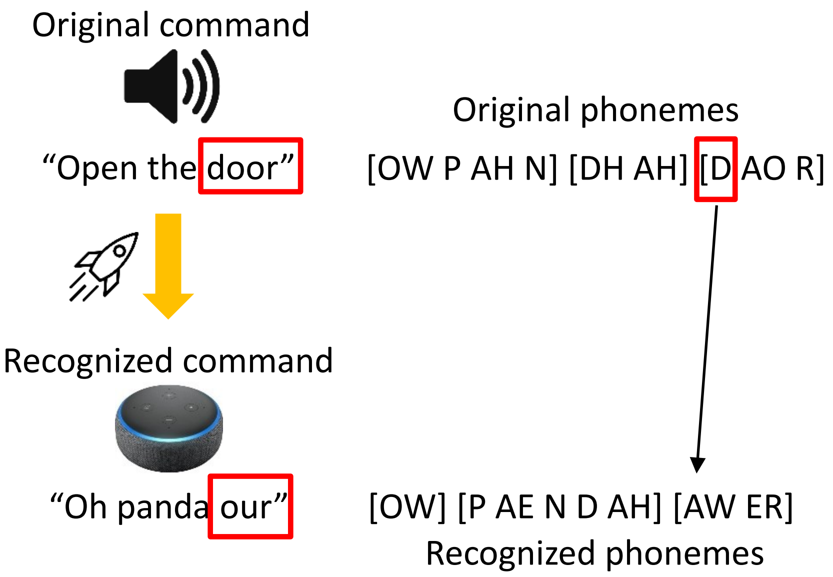
What is happening?
Who’s there?!

UN LOC DAW OUR

🔊

Types of Misinterpretation

- An example command: “Open the door”



Original command

“Open the door”

Recognized command

“Oh panda our”

Original phonemes

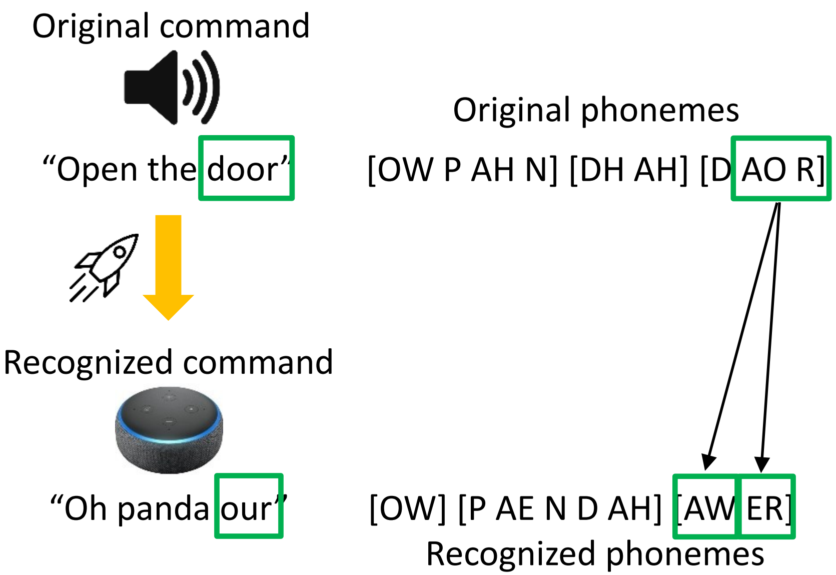
[OW P AH N] [DH AH] [D AO R]

[OW] [P AE N D AH] [AW ER]

Recognized phonemes

✓ Reduction: some phonemes are omitted;

Types of Misinterpretation (contd.)



