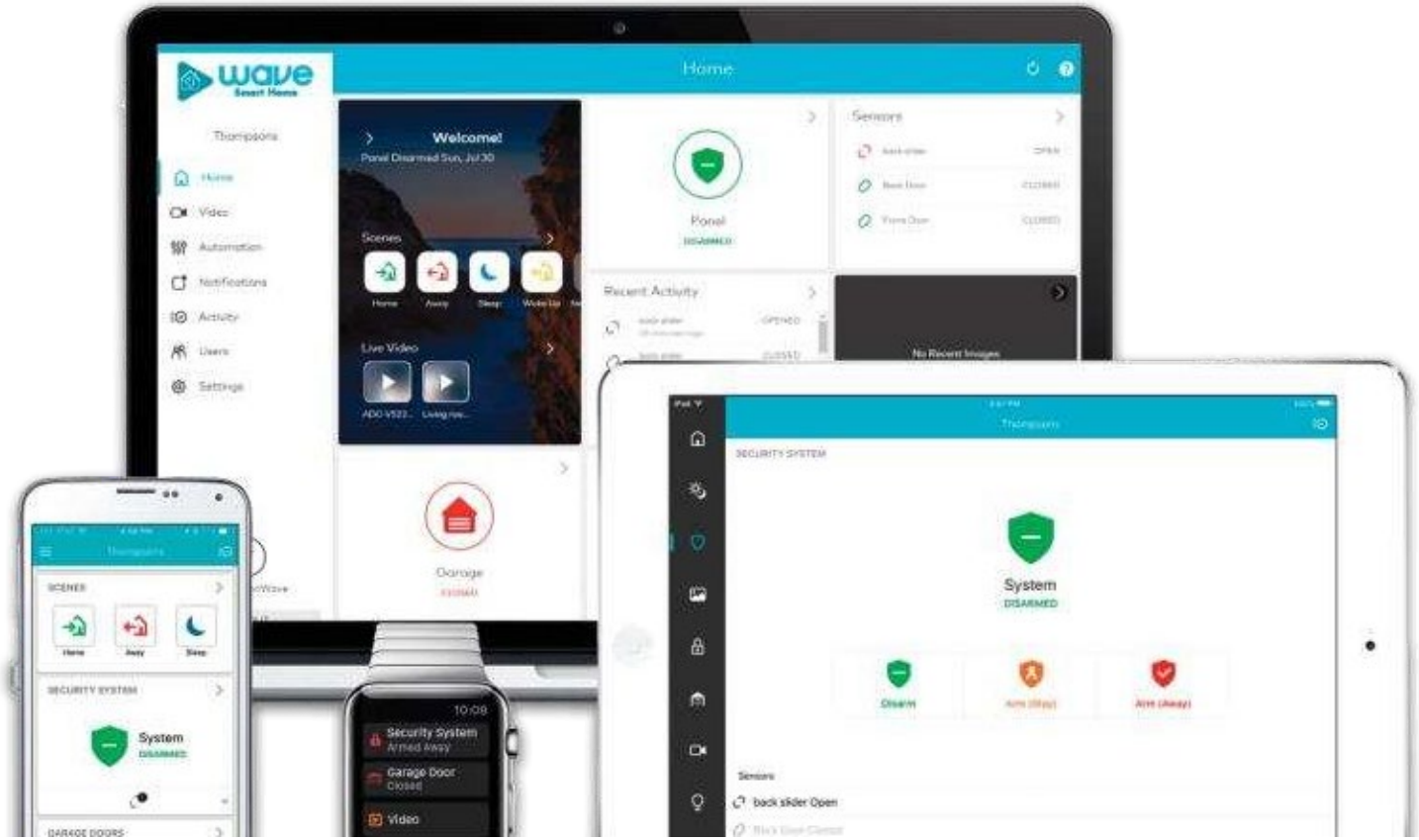
DEVIL'S WHISPER: A GENERAL APPROACH FOR PHYSICAL ADVERSARIAL ATTACKS AGAINST COMMERCIAL BLACK-BOX SPEECH RECOGNITION DEVICES

Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao,
Shengzhi Zhang, Kai Chen, and XiaoFeng Wang

