DolphinAttack: Inaudible Voice Commands

Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu

Zhejiang University

Presenter: Nikita Samarin

DolphinAttack

An approach to inject inaudible voice commands at voice controllable systems by exploiting the ultrasound channel and the vulnerability of the underlying audio hardware.

Voice Controllable System (VCS)

VCS = System + Speech Recognition

Voice Controllable System (VCS)

VCS = System + Speech Recognition

Examples:

- Apple iPhone + Siri
- Google Nexus + Google Now
- Amazon Echo + Alexa
- ...

Voice Controllable System (VCS)

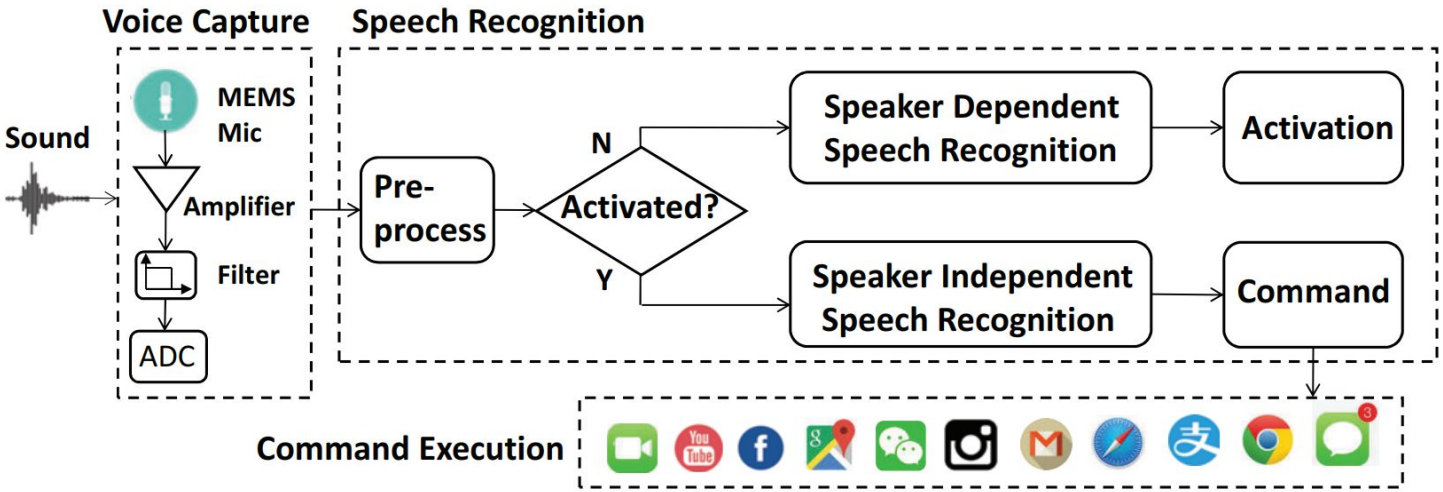


Figure 1: The architecture of a state-of-the-art VCS that can take voice commands as inputs and execute commands.

Voice Controllable System (VCS)

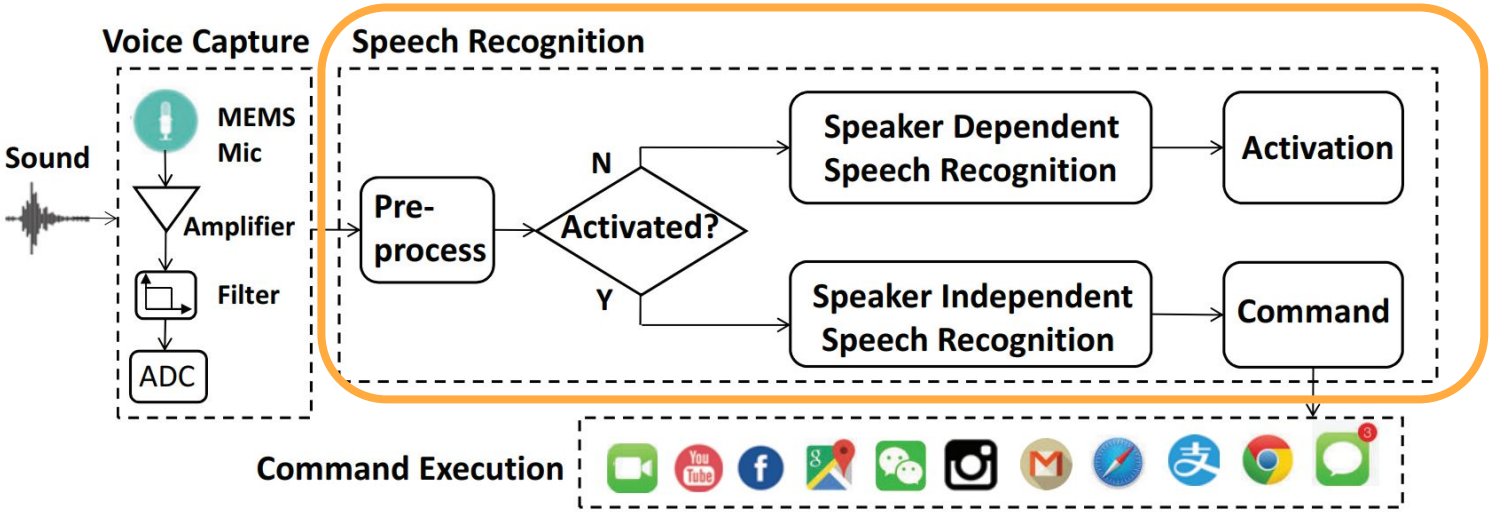
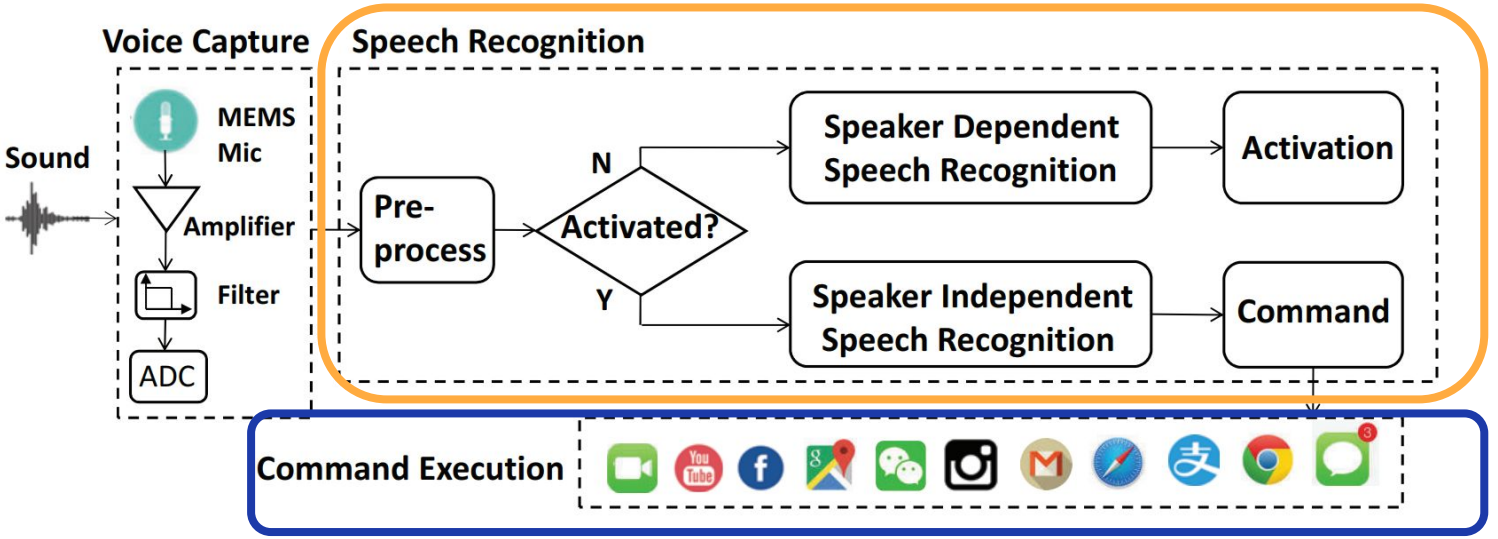