Original command

"Open the door"

Recognized command

"Oh panda our"

Original phonemes

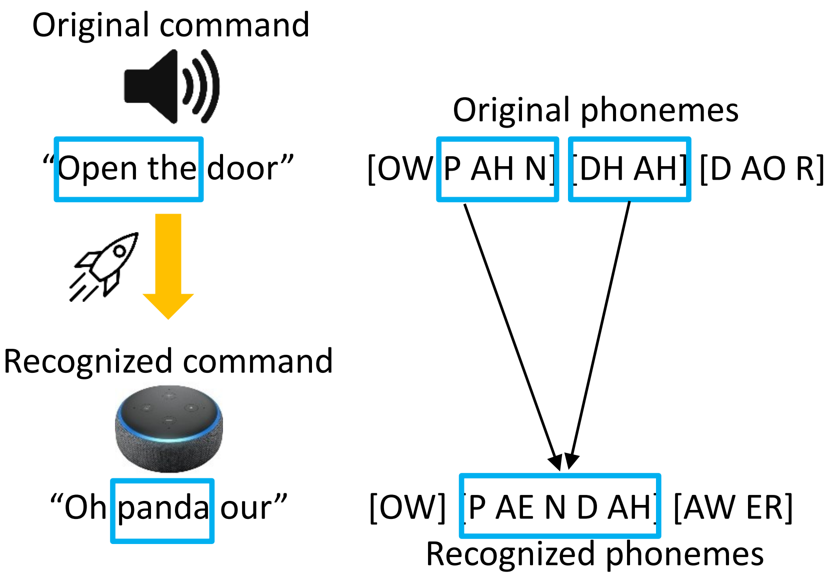
[OW P AH N] [DH AH] [D AO R]

[OW] [P AE N D AH] [AW ER]

Recognized phonemes

- ✓ Reduction: some phonemes are omitted;
- ✓ Replacement: some phonemes are replaced with similar phonemes;

Types of Misinterpretation (contd.)



Original command

"Open the door"

Recognized command

"Oh panda our"

Original phonemes

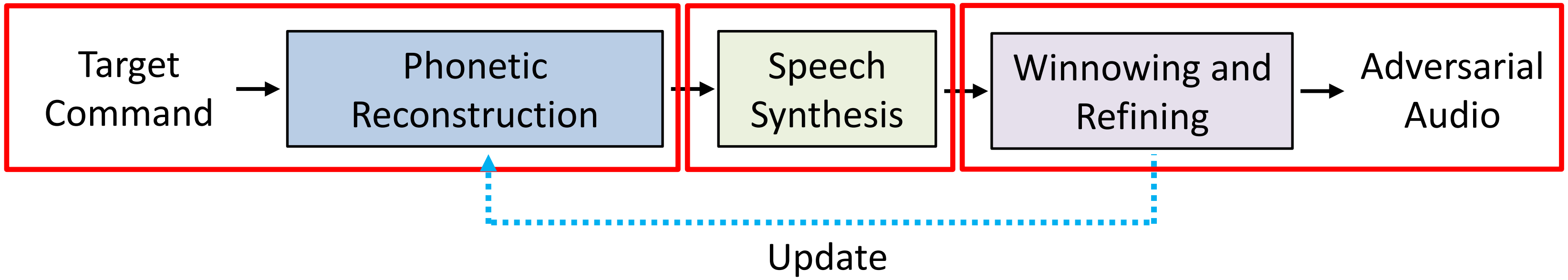
[OW P AH N] [DH AH] [D AO R]

[OW] [P AE N D AH] [AW ER]

Recognized phonemes

- ✓ Reduction: some phonemes are omitted;
- ✓ Replacement: some phonemes are replaced with similar phonemes;
- ✓ Coalescence: some neighboring phonemes are merged together.

