Evaluating Fast Speech Based Adversarial Audio Attack

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

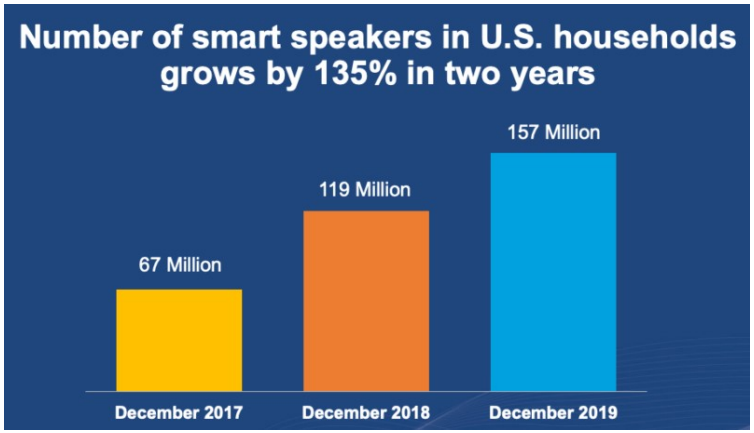
University of Oklahoma

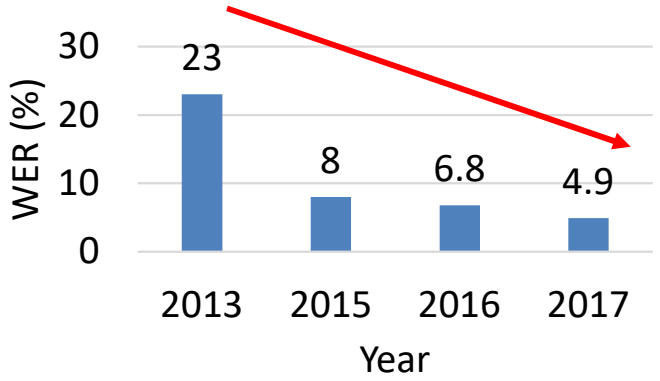
LASER 2021 Workshop

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

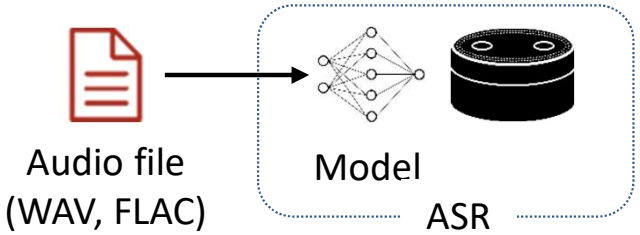
- However, ASR misinterpretations still happen frequently in practice.

Dialects

Accents

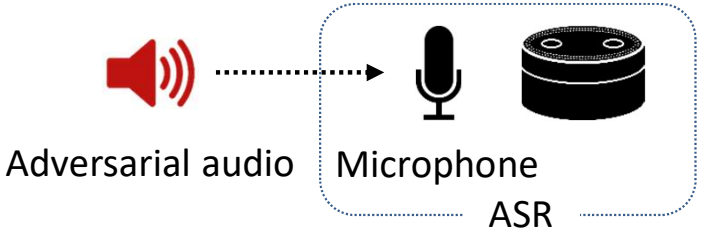
Existing Attacks on ASRs

- 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.