Figure 1: The architecture of a state-of-the-art VCS that can take voice commands as inputs and execute commands.

Voice Controllable System (VCS)

Machine Learning Attacks

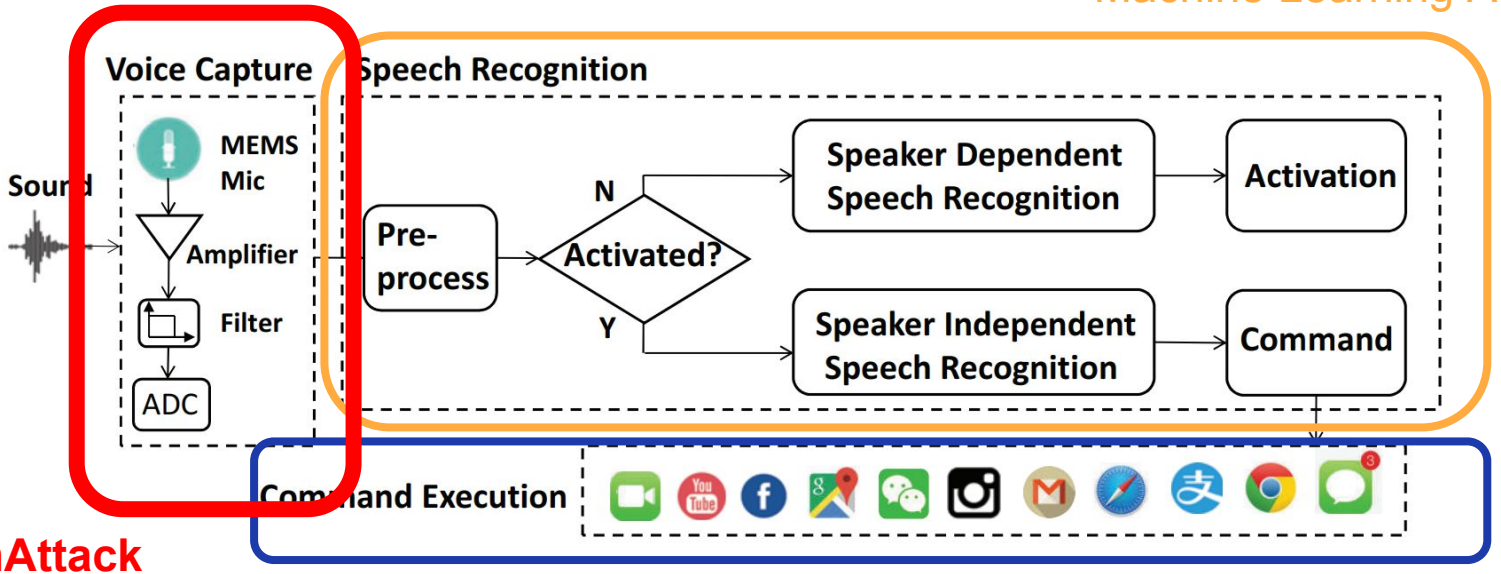


Malware

Figure 1: The architecture of a state-of-the-art VCS that can take voice commands as inputs and execute commands.

Voice Controllable System (VCS)