Presenter: Munim Hasan Wasi
2024-11-13

KAIST



VOICE INPUT IS NEAR UBIQUITOUS

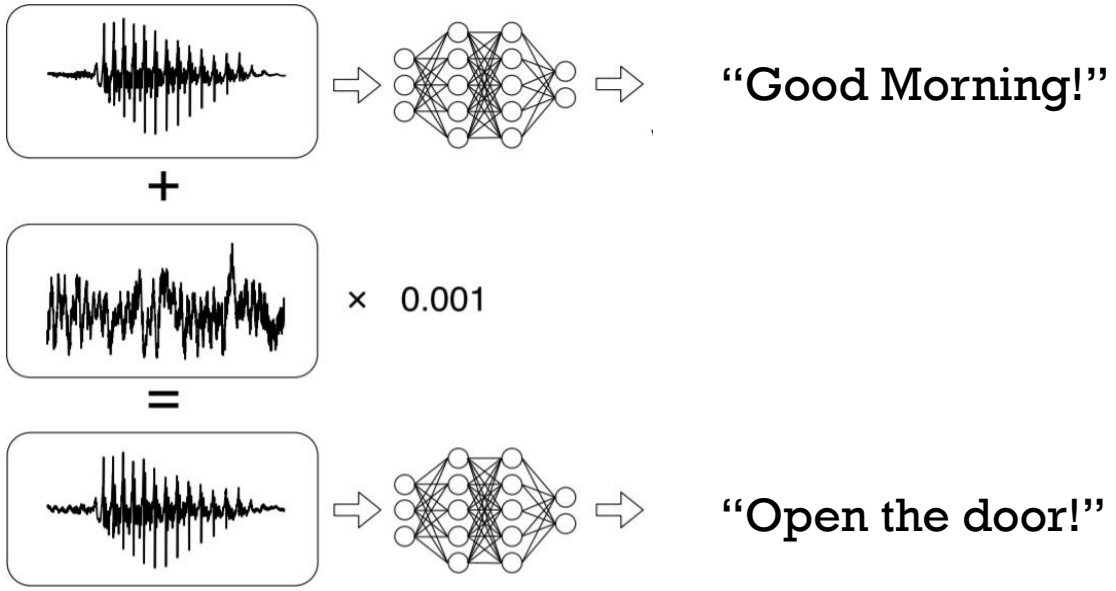
Cell phones, smart home devices, computers…

ATTACKS FOR SPEECH SYSTEMS

Traditional Attacks

Adversarial Attacks

- Audio adversarial examples(AEs)



INTRODUCTION

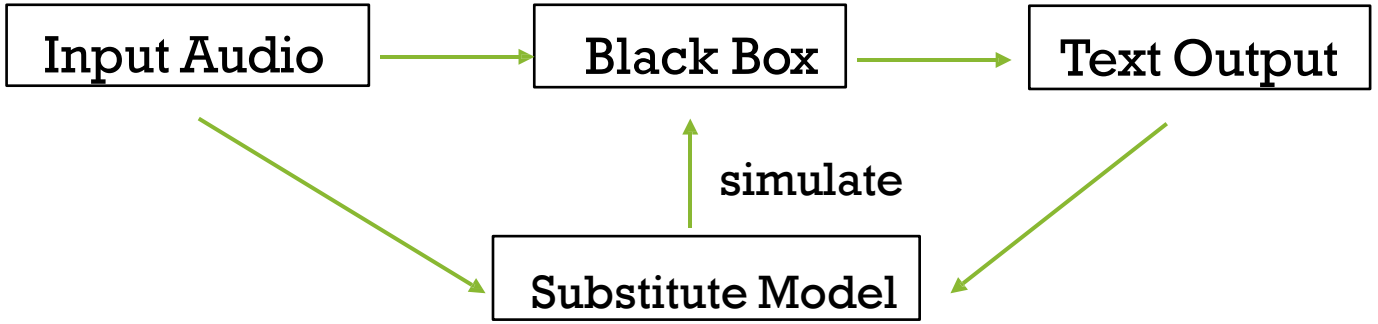
- Can we propose a method to generate audio AEs for several commercial black box speech recognition devices?
  - General
  - Practical
  - Stealthy
  - Automatic

