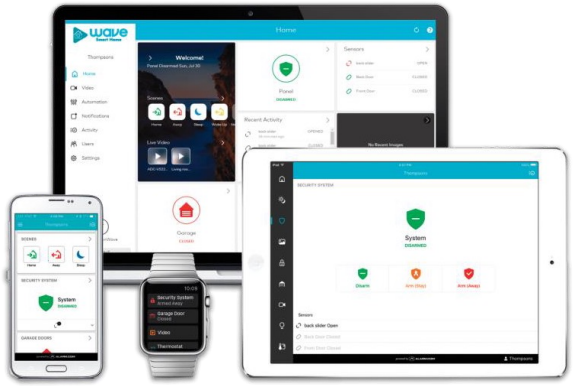
Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices

Yuxuan Chen[1,2,3], Xuejing Yuan[1,2], Jiangshan Zhang[1,2], Yue Zhao[1,2], Shengzhi Zhang[4], Kai Chen[1,2], and XiaoFeng Wang[5]

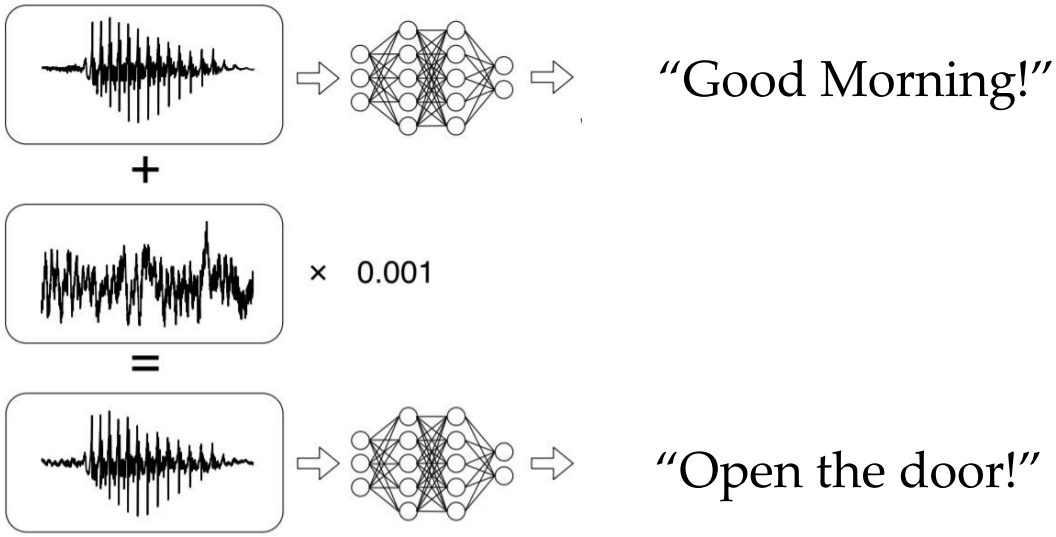
1. SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China
2. School of Cyber Security, University of Chinese Academy of Sciences, China
3. Department of Computer Engineering and Sciences, Florida Institute of Technology, USA
4. Department of Computer Science, Metropolitan College, Boston University, USA
5. School of Informatics and Computing, Indiana University Bloomington, USA

- Voice Input is near ubiquitous
  - Cell phones, smart home devices, computers…



Attacks for Speech Systems

- Traditional Attacks
- Adversarial Attacks
  - Audio adversarial examples(AEs)



Motivation

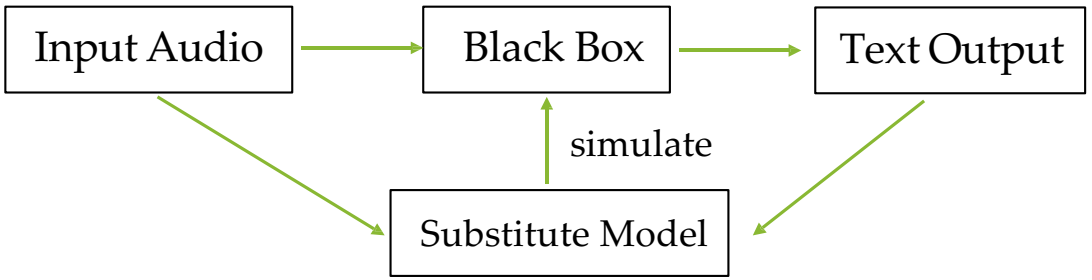
- Can we propose a method to generate audio AEs for several commercial black box speech recognition devices?
  - General
  - Practical
  - Stealthy
  - Automatic