Machine Learning Attacks

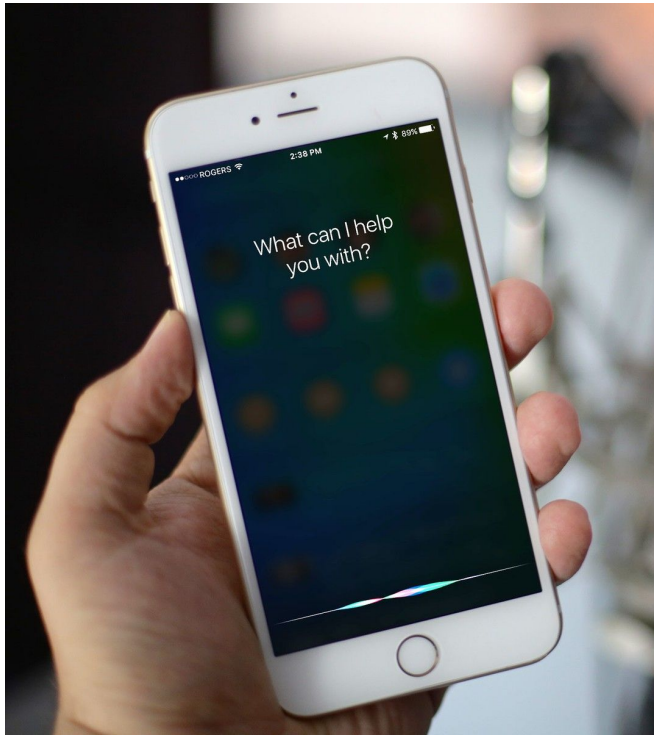


DolphinAttack

Malware

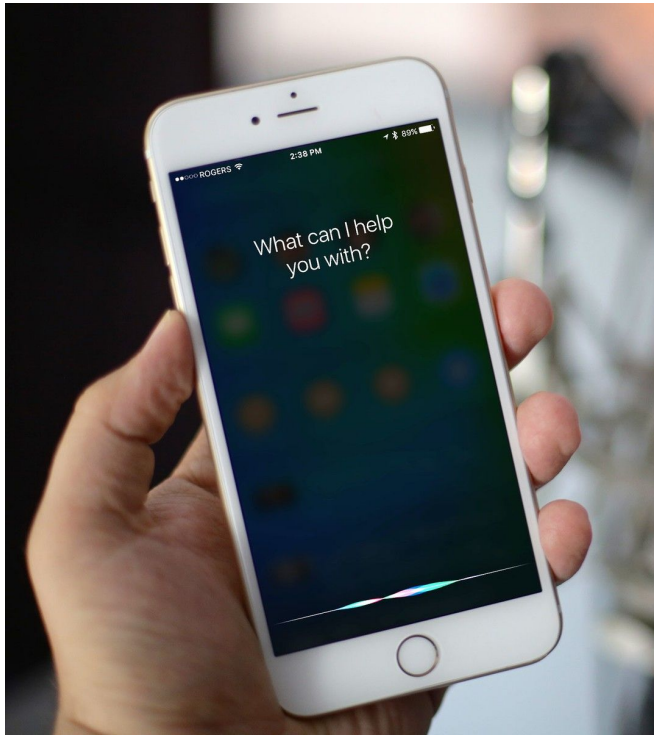
Figure 1: The architecture of a state-of-the-art VCS that can take voice commands as inputs and execute commands.

How can an attacker exploit this attack?



How can an attacker exploit this attack?

- Visiting a malicious website
- Spying
- Injecting fake information
- Denial of service
- … and more!



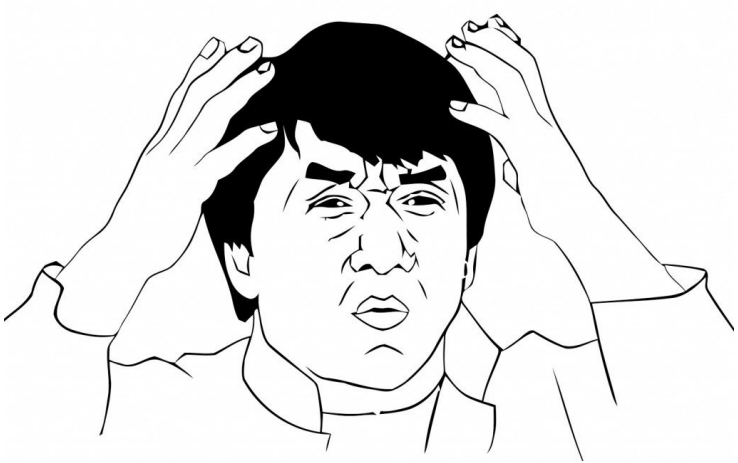
Fundamental Idea

Fundamental Idea

Modulate the low-frequency voice signal (baseband) on an ultrasonic carrier, and demodulate the modulated voice signals at the receiver…

Fundamental Idea

Modulate the low-frequency voice signal (baseband) on an ultrasonic carrier, and demodulate the modulated voice signals at the receiver…



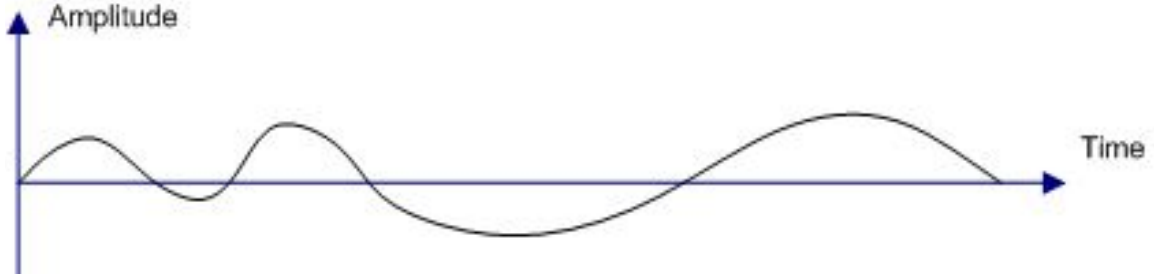
Categories of Sound Waves

- Infrasonic waves
  - $f < 20$ Hz
- Audible sound waves
  - $f = 20$ Hz – 20 kHz
- Ultrasonic waves
  - $f > 20$ kHz

(Amplitude) Modulation

(Amplitude) Modulation

Low-Frequency
Voice Signal
(Baseband)

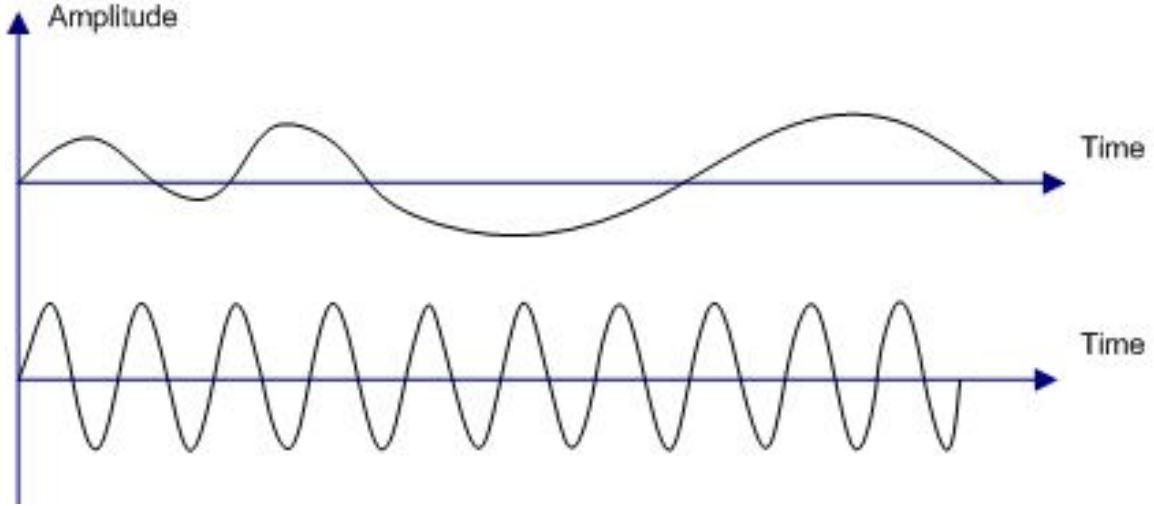


(Amplitude) Modulation

Low-Frequency
Voice Signal
(Baseband)