APPROACH OVERVIEW

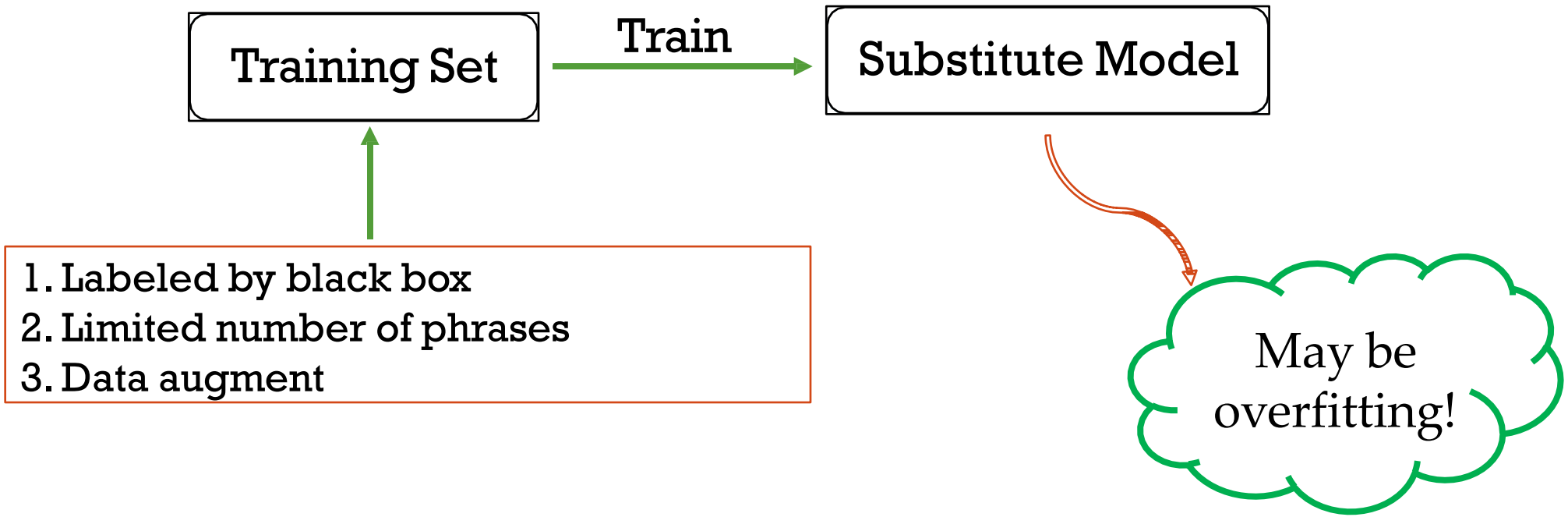
 Black Box Platforms	Commercial Speech APIs (Google API) Commercial Speech Devices (Amazon Echo)
 Transferability Based Approach (TBA)	Gradient descent/white box
 Alternate Models based Generation Approach (AGA)	Local Large Base Model and Substitute Model Ensemble AE generation

SUBSTITUTE MODEL

- Substitute Model
  - Local trained model to simulate black box model
  - Audio corpus labelled by target black box model

