DolphinAttack: Inaudible Voice Commands

GUOMING ZHANG, CHEN YAN, XIAOYU JI, TIANCHEN ZHANG , TAIMIN ZHANG, WENYUAN XU

ZHEJIANG UNIVERSITY

Presented by: Oskars Dauksts
November 14, 2018

Overview

Overview

- Introduction
- Background and Threat model
- Attack Design
- Feasibility experiments across VCS
- Impact quantification
- Defenses
- Conclusion

Introduction

Introduction – Key Terms

- VCS – Voice Controllable Systems
- SR – Speech Recognition
  - Converts spoken words into machine-readable formats
  - Ex: Alexa, Siri, Google Now, etc.
- MEMS Microphones
  - Current standard of most mobile devices
- Amplitude Modulation
  - Technique that modulates voice commands into ultrasonic waves

Introduction

- Can a voice attack be inaudible to human, while still being audible to the device?

- Can injecting a sequence of inaudible voice commands lead to unnoticed security breaches?

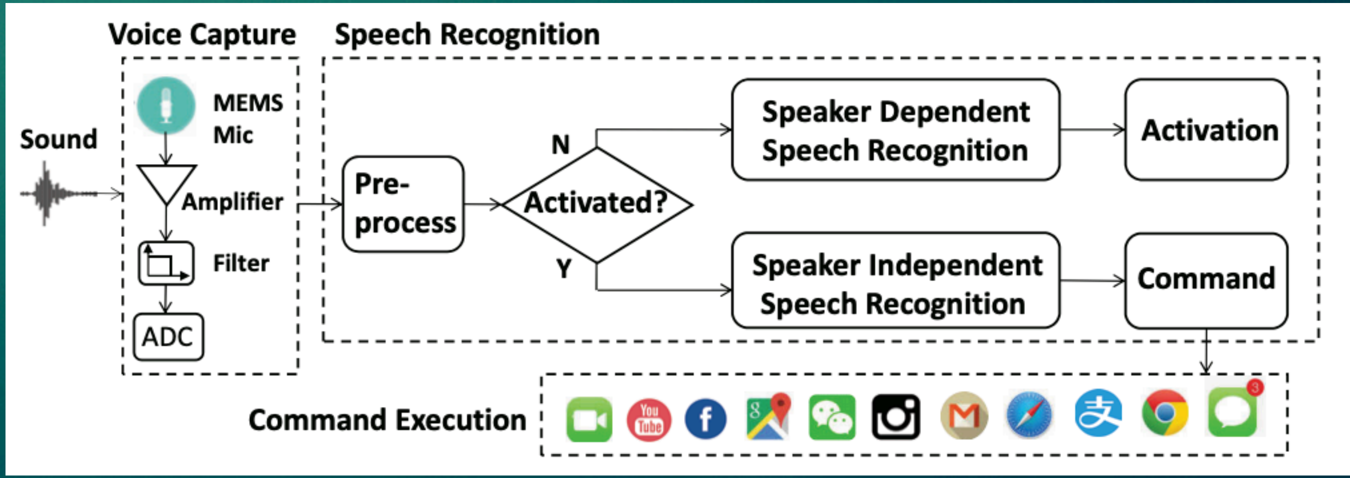
Introduction

- DolphinAttack approach
  - Exploit by utilizing inaudible ultrasound channel (*$F > 20kHz$), that can inject covert voice commands into state-of-art SR systems*
  - *Leverages MEMS microphones*
  - Includes
    - Visiting malicious website – drive-by-download
    - Spying – Listening on speaker
    - Injecting fake information – Sending messages,emcails
    - DOS – Denial of Service
    - Concealing attacks – Dimming screen, reducing volume