+

Ultrasonic
(high-frequency)
Carrier Signal



(Amplitude) Modulation

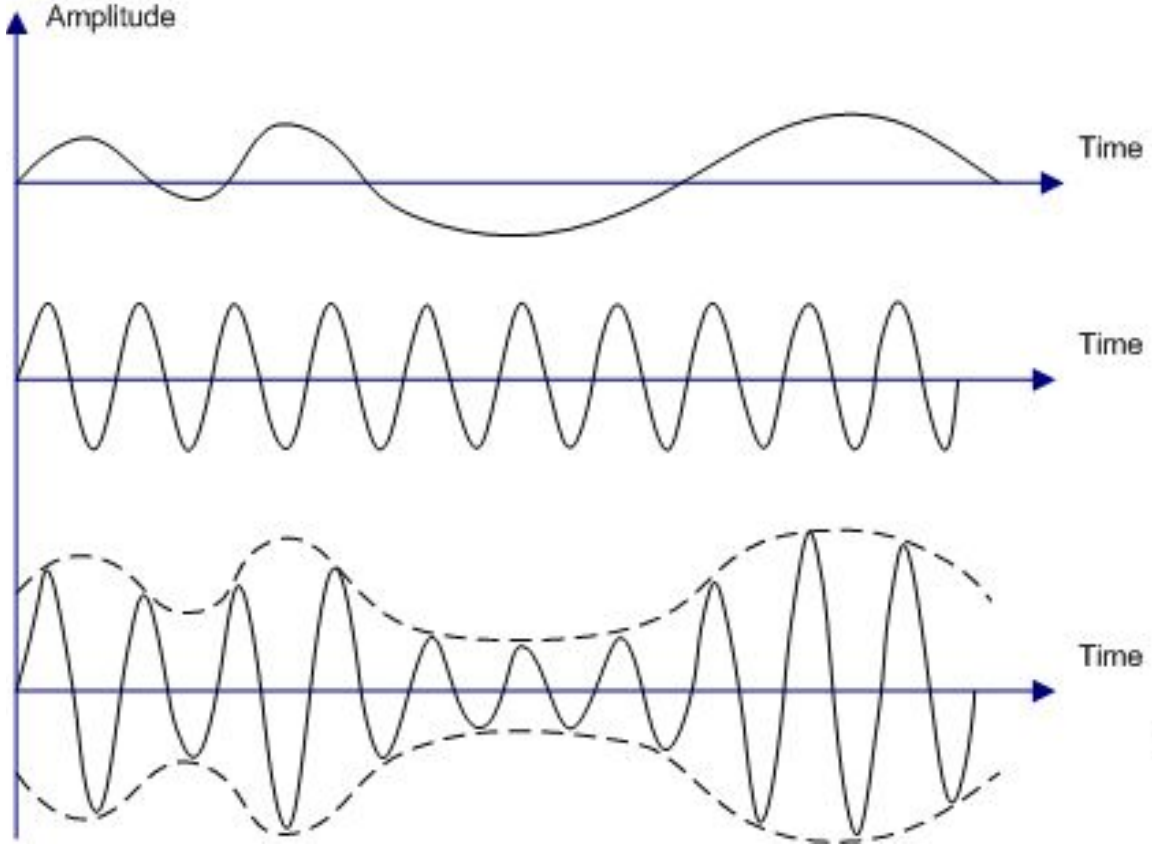
Low-Frequency
Voice Signal
(Baseband)

+

Ultrasonic
(high-frequency)
Carrier Signal

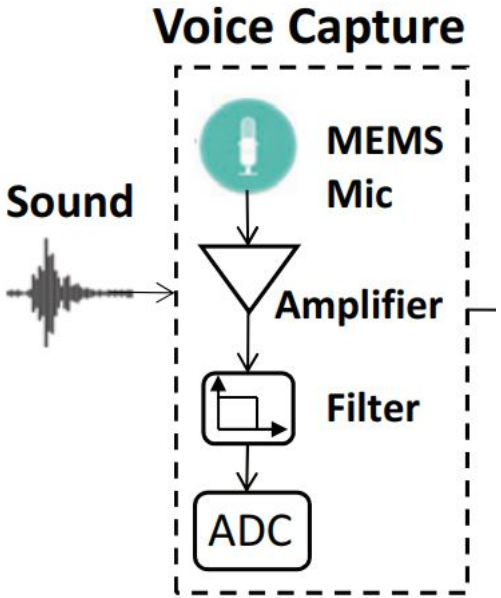
=

Modulated
(high-frequency)
Voice Signal

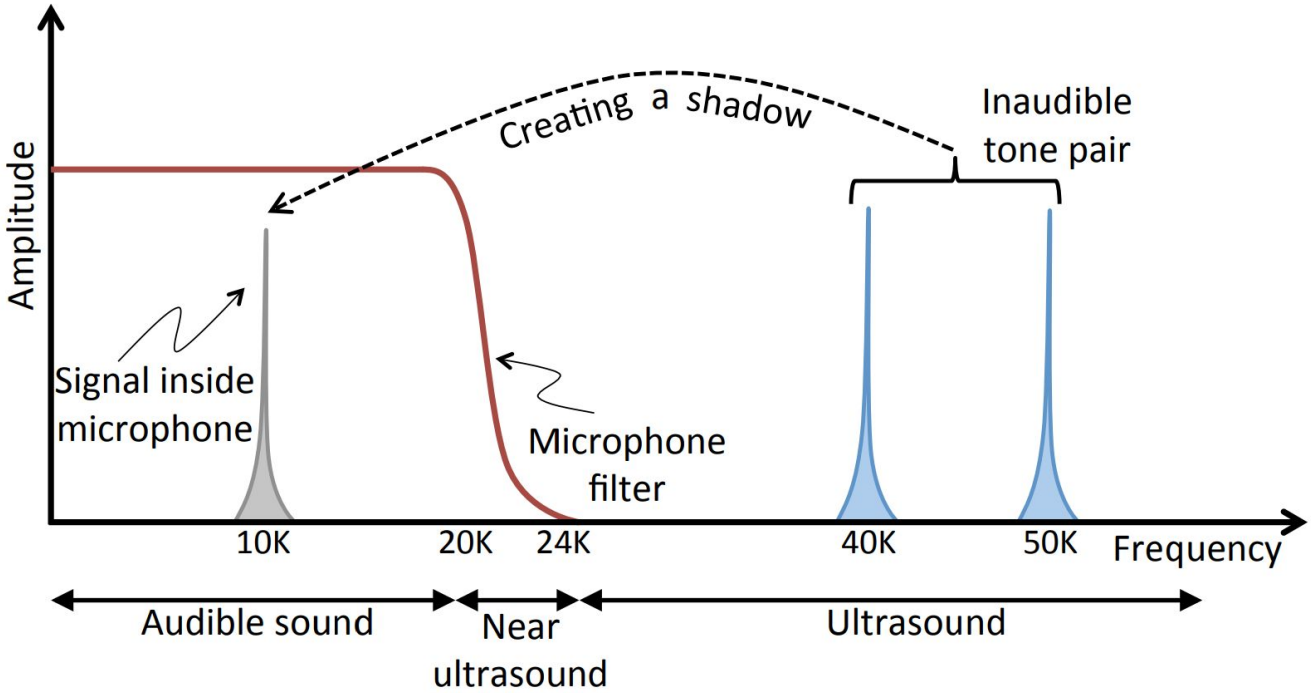


How to recover the voice signal?