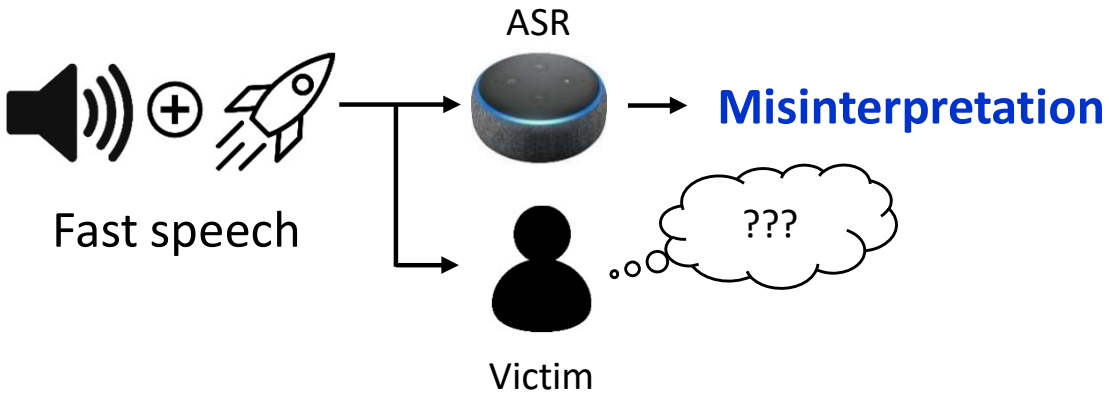
Phoneme VS. Syllable

- What are phoneme?
  - ✓ The smallest units of sound which can distinguish two words,
    e.g., /k/ and /b/ → '**c**at' vs. '**b**at' => 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

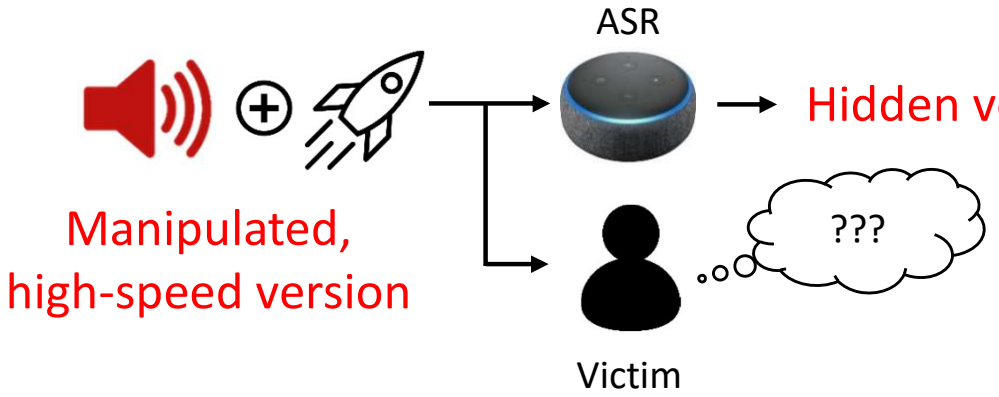
Syllable Structure	Example
V	I
CV	me, see
VC	up, in
CVC	cat, map
CCV	try, sly
CCVC	slip

Motivation

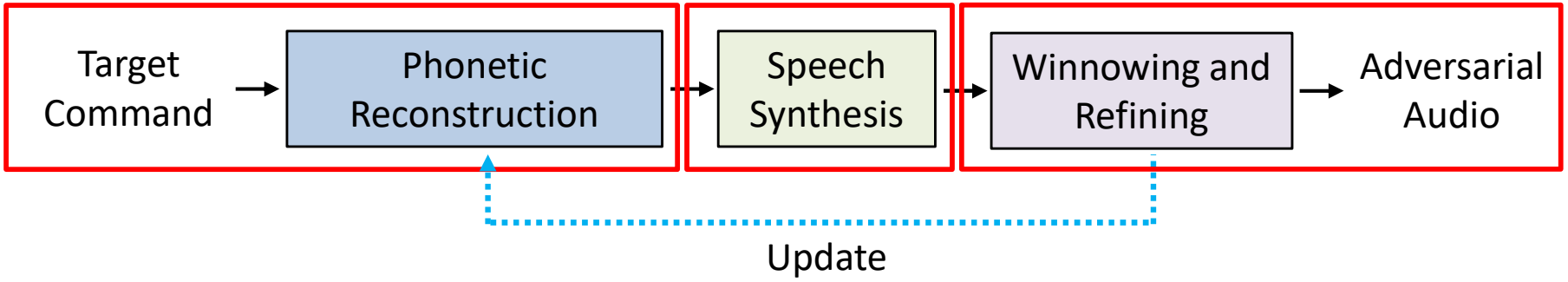
- Impact of fast speech



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



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.

