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?

Neural networks

Speech Recognition Models (SR)
How to attack the voice assistant?

Audio Clip: $I$

$T: \text{"this is for you"}$

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

Audio Clip: $I$

Adversarial Example: $I + \delta$

Perturbation: $\delta$

$T : \text{"this is for you"}$

$T' : \text{"open the door"}$
How to attack the voice assistant?

Audio Clip: $I$

Audio Clip: $I + \delta$

Perturbation: $\delta$

**Adversarial Example:**

$SR(I)$

$x0.01$

$T: \text{“this is for you”}$

$T': \text{“open the door”}$

minimize $dB_1(\delta)$,

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: $I + \delta$

$T$ : “This is for you”

$T'$ : “Open the door”

$x_{0.01}$

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
Adversarial Example

But, failed Over-the-air!
Challenge

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity
Challenge

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity

\[ SR(I + \delta) \text{ vs } SR(H(I + \delta)) \]

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

Channel Effect

Multi-path

Attenuation

Hardware Heterogeneity
Attenuation
Attenuation

![Diagram showing sound waves attenuated over frequency](image)
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

Hardware Heterogeneity

Noise
Hardware Heterogeneity

Anechoic Chamber Testing
Hardware Heterogeneity

Anechoic Chamber Testing

Transmitter

Anechoic Materials

Receiver

Received Power (dB)

Frequency (kHz)

Rx: Nexus 5X

Anechoic Chamber Testing
Hardware Heterogeneity

Anechoic Chamber Testing

Transmitter

Anechoic Materials

Receiver

Received Power (dB)

Frequency (kHz)

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

Transmitter

Receiver

Anechoic Chamber Testing

Anechoic Materials

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

Character Successful Rate (CSR):

<table>
<thead>
<tr>
<th>Device</th>
<th>0.5 m, chamber</th>
<th>0.5 m, office</th>
<th>8 m, office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nexus</td>
<td>0.23</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>HTC</td>
<td>0.2</td>
<td>0.14</td>
<td>0</td>
</tr>
<tr>
<td>iPhone</td>
<td>0.35</td>
<td>0.13</td>
<td>0</td>
</tr>
<tr>
<td>SAMSUNG</td>
<td>0.25</td>
<td>0.13</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Channel Effect

Multi-path

Hardware Heterogeneity

Attenuation
Multi-path

HIVI M200MK3 Speaker

Over-the-air Channel

SAMSUNG S7

Ruler

HIVI M200MK3 Speaker

Over-the-air Channel

SAMSUNG S7

Ruler

HIVI M200MK3 Speaker

Over-the-air Channel

SAMSUNG S7
Multi-path: Near range

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

Tx to Rx: From 0.5m to 8m

Office  Corridor  Home

LOS path reflections

CIR

Time (s)

0 0.02 0.04

0.2

0

-0.2

0.5 m, Corridor
0.5 m, Office
0.5 m, Home

LOS path
Superimposed signal
Reflection2
Reflection1
Multi-path: Near range

Office  Corridor  Home

Tx to Rx: From 0.5m to 8m

LOS path
reflections

CIR
0 0.2
-0.2

Time (s)
0 0.02 0.04

0.5 m, Corridor
0.5 m, Office
0.5 m, Home

|CSI| (dB)
0 -20
-40

0 2 4 6 8

Frequency (kHz)

LOS path

Superimposed signal

Reflection1

Reflection2
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

Tx to Rx: From 0.5m to 8m

Office  Corridor  Home

LOS path
reflections

Q
I

LOS path
Superimposed signal

Reflection2
Reflection1

CIR

Time (s)

8 m, Corridor
8 m, Office

I

16
Multi-path: Long range

Tx to Rx: From 0.5m to 8m

Office  Corridor  Home

LOS path

Reflection1

Reflection2

Superimposed signal

CIR

CSI (dB)

Time (s)

Frequency (kHz)

8 m, Corridor

8 m, Office
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

CSI (dB)

0 -20 -40

8 m, Corridor

8 m, Office

Frequency (kHz)

0 2 4 6 8

Superimposed signal

Reflection1

Reflection2

LOS path

Stronger and unpredictable!
Multi-path: Long range

Character Successful Rate (CSR):

<table>
<thead>
<tr>
<th></th>
<th>0.5 m, chamber</th>
<th>0.5 m, office</th>
<th>8 m, office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nexus</td>
<td>.23</td>
<td>.1</td>
<td>0</td>
</tr>
<tr>
<td>HTC</td>
<td>.2</td>
<td>.1</td>
<td>0</td>
</tr>
<tr>
<td>iPhone</td>
<td>.35</td>
<td>.14</td>
<td>0</td>
</tr>
<tr>
<td>SAMSUNG</td>
<td>.25</td>
<td>.13</td>
<td>0</td>
</tr>
</tbody>
</table>

Highly unpredictable!
Design Inspiration

$I + \delta \rightarrow \text{SR}(H(I + \delta))$

= “Open the door”
Design Inspiration

$I + \delta$

Unknown, but share similarity!

= “Open the door”
Design Inspiration

$I + \delta$

Unknown, but share similarity!

H: public acoustic CIR datasets

$SR(H(I + \delta))$

$SR(H(I + \delta))$

= “Open the door”
Design Inspiration

Unknown, but share similarity!

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

H: public acoustic CIR datasets

"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

Domain (environment-specific) information dominates!

$S_R(H(I + \delta))$

$SR(H(I + \delta))$

$H$: public acoustic CIR datasets

"Open the door"
Adversarial Example Generator

Speech Recognition Models

Domain Discriminator

Clean domain information

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

- Acoustic Graffiti:

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

- Reducing Perturbation’s Coverage:

\[ L1/L2 \text{ regularization} \]

(a) Perturbation: Enhanced

(b) Perturbation: Quality
Evaluation: Audio Quality

• Examples

Classical music


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!