- Exploit the electrical characteristics of microphones and amplifiers...

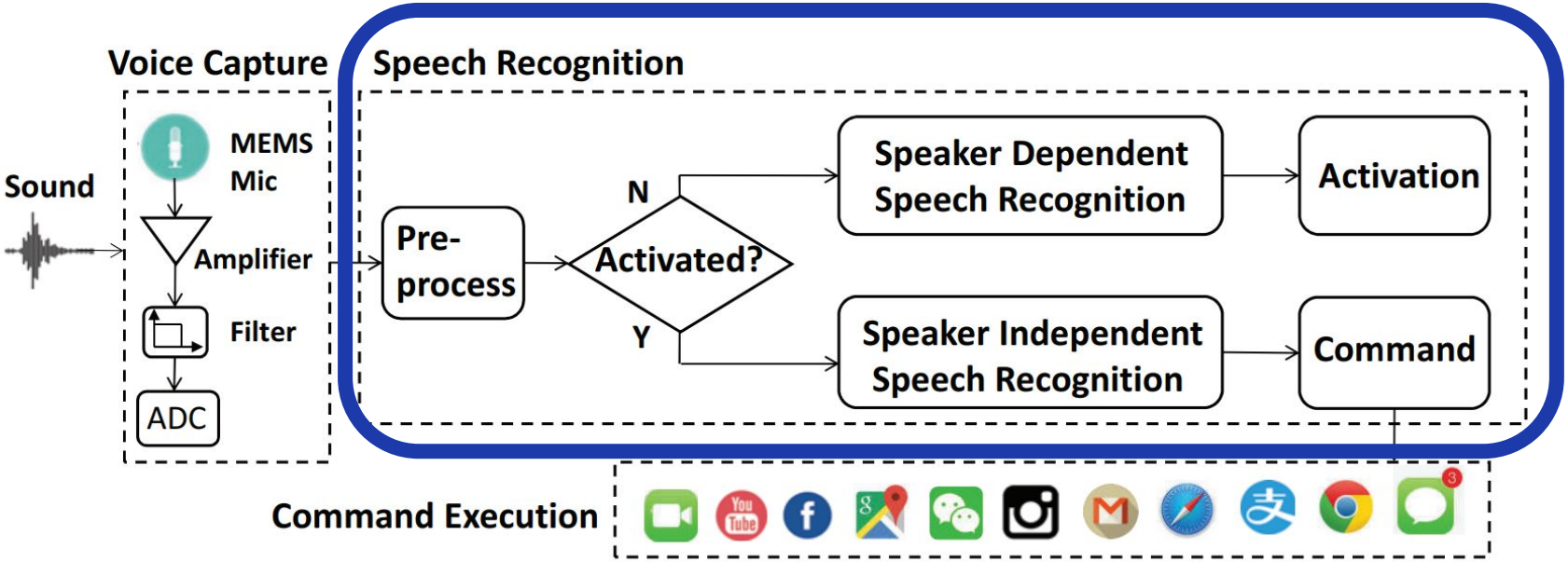


How to recover the voice signal?



Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17). ACM, New York, NY, USA, 2-14. DOI: https://doi.org/10.1145/3081333.3081366

Voice Command Generation



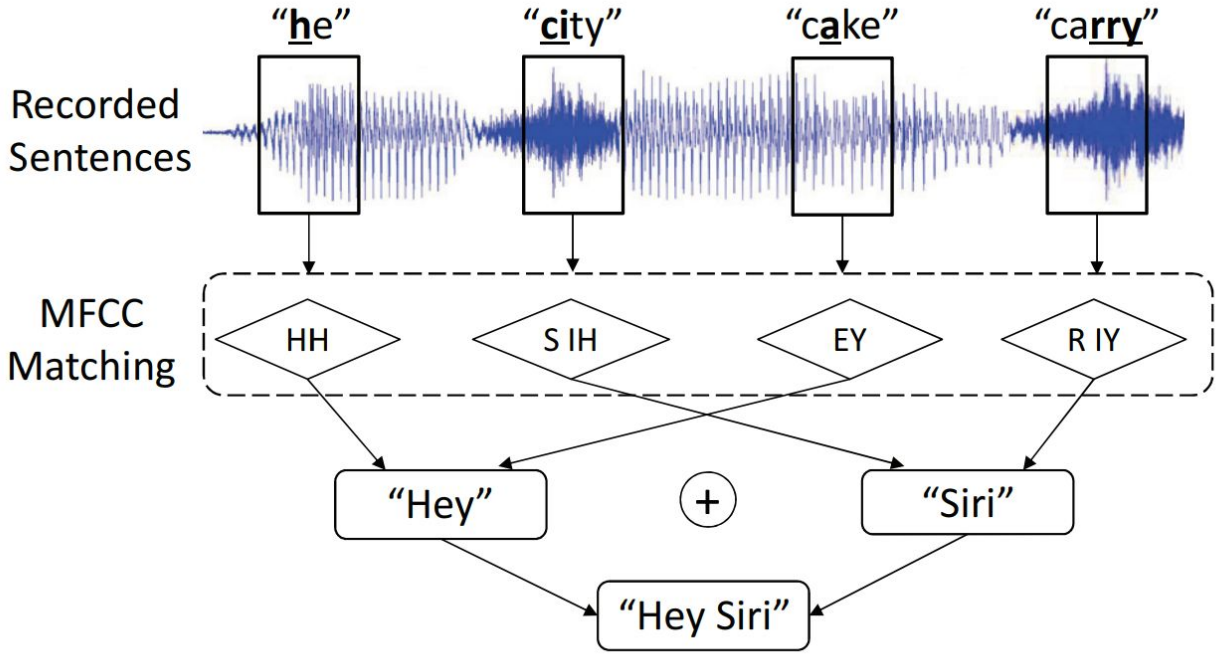
Activation Command Generation - Approach #1

Text-to-speech based brute force

TTS Systems	voice type #	# of successful types	
		Call 12..90	Hey Siri
Selvy Speech [51]	4	4	2
Baidu [8]	1	1	0
Sestek [45]	7	7	2
NeoSpeech [39]	8	8	2
Innoetics [59]	12	12	7
Vocalware [63]	15	15	8
CereProc [12]	22	22	9
Acapela [22]	13	13	1
Fromtexttospeech [58]	7	7	4