Experiment Design

- How to design a comprehensive experiment?
  - ✓ Two types of ASR services:
    - ❖ Transcribe service: Over-the-wire attack
    - ❖ Voice assistant device: Over-the-air attack
  - ✓ Ensure suspiciousness/effectiveness of proposed attack
    - ❖ Can human participant recognize adversarial audio?
    - ❖ Does simply increasing speed of speech enough?
    - ❖ User test is required
- Over-the-wire attack can be automated
  - ✓ Run batch script for submitting audio files
- Over-the-air attack requires more effort
  - ✓ Play an audio file
  - ✓ Verify recognition result on target ASR

Experiment Design (contd.)

- Over-the-wire attack
  - ✓ Target transcribe services

  - ✓ Randomly select 100 popular ASR commands with different length

Command Length	Number of Words	Example	Number of Commands
Short	1	Louder	27
Medium	2-3	Turn up volume	26
Long	>3	Where is my stuff?	47

→ 100

  - ✓ Vary playback speed 2.0x - 3.0x, increment by 0.1x → 11

X

=

1100

Experiment Design (contd.)

- Over-the-air attack

✓ Wake-up words for triggering target ASR

Wake-up words and their adversarial commands

Wake-up Word	Adversarial Command	Playback Speed
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

Experiment Design (contd.)

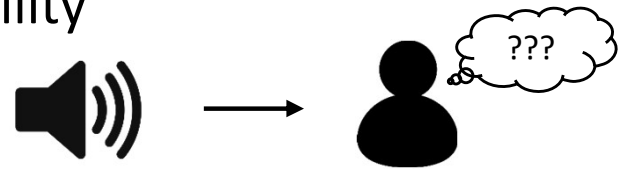
- Over-the-air attack (contd.)
  - ✓ 3 realistic environments
  - ✓ 6 commands for each situation

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

Experiment Design (contd.)

- Human comprehensibility



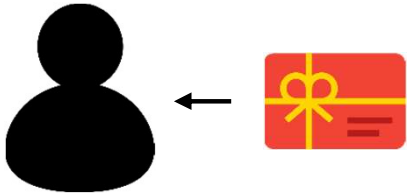
Adversarial Audio Victim

- Recruiting participants

E-mail flyer and in-person

28 participants (11 females, 17 males)

All participants are fluent in English



Gift card is provided after experiment

- Participant tasks
  1. Listen to each voice command
  2. Describe understanding on a questionnaire

- Commands are not disclosed to prevent priming effects

Performance Metrics

- Over-the-wire and Over-the-air attack
  - ✓ Translation accuracy: Percentage of successful attacks

- Human comprehensibility test
  - ✓ Word Error Rate:

$$WER = \frac{Substitution + Deletion + Insertion}{Number\ of\ \boldsymbol{words}\ in\ original\ command}$$

2 Deletions

Original command : The cat sat on the mat ← 6 words

Recognized command: The cat sat bat

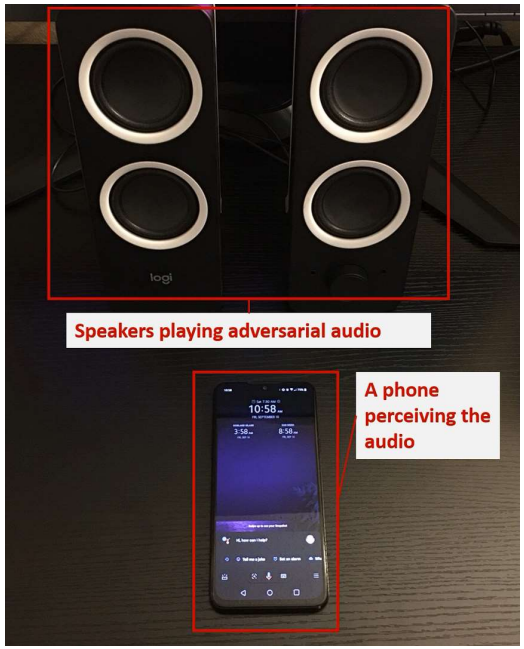
1 Substitution

$$\frac{1+2+0}{6} = 50\%$$

  - ✓ Phoneme Error Rate (PER):

$$PER = \frac{Substitution + Deletion + Insertion}{Number\ of\ \boldsymbol{phonemes}\ in\ original\ command}$$