Approach Overview

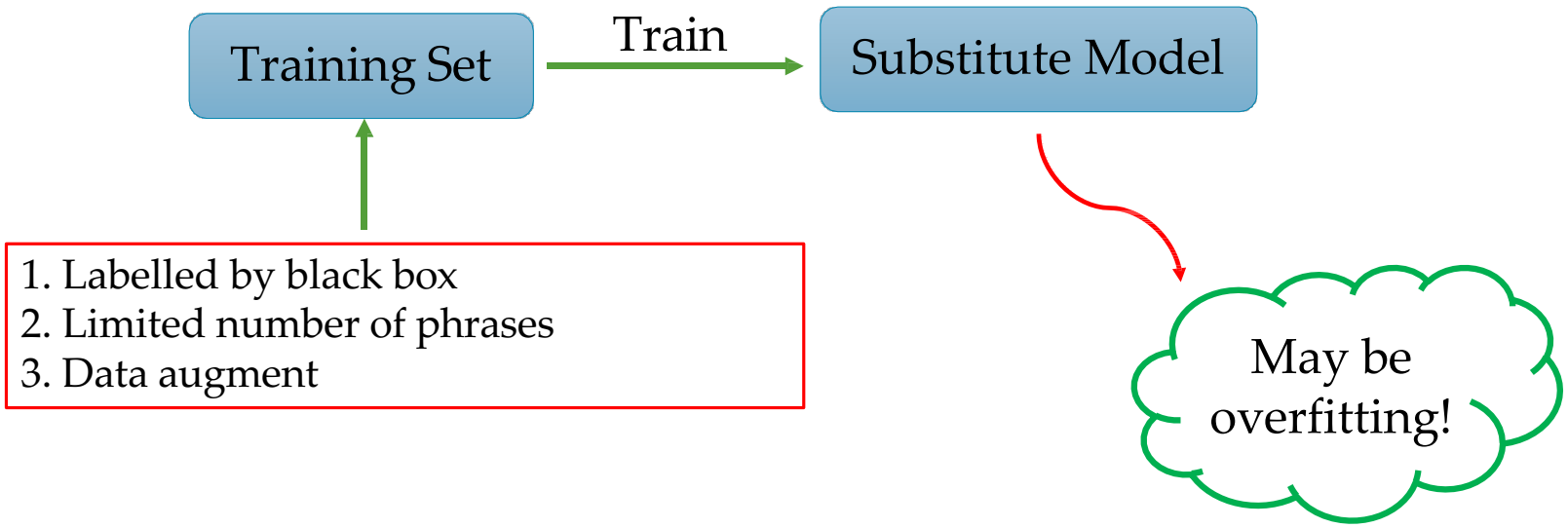
- Black Box Platforms
  - Commercial Speech APIs (Google API)
  - Commercial Speech Devices (Amazon Echo)
- Transferability Based Approach (TBA)
  - Gradient descent/white box
- Alternate Models based Generation Approach (AGA)
  - Local Large Base Model and Substitute Model
  - Ensemble AE generation