System Overview

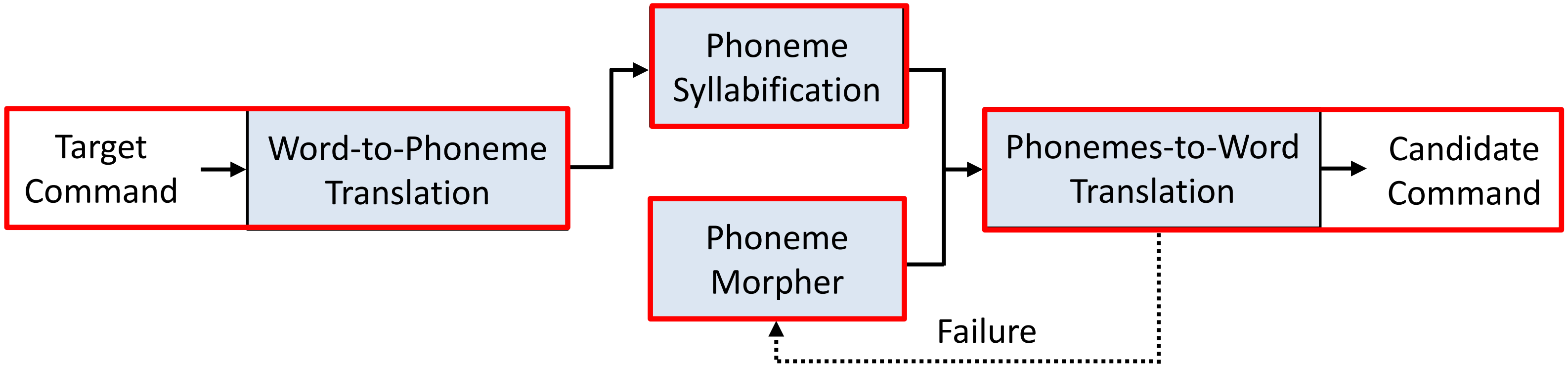


- ✓ Phonetic reconstruction
  - ❖ Extract syllables from target command’s phonetic representation.
  - ❖ Map each word to a new word to generate an adversarial command.

- ✓ Speech synthesis
  - ❖ Generate fast speech of the adversarial command.

- ✓ Winnowing and refining
  - ❖ Verify incomprehensibility and effectiveness.
  - ❖ Update syllabification rules.

Phonetic Reconstruction



Target Command

Word-to-Phoneme Translation

Phoneme Syllabification

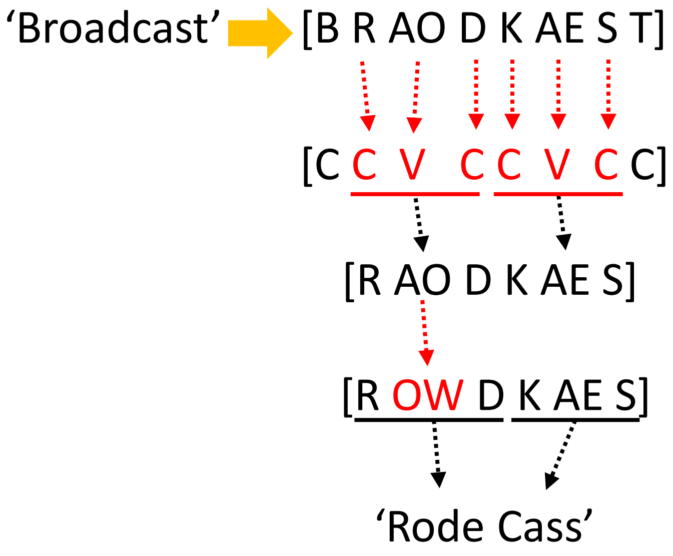
Phoneme Morpher

Phonemes-to-Word Translation

Candidate Command

Failure

- ✓ Word-to-phoneme translation:
- ✓ Phoneme syllabification:
- ✓ Phoneme morpher:
- ✓ Phonemes-to-word translation:



‘Broadcast’ [B R AO D K AE S T]

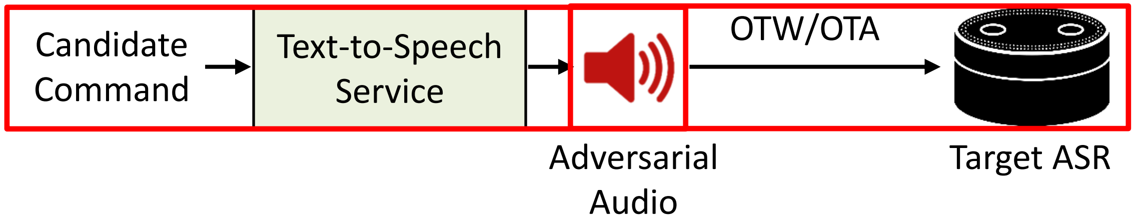
[C C V C C V C C]