Activation Command Generation - Approach #2

Concatenative synthesis (with a few voice recordings)



Threat Model

Threat Model

- Attacker has no access to the target device

Threat Model

- Attacker has no access to the target device
  - But is fully aware of the technical characteristics

Threat Model

- Attacker has no access to the target device
  - But is fully aware of the technical characteristics
- No owner interaction (e.g. unlocking the screen)

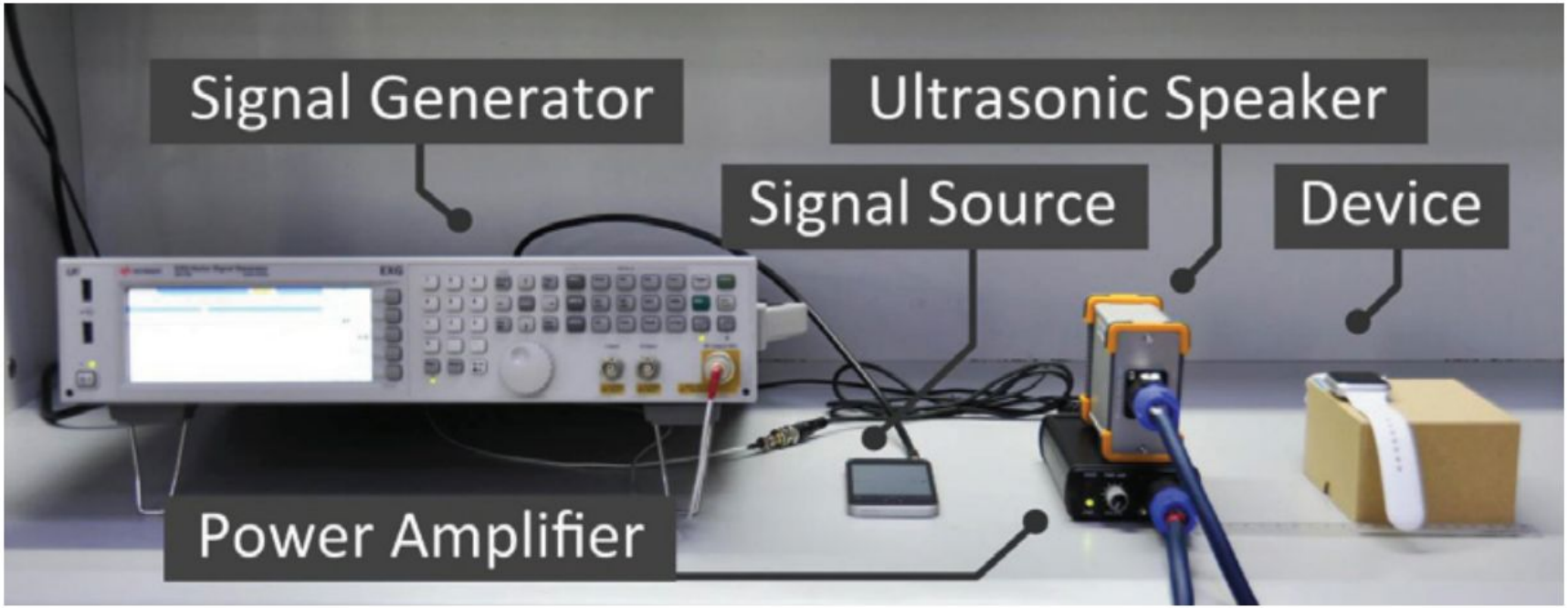
Threat Model

- Attacker has no access to the target device
  - But is fully aware of the technical characteristics
- No owner interaction (e.g. unlocking the screen)
- Attacker will use inaudible voice commands
  - Ultrasound (f > 20 kHz)

Threat Model

- Attacker has no access to the target device
  - But is fully aware of the technical characteristics
- No owner interaction (e.g. unlocking the screen)
- Attacker will use inaudible voice commands
  - Ultrasound (f > 20 kHz)
- Attacker can acquire the required equipment (e.g. speakers designed for transmitting ultrasound)

Experiment Setup (Feasibility Analysis)



Demonstration

Targeted Systems

Table 2: The list of systems and voice commands being tested in Tab. 3.

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.

Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Apple	iPhone 4s	iOS 9.3.5	Siri	√	√	175	110
Apple	iPhone 5s	iOS 10.0.2	Siri	√	√	7.5	10
Apple	iPhone SE	iOS 10.3.1	Siri	√	√	30	25
			Chrome	√	N/A	16	N/A
Apple	iPhone SE †	iOS 10.3.2	Siri	√	√	21	24
Apple	iPhone 6s *	iOS 10.2.1	Siri	√	√	4	12
Apple	iPhone 6 Plus *	iOS 10.3.1	Siri	×	√	–	2
Apple	iPhone 7 Plus *	iOS 10.3.1	Siri	√	√	18	12
Apple	watch	watchOS 3.1	Siri	√	√	111	164
Apple	iPad mini 4	iOS 10.2.1	Siri	√	√	91.6	50.5
Apple	MacBook	macOS Sierra	Siri	√	N/A	31	N/A
LG	Nexus 5X	Android 7.1.1	Google Now	√	√	6	11
Asus	Nexus 7	Android 6.0.1	Google Now	√	√	88	87
Samsung	Galaxy S6 edge	Android 6.0.1	S Voice	√	√	36.1	56.2
Huawei	Honor 7	Android 6.0	HiVoice	√	√	13	14
Lenovo	ThinkPad T440p	Windows 10	Cortana	√	√	58	8
Amazon	Echo *	5589	Alexa	√	√	165	165
Audi	Q3	N/A	N/A	√	N/A	10	N/A

Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Apple	iPhone 4s	iOS 9.3.5	Siri	√	√	175	110
Apple	iPhone 5s	iOS 10.0.2	Siri	√	√	7.5	10
Apple	iPhone SE	iOS 10.3.1	Siri	√	√	30	25
			Chrome	√	N/A	16	N/A
Apple	iPhone SE †	iOS 10.3.2	Siri	√	√	21	24
Apple	iPhone 6s *	iOS 10.2.1	Siri	√	√	4	12
Apple	iPhone 6 Plus *	iOS 10.3.1	Siri	×	√	–	2
Apple	iPhone 7 Plus *	iOS 10.3.1	Siri	√	√	18	12
Apple	watch	watchOS 3.1	Siri	√	√	111	164