Background

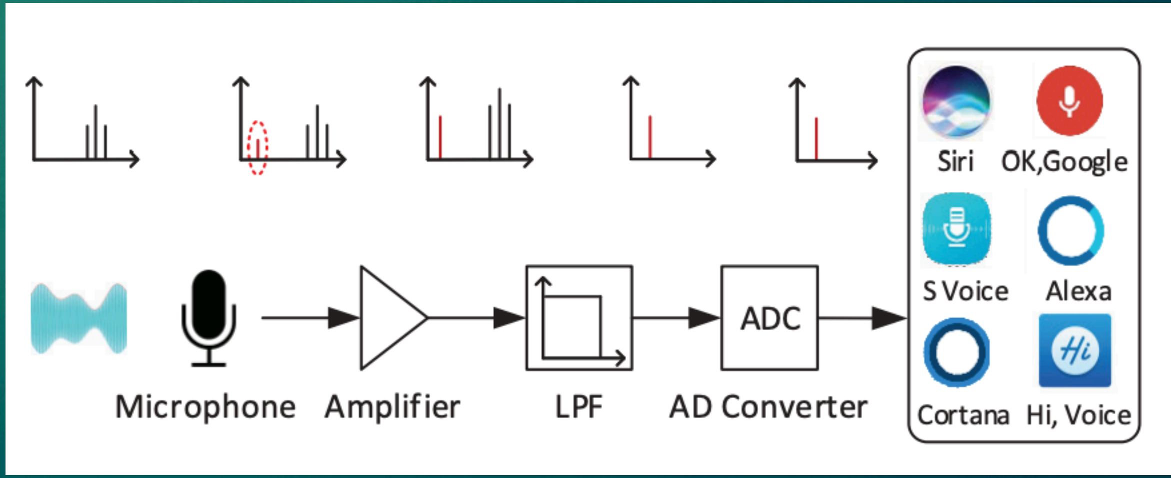
Background - VCS

- Voice capture
  - Records an ambient voice, amplifies, filters and digitizes
  - Pre-process, remove redundant frequencies
- Speech recognition
  - Activation – ‘Hey Siri’
  - Recognition
    - Speaker-dependent (Siri, Google Now)
    - Speaker-independent (Alexa)
- Command Execution
  - Launch application



Background - Microphone

- Microphone/transducer
  - Converts airborne acoustic waves into electrical signals
  - Ideally filters out of range sounds (F < 20Hz - 20kHz < F)
    - Sometimes signals higher/lower are still recorder
- MEMS microphone
  - Dominates the market
    - Size
    - low-power consumption



Threat Model

Threat Model

- No target device access
  - Posses the knowledge of device characteristics
- No owner interaction
- Inaudible
  - *F > 20kHz*
  - Upper bound of human hearing is 20kHz
    - Reason for some devices adopting sampling rate of under 44kHz
- Attacking equipment

Attack Design

Attack Design

- Idea
  - Generate baseband signal of voice commands both for activation and recognition (transmitter device)
  - Modulate baseband signals (transmitter device)
  - Demodulate at the VCS (receiver device)

Attack Design – Voice command generation

- Activation Command Generation
  - Contain wake words ("Hey Siri")
  - Tone to the specific user
    - TTS-Based Brute Force
      - Text-to-Speech can be used to brute force different voice frequencies
    - Concatenative Speech
      - Record an attackers speech and replay the frequencies of the letters to generate a wake command with the user's tone
- General Control Command
  - Ex: 'Call 911' or open 'www.google.com'

Attack Design – Voice Command Modulation

- Baseband signal modulation parameters
  - Modulation depth (m)
    - m = M/A   A – carrier amplitude   M – modulation amplitude
  - Carrier Frequency
    - Lowest Frequency of modulated signal > 20kHz to ensure inaudibility
    - Fcarrier – w > 20kHz   w – FrequencyVoice
  - Voice selection
    - Voice with small bandwidth
      - Female voice has a wider frequency band which leads to frequency leakage

Attack Design – Voice Commands Transmitter

- Powerful transmitter driven by dedicated signal generator
  - Validate and quantify results of a DolphinAttack
- Portable transmitter driven by a mobile device
  - Test walk-by attacks
- Structure of transmitters
  - 1) Signal Source – produces baseband signals of voice commands
  - 2) Modulator – modulates voice signal onto carrier wave
  - 3) Speaker – transforms signal into acoustic waves

Feasibility Experiments Across VCS

Feasibility Experiments across VCS System Selection

- Selected popular VCS and SR systems on the market

Attack	Device/System	Command
Recognition	Phones & Wearable	*Call 1234567890*
Recognition	iPad	*FaceTime 1234567890*
Recognition	MacBook & Nexus 7	*Open dolphinattack.com*
Recognition	Windows PC	*Turn on airplane mode*
Recognition	Amazon Echo	*Open the back door*
Recognition	Vehicle (Audi Q3)	*Navigation* *
Activation	Siri	*Hey Siri*
Activation	Google Now	*Ok Google*
Activation	Samsung S Voice	*Hi Galaxy*
Activation	Huawei HiVoice	*Hello Huawei* *
Activation	Alexa	*Alexa*

* The command is spoken in Chinese due to the lack of English support on these devices.