[R AO D K AE S]

[R OW D K AE S]

‘Rode Cass’

Speech Synthesis



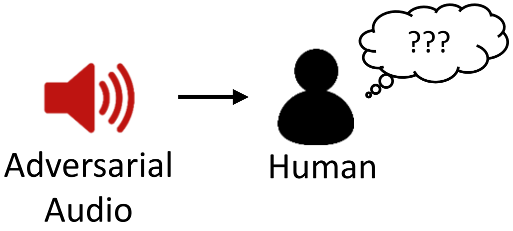
- Generate adversarial audio of a candidate command.
  - ✓ Utilize Google Cloud TTS

- Achieve fast speech by controlling playback speed (2.0x - 3.0x).
  - ✓ Normal speed (≈ 1.0x): Easy to understood by human
  - ✓ Too fast (> 3.0x): ASR fails to recognize due to excessive distortion

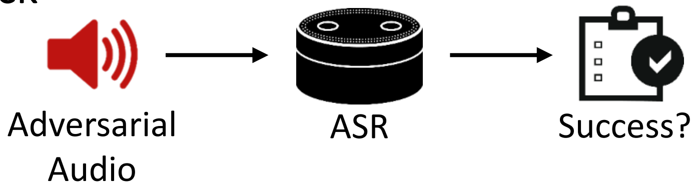
- Generated audio is transmitted to target ASR according to attack scenario

Winnowing and Refining

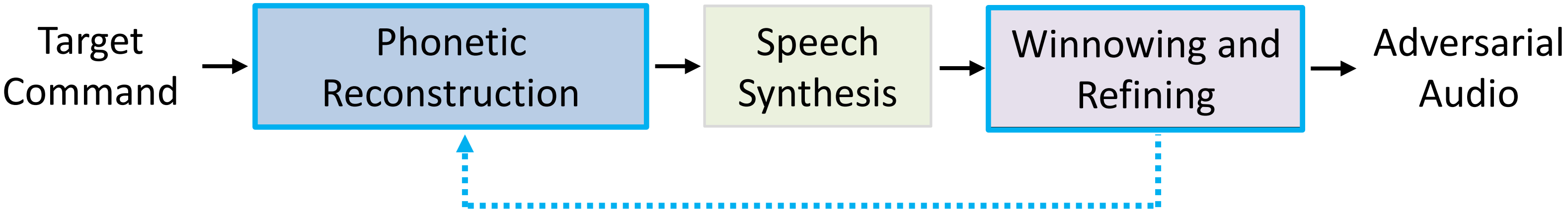
- Winnow out ineffective candidate adversarial audio.
  - ✓ Intelligibility check



  - ✓ Execution check



- Syllabification modifier
  - ❖ If either check fails, the adversary modifies syllabification rules correspondingly.



Evaluation Setup

- Over-the-wire attack
  - ✓ Select 100 ASR commands

amazon Transcribe

IBM Watson

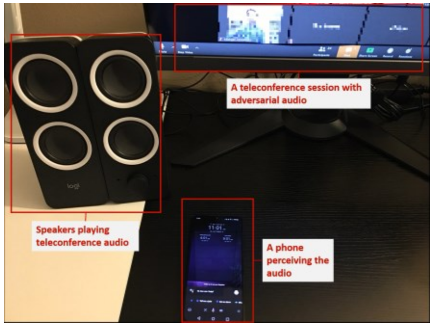
Google Cloud Speech to Text

Microsoft Azure

- Over-the-air attack

Speakers playing adversarial audio
A phone perceiving the audio

Household


A teleconference session with adversarial audio
Speakers playing teleconference audio
A phone perceiving the audio