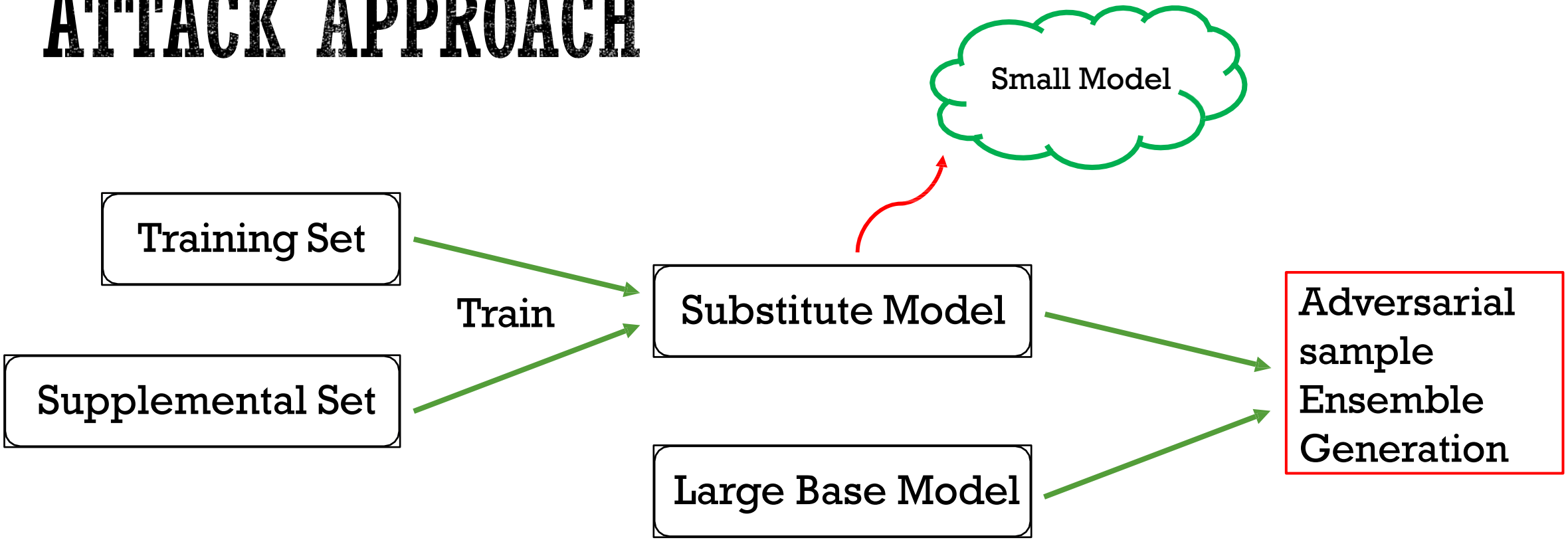


ATTACK APPROACH



- Training set augment: add noise/change voice speed
- Limited data set → potential model overfitting problem

ATTACK APPROACH



- Supplemental set: open-sourced voice data for training
- Large base model: already trained ASR model (e.g. Kaldi Aspire model)

GENERATE ADVERSARIAL SAMPLE

- Momentum based Iterative Fast Gradient Method (MI-FGM)
- Song/music as carrier
- Pdf-id sequence matching method (CommanderSong Yuan et al.)

