Metamorph:
Injecting Inaudible Commands into
Over-the-air Voice Controlled Systems

Tao Chen¹  Longfei Shangguan²  Zhenjiang Li¹  Kyle Jamieson³

¹City University of Hong Kong, ²Microsoft, ³Princeton University
Voice Assistants in Smart Home
Voice Assistants in Smart Home
Voice Assistants in Smart Home

111.8 million people in U.S. use voice assistants and related services!

https://www.emarketer.com/content/us-voice-assistant-users-2019
Are they safe enough?
How to attack the voice assistant?

Speech Recognition Models (SR)

Neural networks
How to attack the voice assistant?

Audio Clip: $I$

$SR(I) \rightarrow T: \text{"this is for you"}$
How to attack the voice assistant?

Audio Clip: $I$

Adversarial Example: $I + \delta$

Perturbation: $\delta$

$T : "this is for you"$

$T' : "open the door"$

$SR(I)$
How to attack the voice assistant?

Audio Clip: $I$

Adversarial Example: $I + \delta$

Perturbation: $\delta$

$x0.01$

$T : "this is for you"$

$T' : "open the door"$

$SR(I)$

minimize $dB_I(\delta)$,

such that $SR(I) = T$,

$SR(I + \delta) = T'$

Nicholas Carlini et al. Audio Adversarial Examples, Deep Learning and Security Workshop, 2018
How to attack the voice assistant?

Audio Adversarial Attack

Audio Clip: $I$

Adversarial Example: $I + \delta$

$SR(I)$

$T$ : “This is for you”

$T'$ : “Open the door”

$x0.01$

Perturbation: $\delta$

minimize $dB_I(\delta)$,

such that $SR(I) = T$,

$SR(I + \delta) = T'$

Nicholas Carlini et al. Audio Adversarial Examples, Deep Learning and Security Workshop, 2018
Is it a real threat? Yes!
Adversarial Example

But, failed Over-the-air!
Challenge

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity
Challenge

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity

\[ SR(I + \delta) \ vs \ SR(H(I + \delta)) \]

H is unknown in advance!
Understand Over-the-air Attack

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity
Attenuation
Attenuation
Attenuation

“Open the door”
Attenuation

“Open the door”

No frequency-selectivity, doesn’t matter at all!
Understand Over-the-air Attack

Channel Effect

Multi-path

Attenuation

Noise

Hardware Heterogeneity
Hardware Heterogeneity

Anechoic Chamber Testing

Transmitter

Receiver

Anechoic Materials
Hardware Heterogeneity

Transmitter

Anechoic Chamber Testing

Received Power (dB)

Frequency (kHz)

Rx: Nexus 5X
Hardware Heterogeneity

Transmitter

Receiver

Anechoic Chamber Testing

Received Power (dB)

Frequency (kHz)

Rx: Nexus 5X
Rx: HTC A9w
Rx: iPhone 8
Rx: SAMSUNG S7
Hardware Heterogeneity

Anechoic Chamber Testing

Transmitter

Anechoic Materials

Receiver

Not strong, device’s inherent feature, compensable!
Hardware Heterogeneity

Character Successful Rate (CSR):

- **Nexus**: CSR = 0.23
- **HTC**: CSR = 0.2
- **iPhone**: CSR = 0.35
- **SAMSUNG**: CSR = 0.25

**0.5 m, chamber**

**0.5 m, office**

**8 m, office**

**Static, predictable and compensable!**
Understand Over-the-air Attack

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity
Multi-path

HIVI M200MK3 Speaker
Over-the-air Channel
SAMSUNG S7

HIVI M200MK3 Speaker
Over-the-air Channel
SAMSUNG S7

HIVI M200MK3 Speaker
Over-the-air Channel
SAMSUNG S7

Ruler
Multi-path: Near range

Tx to Rx: From 0.5m to 8m
Multi-path: Near range

Office  Corridor  Home

Tx to Rx: From 0.5m to 8m

- LOS path
- Reflections
- Reflection 1
- Reflection 2
- Superimposed signal

CIR

Time (s)
Multi-path: Near range

Tx to Rx: From 0.5m to 8m

Office

Corridor

Home

LOS path

reflections

|CIR|

Time (s)

[Graph showing CIR vs. time with different environments labeled: 0.5 m, Corridor, 0.5 m, Office, 0.5 m, Home]

Reflection 1

Reflection 2

Q

 LOS path

Superimposed signal

I

|CSI| (dB)

