Inaudible Voice Commands:
The Long-Range Attack and Defense

Nirupam Roy
Sheng Shen
Haitham Hassanieh
Romit Roy Choudhury

University of Illinois at Urbana-Champaign
50 million voice assistants are sold in US
Inaudible Acoustics

Normal Sound
(< 24 kHz)

Ultrasound
(> 25 kHz)

“Inaudible Acoustics”
(> 25 kHz)

“Alexa, open the garage door!”
Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

MobiSys’17 (Best Paper)       CCS’17       arXiv
Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

Today’s Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?
Today’s Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?
Microphone frequency spectrum

Amplitude

10k  20k  30k  40k  50k  60k  70k  80k  90k  100k

Diaphragm → Amplifier → Filter → ADC → Digitized signal
Amplifier
Filter
ADC

Microphone
filter

Amplitude

10k 20k 30k 40k 50k 60k 70k 80k 90k 100k

Microphone frequency spectrum

Diaphragm → Air Vibration → Electric Voltage

Filter → ADC
\[ V_{out} = a_1 V_{in} \]
\[ V_{out} = a_1 V_{in} \]

The output voltage \( V_{out} \) is a linear combination of the input voltage \( V_{in} \) and higher-order terms:

\[ V_{out} = a_1 V_{in} + a_2 V_{in}^2 + \ldots \]

The graph on the right shows the nonlinear behavior of the system, with the input frequency on the x-axis and the output on the y-axis. The diagram on the bottom left illustrates the process of sound amplification, from the diaphragm to the amplifier, with a nonlinear response curve indicating the amplification of the input signal.
\[ V_{out} = a_1 V_{in} \]

\[ V_{out} = a_1 V_{in} + a_2 V_{in}^2 + \ldots \]
\[ V_{out} = a_1 V_{in} + a_2 V_{in}^2 \]

\[
\left( \sin F_1 + \sin F_2 \right)^2 = -\cos 2F_1 \\
- \cos 2F_2 \\
- \cos (F_1 + F_2) \\
+ \cos (F_1 - F_2)
\]
\[ V_{out} = a_1 V_{in} + a_2 V_{in}^2 \]

\[
(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1 + F_2) + \cos (F_1 - F_2)
\]
$V_{out} = a_1 V_{in} + a_2 V_{in}^2$

$(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1 + F_2) + \cos (F_1 - F_2)$
\[ V_{out} = a_1 V_{in} + a_2 V_{in}^2 \]

\[
(\sin F_1 + \sin F_2)^2 = -\cos 2F_1 - \cos 2F_2 - \cos (F_1 + F_2) + \cos (F_1 - F_2)
\]
The diagram illustrates the frequency spectrum and amplitude response of a microphone with a filter. It shows audible and inaudible frequency ranges, with specific notations for $(F_1 - F_2)$, $F_2$, and $F_1$. The microphone filter is depicted as affecting the amplitude at various frequencies.
V(t) = “Alexa, open the garage door!”
V(t) = “Alexa, open the garage door!”
3-5 ft
Can someone attack from a longer range?
Can someone attack from a longer range?
Can someone attack from a longer range?

High power makes ultrasonic speakers audible
Speakers have nonlinearity too!
Voice Command: $v(t)$
Speaker Nonlinearity

\[ V_{in} = v(t)\sin(\omega_1 t) \]

\[ a_1 V_{in} + a_2 V_{in}^2 \]
$V_{in} = v(t)\sin(\omega_1 t)$

Speaker Nonlinearity

$a_1 V_{in} + a_2 V_{in}^2$

$a_1 v(t)\sin(\omega_1 t)$
Amplitude

$V_{in} = v(t) \sin(\omega_1 t)$

Speaker Nonlinearity

$a_1 V_{in} + a_2 V_{in}^2$

$a_1 v(t) \sin(\omega_1 t) + a_2 \left( v^2(t) - v^2(t) \cos(2\omega_1 t) \right)$
$V_{in} = v(t) \sin(\omega_1 t)$

Speaker Nonlinearity:

$\alpha_1 V_{in} + \alpha_2 V_{in}^2$

General to all speakers!

Our Solution: "Leakage Optimization"
Speaker Nonlinearity $\rightarrow$ Audible Leakage

Speaker input

$V(f)$

Bandwidth: $B$

Speaker output

$V(-f) \ast V(f)$

Bandwidth: $B$
Speaker Nonlinearity $\rightarrow$ Audible Leakage

- Speaker input
- Speaker output

Amplitude

Frequency
Speaker Nonlinearity $\rightarrow$ Audible Leakage

Speaker input

Speaker output
Speaker Nonlinearity $\rightarrow$ Audible Leakage

Speaker input

Speaker output
Speaker Nonlinearity $\rightarrow$ Audible Leakage

Chopping compresses the leakage band
Maximize $\min_f[T(f) - L(f)]$

subject to $f_0 \leq f_1 \leq f_2 \leq \ldots \leq f_N$
Evaluation
Inaudible voice commands: Long range

Speaker array running leakage optimization

25 feet
Evaluation

Wake-word hit rate

![Graph showing wake-word hit rate vs. attack distance for different voice assistants: Alexa, S-Voice, Siri.](image)
Evaluation

Wake-word hit rate

Command detection accuracy
Evaluation

Maximum activation distance for different input power
Today’s Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?
Talk Outline

0. [BackDoor], [DolphinAttack], [Princeton Video]

Today’s Talk:

1. How to launch long-range (realistic) attacks?

2. How to defend against these attacks?
Core Question:

Is this a “non-linear signal” or normally recorded signal?

Voice signal: $\nu(t)$

Inaudible Voice Attack:

$$[\nu(t) \cdot \sin(\omega_1 t) + c \cdot \sin(\omega_1 t)]^2 = \nu(t) + c'\nu^2(t) + \cdots$$
Core Question:
Is this a “non-linear signal” or normally recorded signal

\[ [v(t) \cdot \sin(\omega_1 t) + c \cdot \sin(\omega_1 t)]^2 = v(t) + c'v^2(t) + \cdots \]
Difficult to decouple “voice signal” and “non-linear signal”

Human voice signals present opportunities …
Opportunity #1: Voice > 50 Hz

Human Voice

$v(t)$

Amplitude

Frequency

$v^2(t)$

$f$

$2f$

$3f$

$4f$
Opportunity #1: Voice > 50 Hz

Human Voice

Energy at sub-50Hz band
Opportunity #2: Correlation

Energy variation in $v(t)$

Energy variation in $v^2(t)$

Correlation
Opportunity #3: Amplitude Skewness

Amplitude skew = $v(t) + v^2(t)$
5000 Test Cases

Amplitude skew

Correlation

Sub-50Hz power

Real voice

Attack voice
Overall Detection Accuracy

![Bar chart showing overall detection accuracy across different loudness levels (dbSPL). Accuracy is consistently high at 1.0 for loudness levels of 50, 60, 70, 80, and 90 dbSPL.]
To summarize...

Inaudible Acoustics (> 25 kHz): “Alexa, open the garage door!”

Ok
Ok
Ok