Momentum

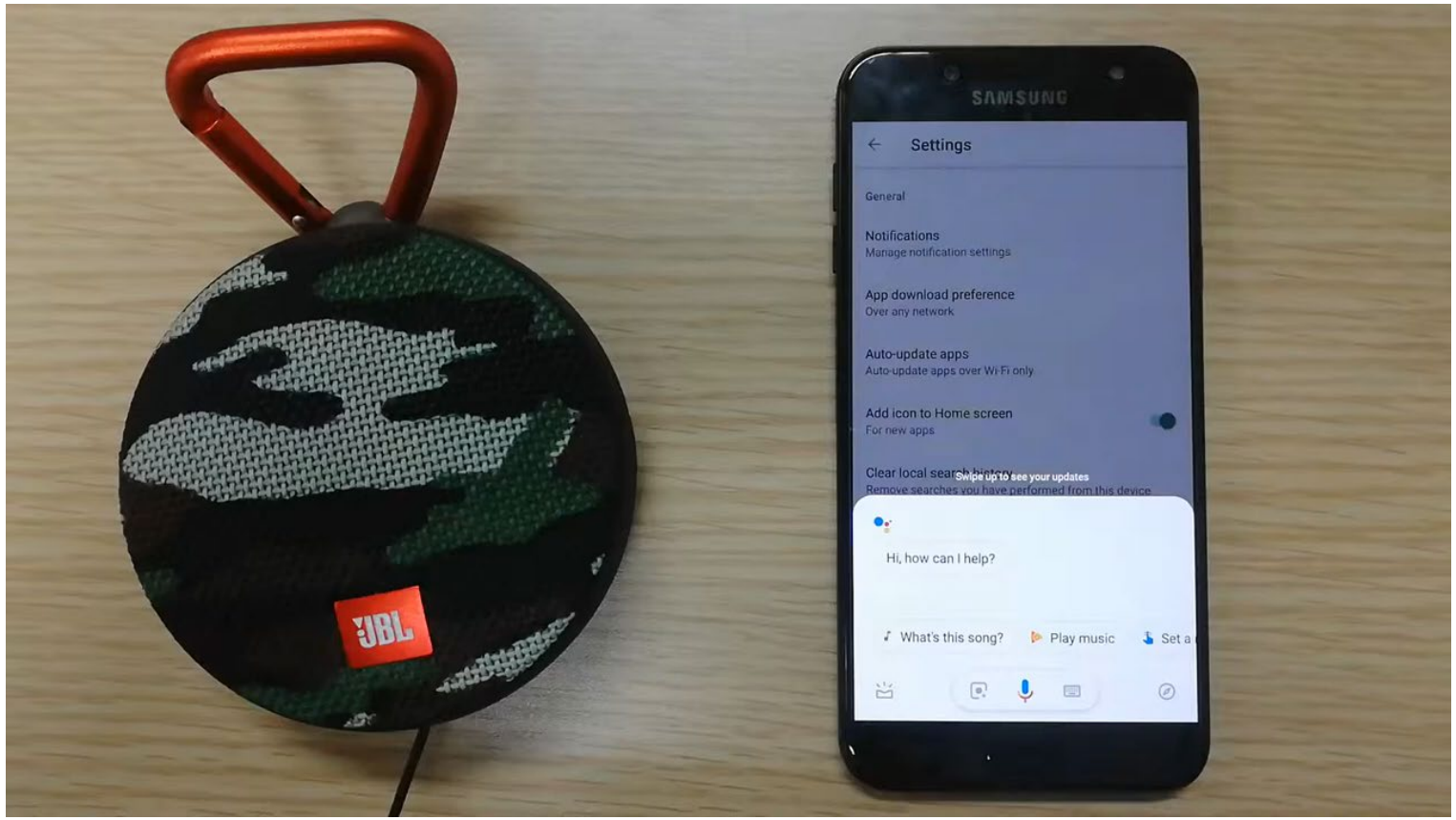
$$g_{t+1} = \mu \cdot g_t + \frac{J(x_t^*, y)}{\|\nabla_x J(x_t^*, y)\|_1}$$