Substitute Model

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

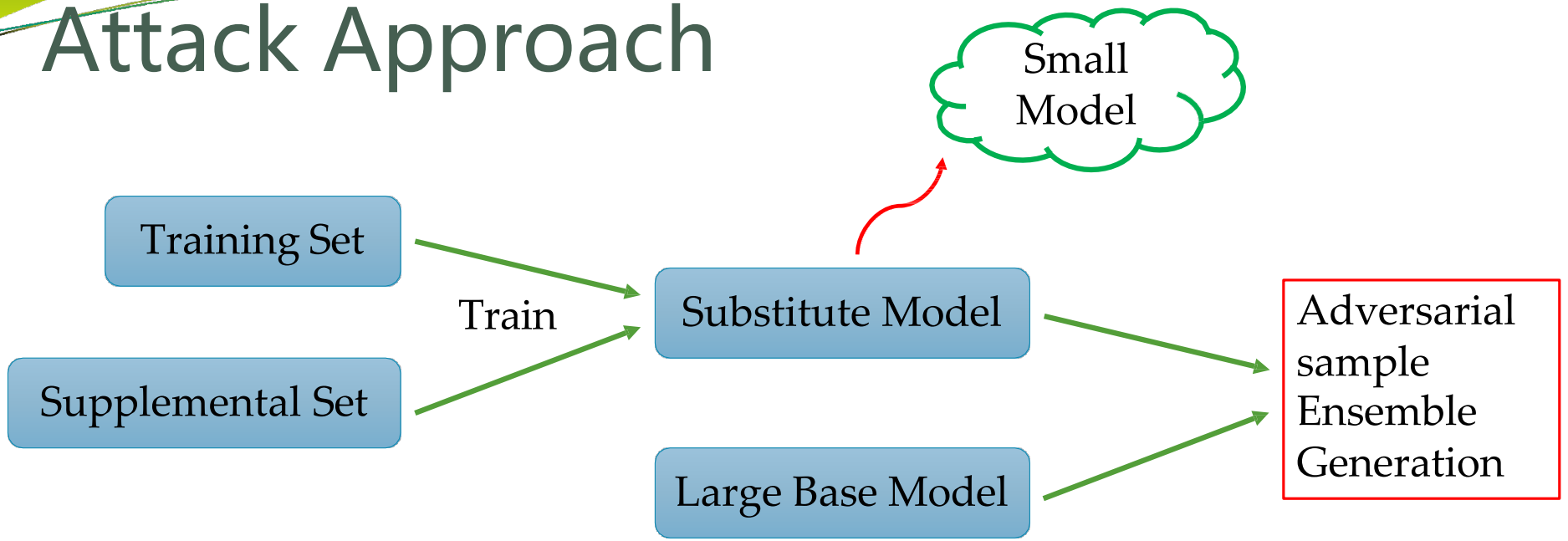


Attack Approach



- 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.)