Over-the-air Experiment Environment



Household

Teleconference

Vehicle

- ✓ Distance between speaker and target ASR: 30 – 200 cm
- ✓ Indoor noise level: 15-20 dB
- ✓ In-vehicle noise level: 60 dB (engine on)

Over-the-air Attack Demonstration

Video available at commandergabble.info

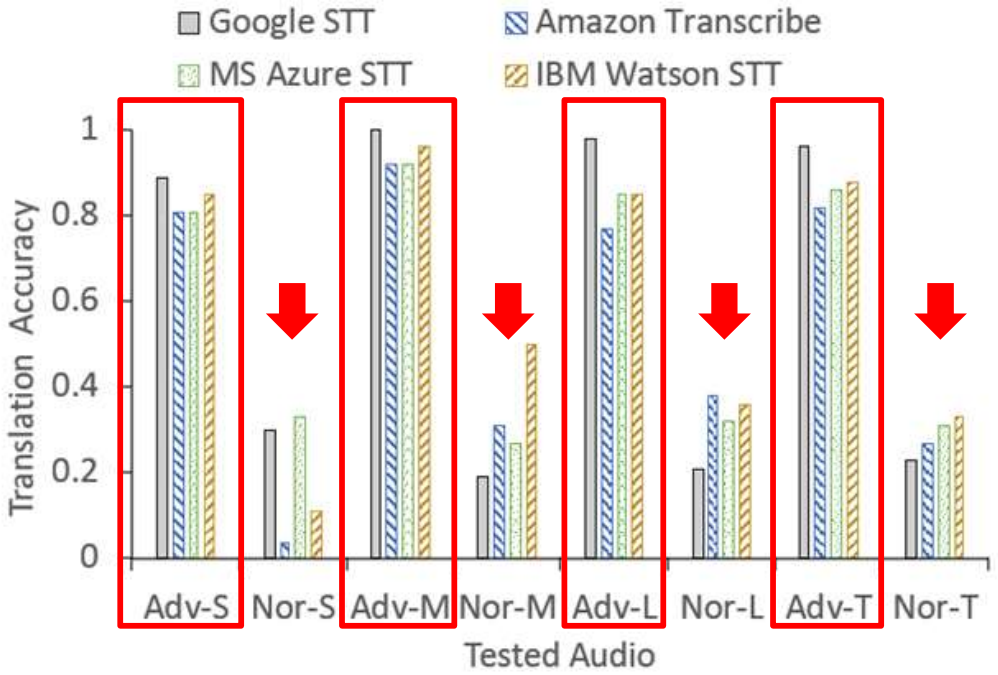
Experimentation Artifact

- Artifact is available in public
  - ✓ Evaluated as a functional artifact

- We provide:
  - ✓ Adversarial audio examples
  - ✓ Experiment procedure for key experiments
  - ✓ Adversarial audio files utilized in the experiments
  - ✓ Questionnaire for human comprehensibility test

commandergabble.info ← Feel free to visit our web page!

Over-the-wire Translation Accuracy



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

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 Result

- Each participant listens to 42 fast speech audios
  - ✓ 21 adversarial audios
    (3 wake-up words + 6 scenario specific commands x 3 scenarios)
  - ✓ 21 corresponding benign audios (fast speech)

- Order of played audios is randomized

- None could comprehend any adversarial audio file
  - ✓ Measured WERs and PERs are consistently above 0.5
  - ✓ More than half are greater then or equal to 1.0

- WER and PER for each adversarial audio are greater than the corresponding normal 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?