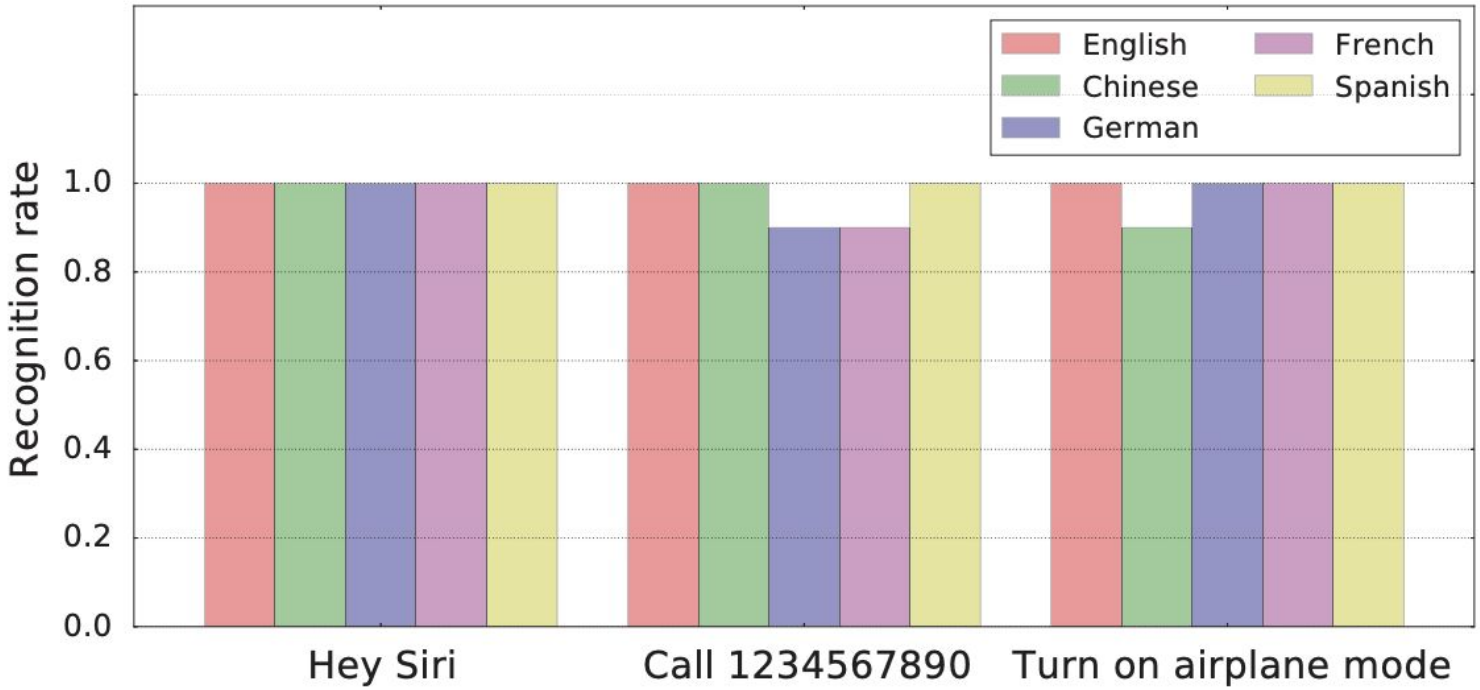
Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Apple	iPhone 4s	iOS 9.3.5	Siri	√	√	175	110
Apple	iPhone 5s	iOS 10.0.2	Siri	√	√	7.5	10
Apple	iPhone SE	iOS 10.3.1	Siri	√	√	30	25
			Chrome	√	N/A	16	N/A
Apple	iPhone SE †	iOS 10.3.2	Siri	√	√	21	24
Apple	iPhone 6s *	iOS 10.2.1	Siri	√	√	4	12
Apple	iPhone 6 Plus *	iOS 10.3.1	Siri	×	√	–	2
Apple	iPhone 7 Plus *	iOS 10.3.1	Siri	√	√	18	12
Apple	watch	watchOS 3.1	Siri	√	√	111	164

Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Apple	iPhone 4s	iOS 9.3.5	Siri	√	√	175	110
Apple	iPhone 5s	iOS 10.0.2	Siri	√	√	7.5	10
Apple	iPhone SE	iOS 10.3.1	Siri	√	√	30	25
			Chrome	√	N/A	16	N/A
Apple	iPhone SE †	iOS 10.3.2	Siri	√	√	21	24
Apple	iPhone 6s *	iOS 10.2.1	Siri	√	√	4	12
Apple	iPhone 6 Plus *	iOS 10.3.1	Siri	×	√	–	2
Apple	iPhone 7 Plus *	iOS 10.3.1	Siri	√	√	18	12
Apple	watch	watchOS 3.1	Siri	√	√	111	164

Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Huawei	Honor 7	Android 6.0	HiVoice	√	√	13	14
Lenovo	ThinkPad T440p	Windows 10	Cortana	√	√	58	8
Amazon	Echo *	5589	Alexa	√	√	165	165
Audi	Q3	N/A	N/A	√	N/A	10	N/A

Manuf.	Model	OS/Ver.	SR System	Attacks		Max Dist. (cm)	
				Recog.	Activ.	Recog.	Activ.
Huawei	Honor 7	Android 6.0	HiVoice	√	√	13	14
Lenovo	ThinkPad T440p	Windows 10	Cortana	√	√	58	8
Amazon	Echo *	5589	Alexa	√	√	165	165
Audi	Q3	N/A	N/A	√	N/A	10	N/A