Frequency (kHz)

[Graph showing CSI vs. frequency with different environments labeled: 0.5 m, Corridor, 0.5 m, Office, 0.5 m, Home]
Multi-path: Near range

Office | Corridor | Home

Tx to Rx: From 0.5m to 8m

Also not strong and similar!
Multi-path: Long range

Office  Corridor  Home

Tx to Rx: From 0.5m to 8m
Multi-path: Long range

Office | Corridor | Home

Tx to Rx: From 0.5m to 8m

LOS path
reflections

CIR

Time (s)

0 0.02 0.04

8 m, Corridor
8 m, Office

LOS path
Superimposed signal

Reflection2
Reflection1
Multi-path: Long range

Office  Corridor  Home

Tx to Rx: From 0.5m to 8m

LOS path

reflections

CIR

Time (s)

0 0.02 0 0.02 0.04

8 m, Corridor
8 m, Office

LOS path

Superimposed signal

Reflection2

Reflection1

CSI (dB)

Frequency (kHz)

0 2 4 6 8

0 -20 -40

8 m, Corridor
8 m, Office
Multi-path: Long range

Tx to Rx: From 0.5m to 8m

Office

Corridor

Home

Stronger and unpredictable!
Multi-path: Long range

Character Successful Rate (CSR):

![Bar chart showing CSR for different devices and distances](chart.png)

- **0.5 m, chamber**: Nexus 0.23, HTC 0.2, iPhone 0.35, SAMSUNG 0.25
- **0.5 m, office**: Nexus 0.1, HTC 0.1, iPhone 0.14, SAMSUNG 0.13
- **8 m, office**: Nexus 0, HTC 0, iPhone 0, SAMSUNG 0

Highly unpredictable!
Design Inspiration

\[ I + \delta \rightarrow SR(H(I + \delta)) \]

= “Open the door”
Design Inspiration

I + δ

Unknown, but share similarity!

SR(I + δ))

= “Open the door”
Design Inspiration

I + δ

Unknown, but share similarity!

H: public acoustic CIR datasets

SR(H(I + δ))

= “Open the door”
Design Inspiration

\[
\begin{align*}
\text{arg min}_\delta & \quad \alpha \cdot dB_I(\delta) + \frac{1}{M} \sum_i \text{Loss}(SR(H_i(I + \delta)), T') \\
\text{Unknown, but share similarity!} & \\
\text{H: public acoustic CIR datasets} & \\
\end{align*}
\]

= “Open the door”
Design Inspiration

Transcript and Character Successful Rate:

\[
\arg \min_{\delta} \alpha \cdot dB_I(\delta) + \frac{1}{M} \sum_i \text{Loss}(SR(H_i(I + \delta)), T')
\]
Design Inspiration

I + δ → \( SR(H(I + δ)) \)

Domain (environment-specific) information dominates!

\( SR(H(I + δ)) \)

H: public acoustic CIR datasets

“Open the door”
Metamorph: Meta-Enha

Adversarial Example Generator

\[ \arg \min_\delta \alpha \cdot dB_I(\delta) + \frac{1}{M} \sum_i Loss(SR(H_i(I + \delta)), T') - \beta \cdot L_d \]
Metamorph: Meta-Qual

- Acoustic Graffiti:

\[ \text{distance}(\delta, \hat{N}) \]

- Reducing Perturbation’s Coverage:

\( L1/L2 \) regularization

![Graphs](attachment:image.png)
Evaluation: Audio Quality

• Examples

Classical music

Original: [no transcription]

Meta-Enha: “hello world”

Meta-Qual: “hello world”

Human speech

Original: “your son went to serve at a distant place and became a centurion”

Meta-Enha: “open the door”

Meta-Qual: “open the door”
Evaluation: Attack Successful Rate

- Attack Target: “DeepSpeech” (White-Box)
Evaluation: Attack Successful Rate

- Line-of-Sight (LOS) Attack

Meta-Enha: > 90% attack successful rate
Evaluation: Attack Successful Rate

- No-Line-of-Sight (NLOS) Attack

Meta-Enha: over 85% attack successful rate across 11/20 NLOS location!
Conclusion

1. Investigate over-the-air audio adversarial attacks systematically.
2. Propose a “generate-and-clean” two-phase design and improve the audio quality.
3. Develop a prototype and conduct extensive evaluations.

Visit acoustic-metamorph-system.github.io for more information!