$$x_{t+1}^* = x_t^* + Clip_{\varepsilon}(\alpha \cdot g_{t+1})$$

KAIST

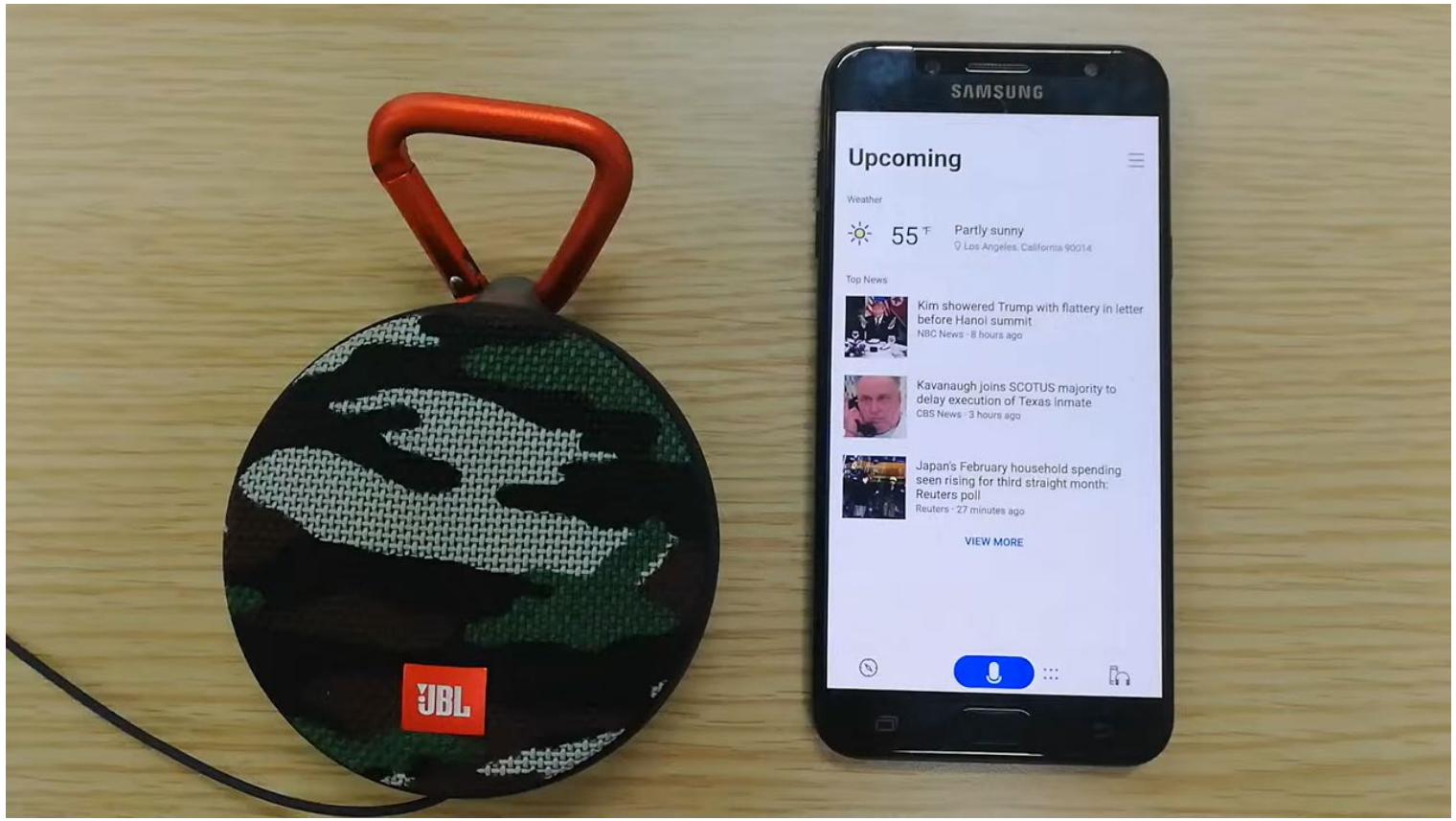
ADVERSARIAL EXAMPLE – AMAZON ECHO

ADVERSARIAL EXAMPLE – GOOGLE ASSISTANT

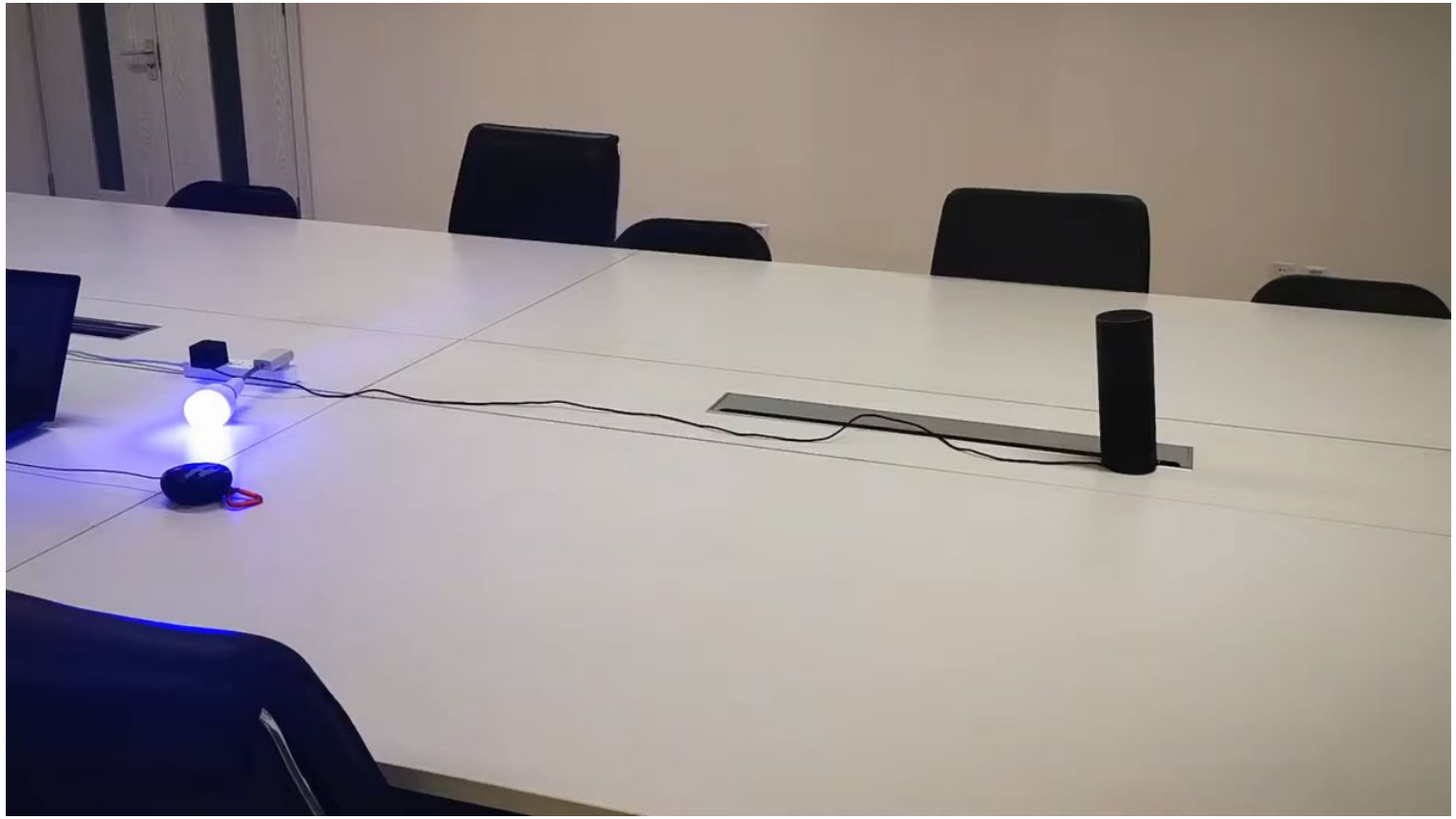


ADVERSARIAL EXAMPLE - GOOGLE HOME MINI

ADVERSARIAL EXAMPLE - CORTANA

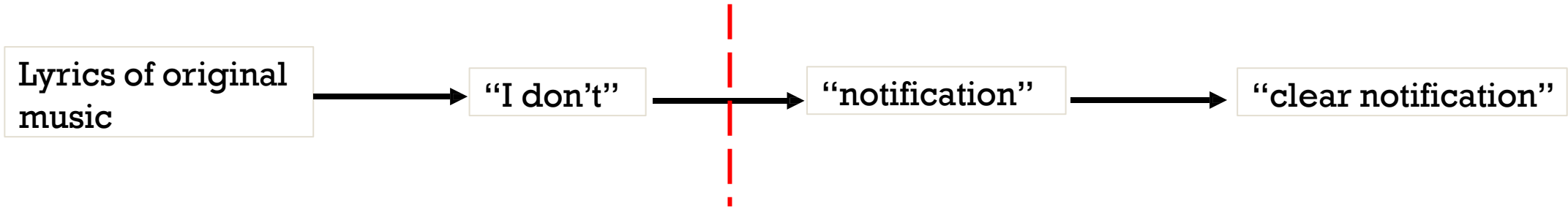


ADVERSARIAL EXAMPLE - ATTACK IN REALISTIC DISTANCE (UP TO 2 METERS)



UNDERSTAND THE ATTACK

- Base Model
  - Generate features in a coarse-grained manner
- Substitute Model
  - Fine-tunes the features to attack black box model
- Case example, target phrase "clear notification"



RESULTS EVALUATIONS

- Device Attack Effectiveness
  - Target devices: Google Assistant (cell phone), Google Home, Microsoft Cortana (cell phone), Amazon Echo, IBM Wav-To-Air
  - Effective distance: 5 cm- 200 cm

Black -box	Google		Microsoft	Amazon	IBM
	Assistant	Home	Cortana	Echo	WAA
TBA	4/10	4/10	2/10	0/10	3/10
AGA	10/10	9/10	10/10	10/10	10/10
SNR (dB)	9.03	8.81	10.55	12.10	7.86

Note: (1) "WAA" is used to represent "Wav-Air-API" attack. (2) The results were all based on the tests conducted in October 2019.