Feasibility Experiments across VCS Experiment Setup

- Equipment
  - Same equipment was used across all devices
- Setup
  - All experiments except for the automobiles SR systems were tested in the same environment

Feasibility Results and Limits

- Hardware Dependence
  - Difference in hardware shows great variance in
    - Attack distance
    - Success Rate
- SR system Dependence
  - Different audio handling
- Commands Matter
  - Length of the command defines success rate
- Carrier Wave Frequency

Experiment Results

Manuf.	Model	OS/Ver.	SR System	Attacks		Modulation Parameters		Max Dist. (cm)	
				Recog.	Activ.	f_c (kHz) & [Prime f_c] ‡	Depth	Recog.	Activ.
Apple	iPhone 4s	iOS 9.3.5	Siri	√	√	20–42 [27.9]	≥ 9%	175	110
Apple	iPhone 5s	iOS 10.0.2	Siri	√	√	24.1 26.2 27 29.3 [24.1]	100%	7.5	10
Apple	iPhone SE	iOS 10.3.1	Siri	√	√	22–28 33 [22.6]	≥ 47%	30	25
			Chrome	√	N/A	22–26 28 [22.6]	≥ 37%	16	N/A
Apple	iPhone SE †	iOS 10.3.2	Siri	√	√	21–29 31 33 [22.4]	≥ 43%	21	24
Apple	iPhone 6s *	iOS 10.2.1	Siri	√	√	26 [26]	100%	4	12
Apple	iPhone 6 Plus *	iOS 10.3.1	Siri	×	√	— [24]	—	—	2
Apple	iPhone 7 Plus *	iOS 10.3.1	Siri	√	√	21 24-29 [25.3]	≥ 50%	18	12
Apple	watch	watchOS 3.1	Siri	√	√	20–37 [22.3]	≥ 5%	111	164
Apple	iPad mini 4	iOS 10.2.1	Siri	√	√	22–40 [28.8]	≥ 25%	91.6	50.5
Apple	MacBook	macOS Sierra	Siri	√	N/A	20-22 24-25 27-37 39 [22.8]	≥ 76%	31	N/A
LG	Nexus 5X	Android 7.1.1	Google Now	√	√	30.7 [30.7]	100%	6	11
Asus	Nexus 7	Android 6.0.1	Google Now	√	√	24–39 [24.1]	≥ 5%	88	87
Samsung	Galaxy S6 edge	Android 6.0.1	S Voice	√	√	20–38 [28.4]	≥ 17%	36.1	56.2
Huawei	Honor 7	Android 6.0	HiVoice	√	√	29–37 [29.5]	≥ 17%	13	14
Lenovo	ThinkPad T440p	Windows 10	Cortana	√	√	23.4–29 [23.6]	≥ 35%	58	8
Amazon	Echo *	5589	Alexa	√	√	20-21 23-31 33-34 [24]	≥ 20%	165	165
Audi	Q3	N/A	N/A	√	N/A	21–23 [22]	100%	10	N/A

‡ Prime f_c is the carrier wave frequency that exhibits highest baseband amplitude after demodulation. — No result

† Another iPhone SE with identical technical spec.