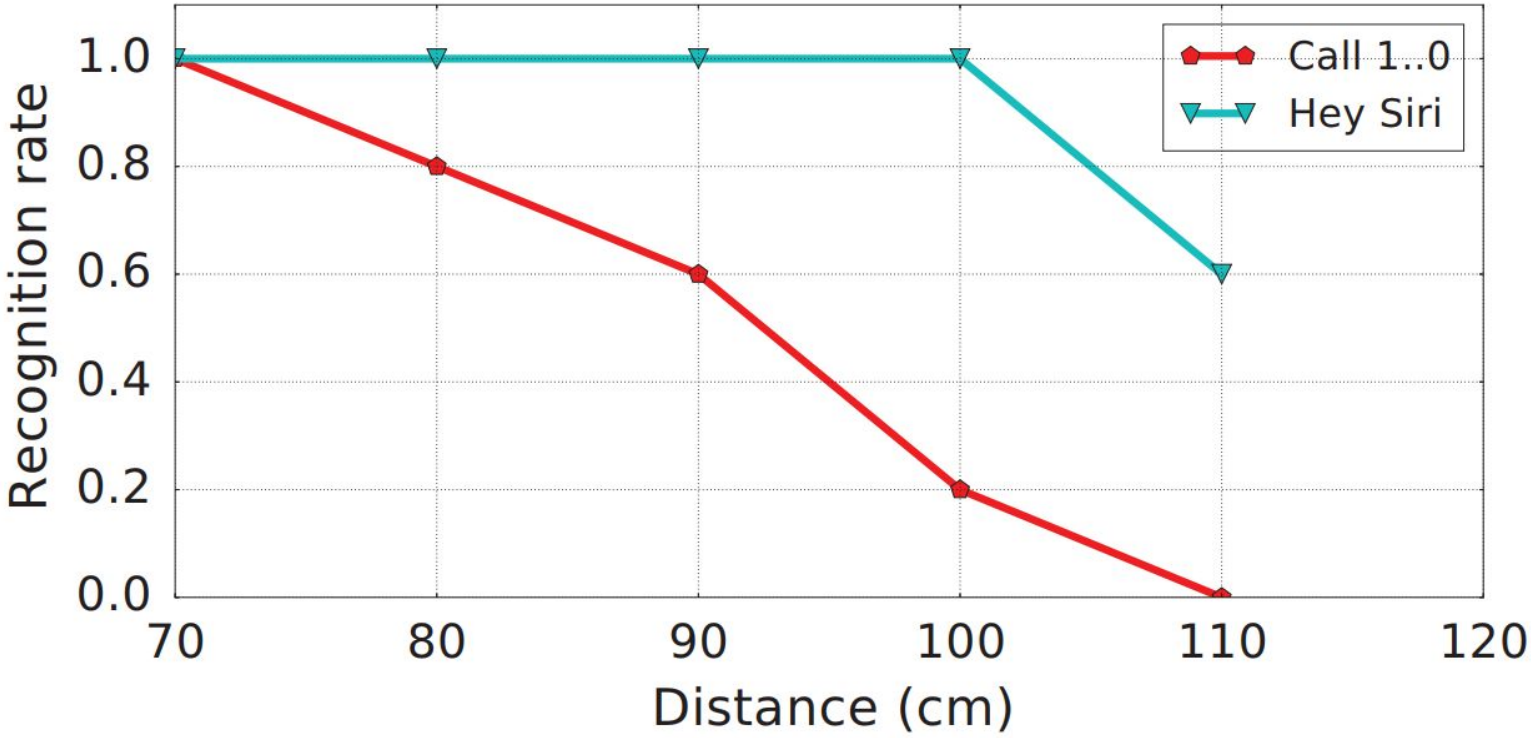
Influence of Languages (Apple Watch)



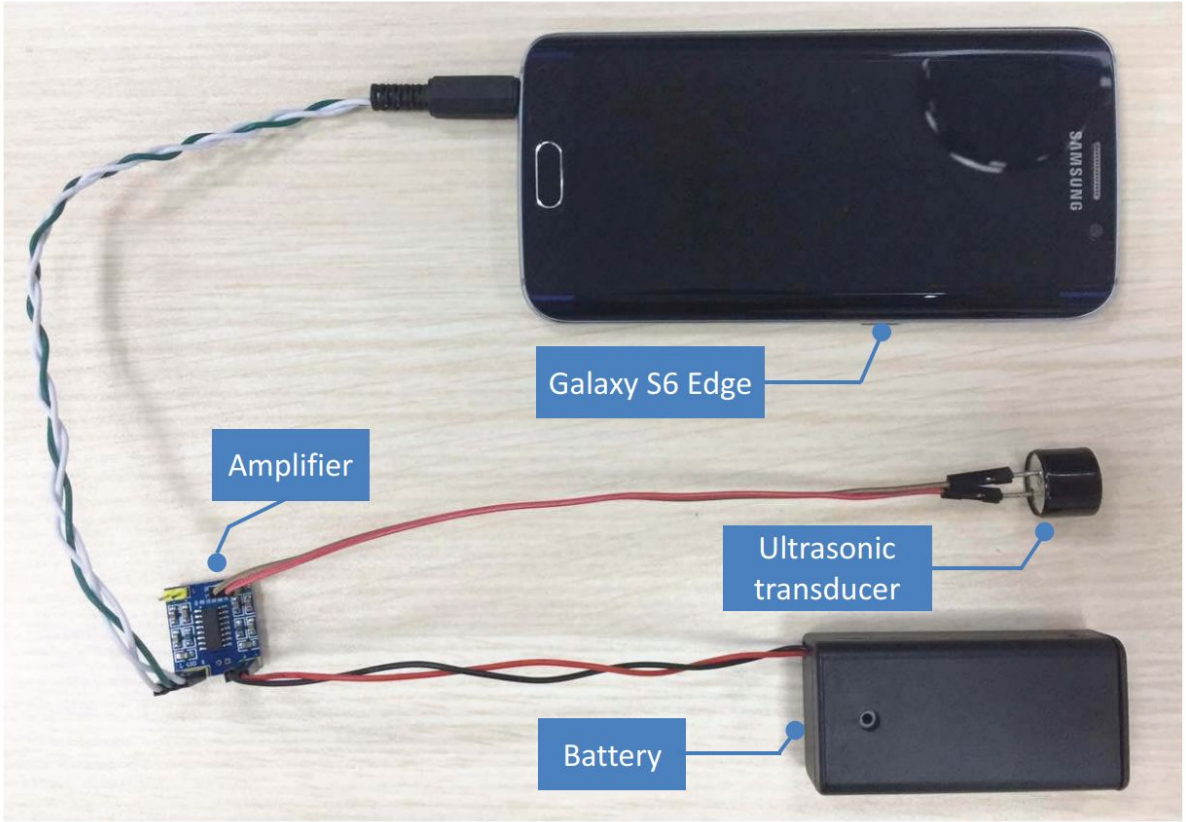
Impact of Background Noises (Apple Watch)

Scene	Noises (dB)	Recognition rates	
		Hey Siri	Turn on airplane mode
Office	55–65	100%	100%
Cafe	65–75	100%	80%
Street	75–85	90%	30%

Impact of Attack Distances (Apple Watch)



Great! What about something more portable?



Portable Setup

"Turn on airplane mode" (without amplifier)

Table 5: Portable device attack results. Attacking an Apple watch using a Galaxy S6 Edge smartphone that is 2 cm away.

f_c (kHz)	20	21	22	23	24
Word recognition rate	80%	100%	16%	100%	0%
Sentence recognition rate	80%	100%	0%	100%	0%

Portable Setup

*With the amplifier module, the maximum distance of effective attacks is increased to **27 cm**.*

Proposed Defenses

Proposed Defenses

- Hardware-based

Proposed Defenses

- Hardware-based
  - Microphone enhancement

Proposed Defenses

- Hardware-based
  - Microphone enhancement
  - Inaudible voice command cancellation

Proposed Defenses

- Hardware-based
  - Microphone enhancement
  - Inaudible voice command cancellation
- Software-based

Proposed Defenses

- Hardware-based
  - Microphone enhancement
  - Inaudible voice command cancellation
- Software-based
  - Distinguish modulated voice commands and genuine ones using machine learning (e.g. SVM)

Thank you!

Questions?