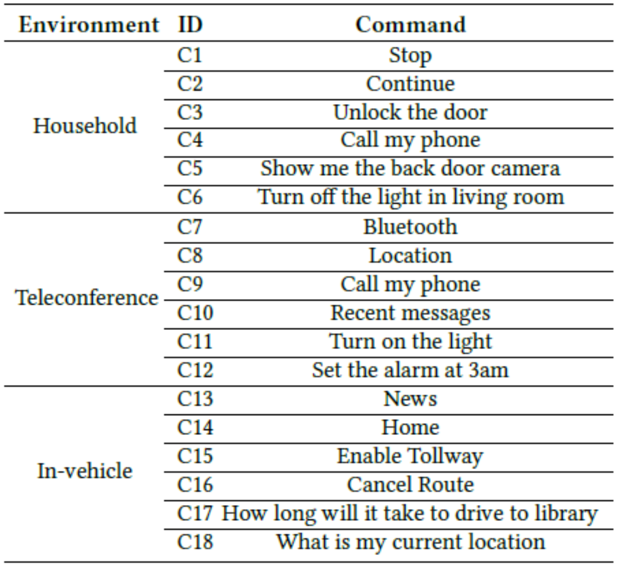
Teleconference

Phone 1 perceiving the audio while navigating to the destination
Phone 2 playing the adversarial audio through the car speaker

Vehicle

  - ✓ 6 commands for each environment

Over-the-air attack commands

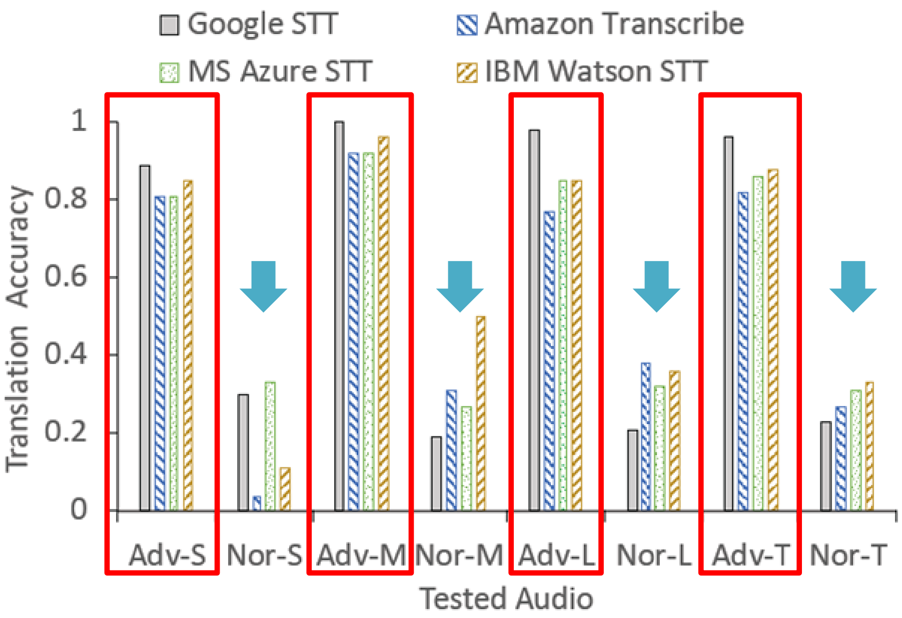


Environment	ID	Command
Household	C1	Stop
	C2	Continue
	C3	Unlock the door
	C4	Call my phone
	C5	Show me the back door camera
	C6	Turn off the light in living room
Teleconference	C7	Bluetooth
	C8	Location
	C9	Call my phone
	C10	Recent messages
	C11	Turn on the light
	C12	Set the alarm at 3am
In-vehicle	C13	News
	C14	Home
	C15	Enable Tollway
	C16	Cancel Route
	C17	How long will it take to drive to library
	C18	What is my current location

Over-the-wire Translation Accuracy

Length of commands

Command Length	Number of Words
Short	1
Medium	2-3
Long	>3



OTW Translation accuracy for fast speech audio files

- ❖ Most of adversarial audios are correctly recognized.
- ❖ Highest accuracy (95%) for medium length commands.
- ❖ Low accuracy (28%) for normal commands.