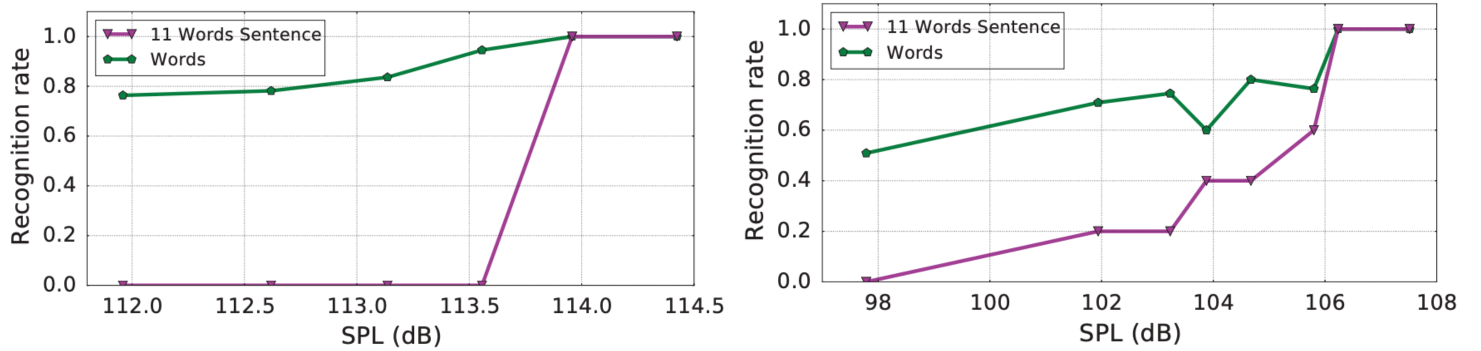
* Experimented with the front/top microphones on devices.

Impact Quantification

Impact Quantification - Limits

- Influence of languages
  - Accents
- Impact of background noises
- Impact of sound pressure
- Impact of Attack Distances

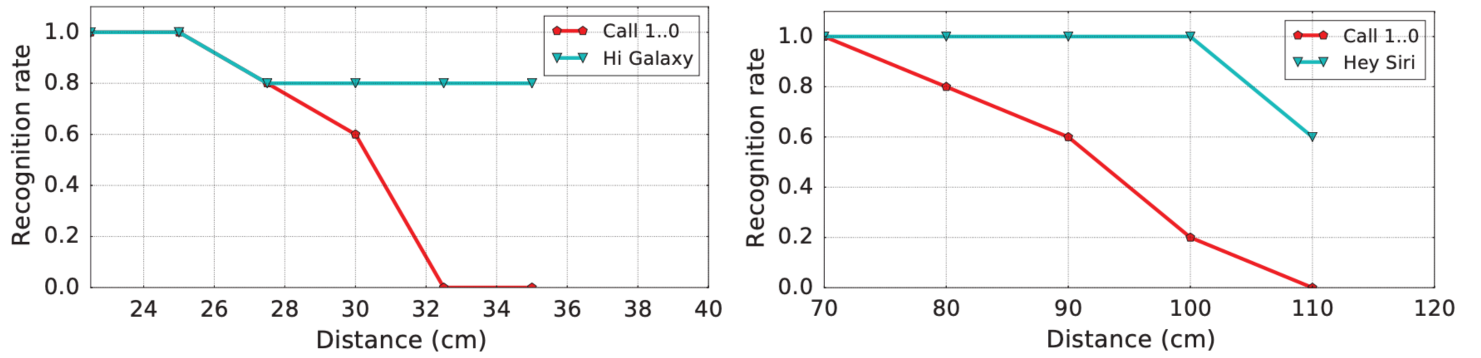
Impact Quantification – SPL and Distance



(a) The recognition rates of the Galaxy S6 Edge

(b) The recognition rates of the Apple watch

Figure 15: The impact of sound pressure levels on the recognition rates for two portable devices.



(a) The recognition rates of the Galaxy S6 Edge

(b) The recognition rates of the Apple watch

Figure 16: The impact of attack distances on the recognition rates for two portable devices.

Defenses

Defenses

- Hardware-based
  - Enhanced microphones to eliminate frequencies above 20kHz
  - Add a module prior to LPF to detect modulated commands and cancel them out
- Software-based
  - Since the original signal produces is much lower than 20kHz we can detect extensive alteration in the frequency (i.e. F > 20kHz)

Conclusion

Conclusion

- DolphinAttack
  - Inaudible attack to SR systems
  - Leverages amplitude modulation
- In order to avoid abuse of DolphinAttack two defenses were suggested
  - Hardware-based
  - Software-based