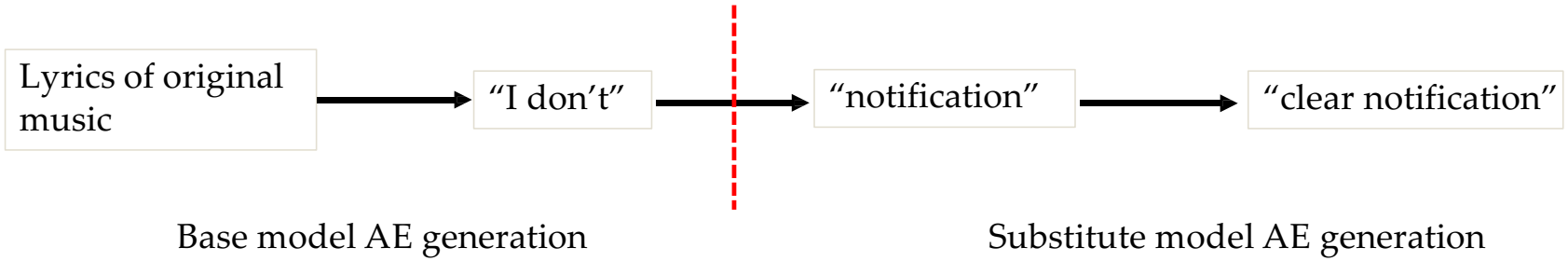
Momentum

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

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

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

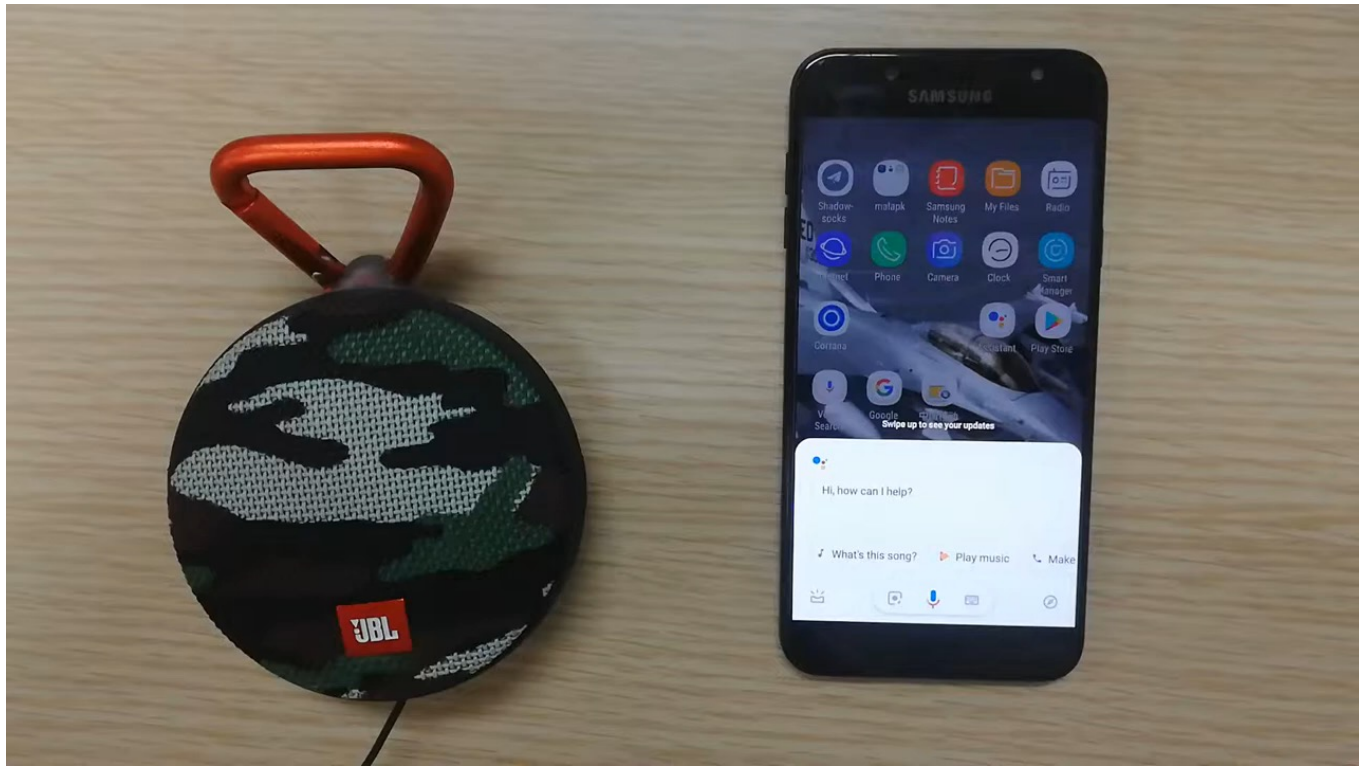
Note: (1) "WAA" is used to represent "Wav-Air-API" attack. (2) The results were all based on the tests conducted in October 2019.