Over-the-air Attack Success Rate

- Target ASRs

Google
Pixel 4

Amazon
Echo Dot

Lenovo
ThinkPad X1

- Adversarial wake-up word test

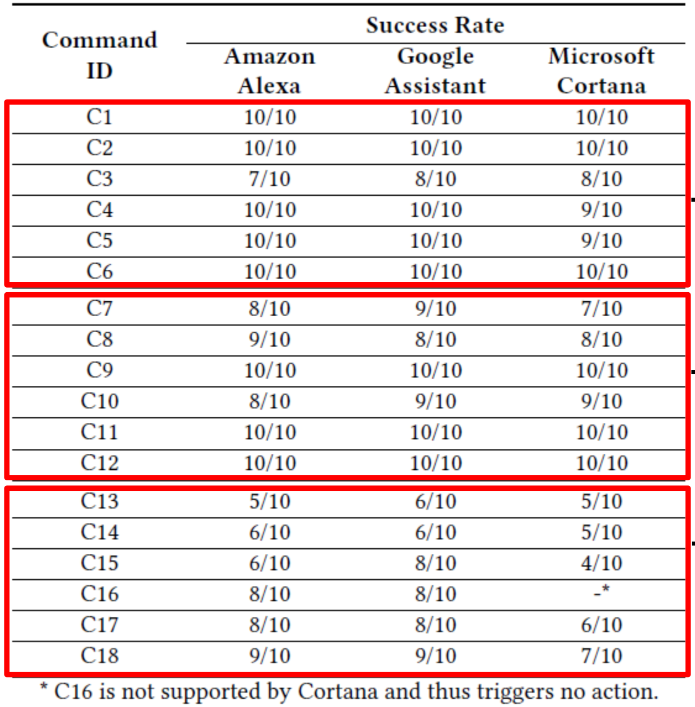
Wake-up words and their adversarial commands

Wake-up Word	Adversarial Command	Playback Speed	Successful?
Ok Google	kaye go oh	2.0x-2.1x	✓
Alexa	a leh sa	2.0x-2.1x	✓
Hey Cortana	hye core ta	2.0x-2.1x	✓

❖ All wake-up words are correctly recognized by target ASRs.

Over-the-air Attack Success Rate (contd.)

Attack performance on different ASRs



Command ID	Success Rate		
	Amazon Alexa	Google Assistant	Microsoft Cortana
C1	10/10	10/10	10/10
C2	10/10	10/10	10/10
C3	7/10	8/10	8/10
C4	10/10	10/10	9/10
C5	10/10	10/10	9/10
C6	10/10	10/10	10/10
C7	8/10	9/10	7/10
C8	9/10	8/10	8/10
C9	10/10	10/10	10/10
C10	8/10	9/10	9/10
C11	10/10	10/10	10/10
C12	10/10	10/10	10/10
C13	5/10	6/10	5/10
C14	6/10	6/10	5/10
C15	6/10	8/10	4/10
C16	8/10	8/10	-*
C17	8/10	8/10	6/10
C18	9/10	9/10	7/10

* C16 is not supported by Cortana and thus triggers no action.

✓ Average success rates for three ASRs:

❖ Home: 95%, 97%, 93%

❖ Teleconference: 92%, 93%, 90%

❖ Noisy environment results decreased success rates.

- Human comprehensibility test
  - ✓ Recruited 28 volunteers
  - ✓ None could comprehend any adversarial audio

Conclusion

- ✓ We systematically explore misinterpretations introduced by fast speech and analyze the consequent phonetic structure variations.

- ✓ By combining phoneme manipulation with fast speech, we develop *CommanderGabble* for a model-agnostic and easily-constructed adversarial attack against ASR systems.

- ✓ We perform extensive experiments to evaluate feasibility robustness, and suspiciousness of *CommanderGabble*.

The UNIVERSITY of OKLAHOMA

Thank you!
Any questions?

← Feel free to check our artifact web page!

commandergabble.info