RESULTS EVALUATIONS

- Robustness test
  - Successful tests over total 30 tests
  - Same test environment (distance/volume/position)
  - 76% (38/50) of the commands: 10 successful cases over 30 total playing tests (1/3)

Target Model	Target Command	Success Rate
Google Home	Ok Google, call 911	25/30
Google Home	Ok Google, Turn on the Bluetooth	21/30
Amazon Echo	Echo, turn off the light	28/30
Microsoft Cortana	Hey Cortana, open the website	29/30

HUMAN PERCEPTION EVALUATIONS

- Human Perception Test
  - Amazon MTurk Survey
  - 84% of users think the samples sounds like normal speech or noise speech
  - only 1.4% of users could tell over 50% words in the target commands

EVALUATIONS OF OTHER APPROACHES

- Devil's Whisper vs Naïve command/music mixture?
  - Simple music + command combination using Adobe tools
  - Under similar success rate with Devil's Whisper
  - Only 6% users think the samples sounds like normal speech or noise speech

Command Voice + Original Song

VS

Devil's Whisper

RELATED RESEARCH
FOUNDATIONS AND EARLY ADVERSARIAL ATTACKS

- White-Box Attacks on ASR Systems
  - Carlini et al. (2016) attacked Mozilla's DeepSpeech by embedding adversarial audio that could be misinterpreted when model parameters were known.
  - Yuan et al. - CommanderSong (2018): demonstrated embedding adversarial commands within songs for ASR attacks in controlled, white-box setups.
- Initial Black-Box Exploration
  - Papernot et al. (2017) explored black-box attacks in image recognition by training substitute models, a foundational approach for later black-box methods.
  - Taori et al. (2019): attempted black-box attacks on commercial ASR systems but found transferability and effectiveness to be limited.

RELATED RESEARCH
POST-DEVIL'S WHISPER

- Advancements in Black-Box Adversarial Attacks
  - Zheng et al. (2021): Minimal-information black-box attacks via Occam and NI-Occam, achieving high success rates with minimal interaction.
- Security of ASR Systems: Taxonomy and Modular Analysis
  - Chen et al. (2022): Introduced a systematic framework for understanding ASR attacks and defenses, aligning with image recognition security methods and enhancing transferability insights.

CONCLUSIONS

- First adversarial attack against commercial speech devices
  - Targeting commercial and home-use devices like Echo or Google Home
  - Overcome black box challenges (No inside information/parameters)
  - Towards more complicated speech system rather than image system
- Novel model ensemble AE generation approach
- Effective and Stealthy
  - Show physical robustness in real world scenario
  - Human survey

THANK YOU

KAIST

GOOD QUESTIONS

- Is this attack scalable in the real scenario?
- Given that Devil's Whisper AEs are consistently successful, how might future model updates or technological advancements impact AE recognition and defense strategies?
- Given the high computational requirements for attacking large language models (LLMs), what practical constraints or resources would be necessary to make black-box attacks on LLMs feasible for attackers without extensive resources?
- Could integrating sensor fusion mechanisms, which use input from multiple microphones or environmental sensors, be an effective way to flag suspicious audio patterns? Would this add complexity or latency to the ASR process?
- As this paper focused on the signal synthesis part, I also feel like the prompts are quite important as well. For example if the ASRs now are patched with those AEs, they might not have been strong against prompts which can indirectly leak privacy information. What is your thought on this?

BEST QUESTIONS

- Hyunmin Ju
  - Could this approach be adapted to fool real-time ASR models, such as those used in security and surveillance?
- Boris Testud
  - When I looked at the videos showcasing the adversarial examples, I thought that in some examples, the commands were quite discernable. What do you think?
- Changgun Kang
  - Rather than injecting a specific command into the ASR system, can we perform a DoS attack using an adversarial example on the ASR system?