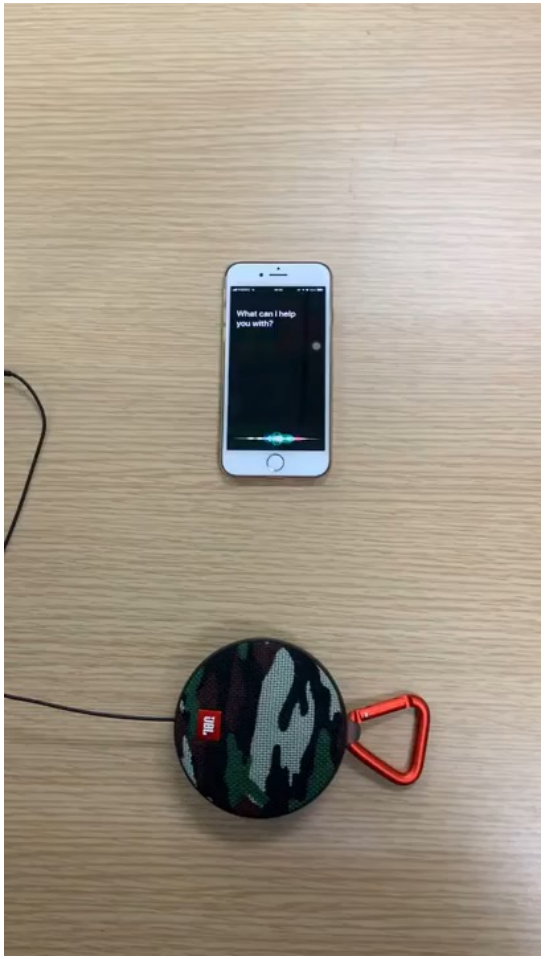
Results Evaluations

- 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

Demo Videos


JBL
SAMSUNG
Hi, how can I help?
What's this song?
Play music
Make


What can I help you with?
JBL

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

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

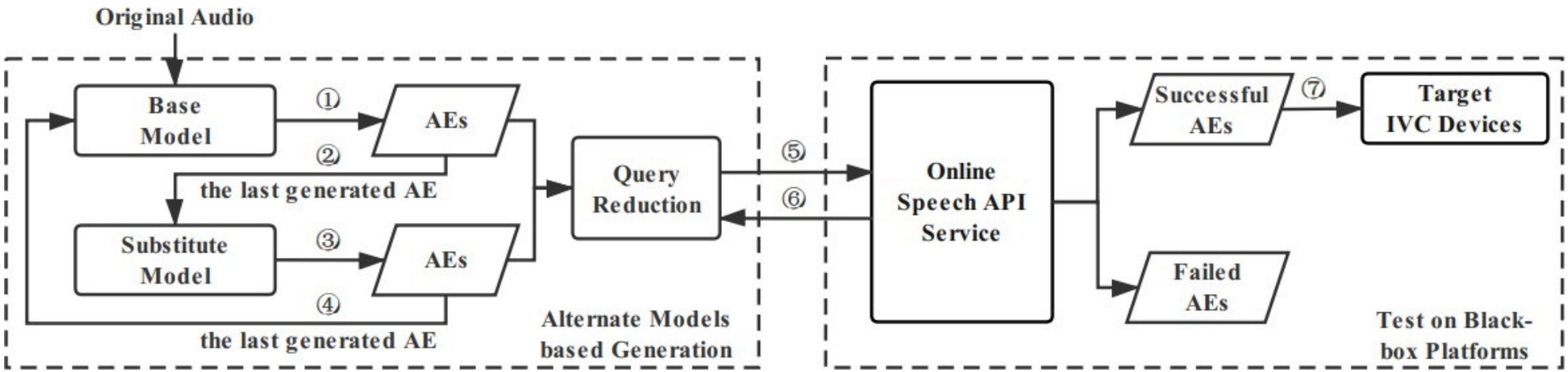
Limitations

- AE generation for only specific commands
  - Not general approach
- Cannot overcome fundamental black box challenge
  - Victim can update or modify the model

Thank you!

Back up slides

Attack Workflow



- AE Generation
  - Generate AEs with base model and substitute model.
  - Efficient query of the black-box API

Original Song Selection

- 10 songs from soft and classic music of CommanderSong
- Run Devil's Whisper approaches on 10 songs
  - "Okay Google, navigate to my home" – Google API
  - "Hey Cortana, turn off the bedroom light" – Bing API
- 6 songs were selected

Implementations

- Target API systems and speech devices
  - Google Speech API, Bing API, Amazon Transcribe API, IBM Speech API
  - Google Home/Assistant, Cortana App, Amazon Echo, IBM Wav-to-Air
- Phrase Selection
  - Commonly used commands “turn off light”, “call 911”, “open the door” etc.
  - 10 commands for each target

Okay Google, take a picture.
Okay Google, navigate to my home.
Okay Google, turn off the light.
Okay Google, play music.
Hey Cortana, open the website.
Hey Cortana, make it warmer.
Echo, turn off the computer.
Echo, call my wife.

Evaluations of Other Approaches

- Train the substitute model with a larger training set
  - Larger training data -> better approximation -> better results ?
  - New dataset: 5.17 times larger

Command	G1	G2	G3
Okay Google, play music.	X	X	X
Okay Google, take a picture.	X	X	X
Okay Google, turn off the light.	✓	✓	X
Okay Google, navigate to my home.	X	X	X

  - G1: Google Command API, G2: Google Assistant, G3: Google Home

Evaluations of Other Approaches

- Ensemble AE generation without model approximation
  - Kaldi Aspire model and Mini LibriSpeech model: ensemble AE generation
  - No model simulation (No substitute model training)

Command	G1	G2	G3
Okay Google, call 911.	X	X	X
Okay Google, take a picture.	X	X	X
Okay Google, set an alarm on 8 am.	X	X	X
Okay Google, navigate to my home.	✓	✓	X

  - G1: Google Command API, G2: Google Assistant, G3: Google Home