DOLPHIN ATTACK: INAUDIBLE VOICE COMMANDS

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

BACKGROUND- DOLPHIN ATTACK

An approach to inject inaudible voice commands at VCS by exploiting the ultrasound channel (i.e., f > 20 kHz) and the vulnerability of the underlying audio hardware

BACKGROUND
SPEECH RECOGNITION

- Allows machines or programs to identify spoken words and convert them into machine-readable formats
- It has become an increasingly popular human-computer interaction mechanism because of its accessibility, efficiency, and recent advances in recognition accuracy

BACKGROUND - VCS

- Voice Controllable System
  - Speech recognition combined with a system

Apple iPhone – Siri

Amazon Echo – Alexa

VOICE CONTROLLABLE SYSTEM

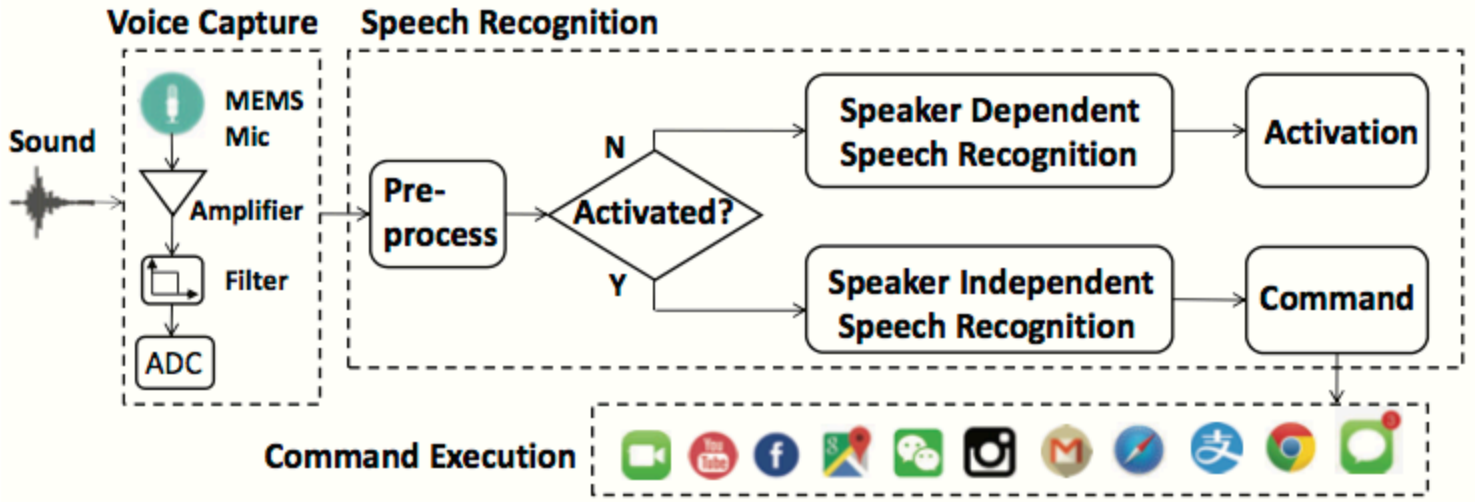


Figure 1: The architecture of a state-of-the-art VCS that can take voice commands as inputs and execute commands.

ATTACKS ON VCS

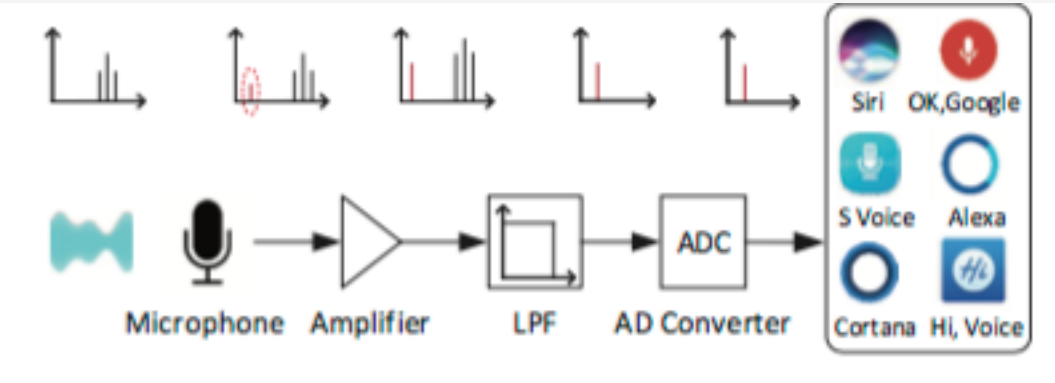
- Visiting a malicious site
  - Drive-by-download attack
  - Exploit device with 0-day vulnerabilities
- Spying
  - Initiate video/phone calls to gain visual/sound of device surroundings

ATTACKS ON VCS

- Injecting fake information
  - Inject command to send fake texts/emails
  - Publish fake posts
  - Add fake events in calendar
- Denial of service
  - Airplane mode
- Concealing attacks
  - Dimming screen and lowering volume

BACKGROUND - MICROPHONE

- Voice capture system that converts airborne acoustic waves to electrical signals
- Two main types
  - Electret Condenser Microphone (ECMs)
  - Micro Electro Mechanical System (MEMS)



BACKGROUND
SOUND WAVES

- Human audible
  - 20 Hz < f <20 kHz
- Ultrasonic
  - f > 20 kHz

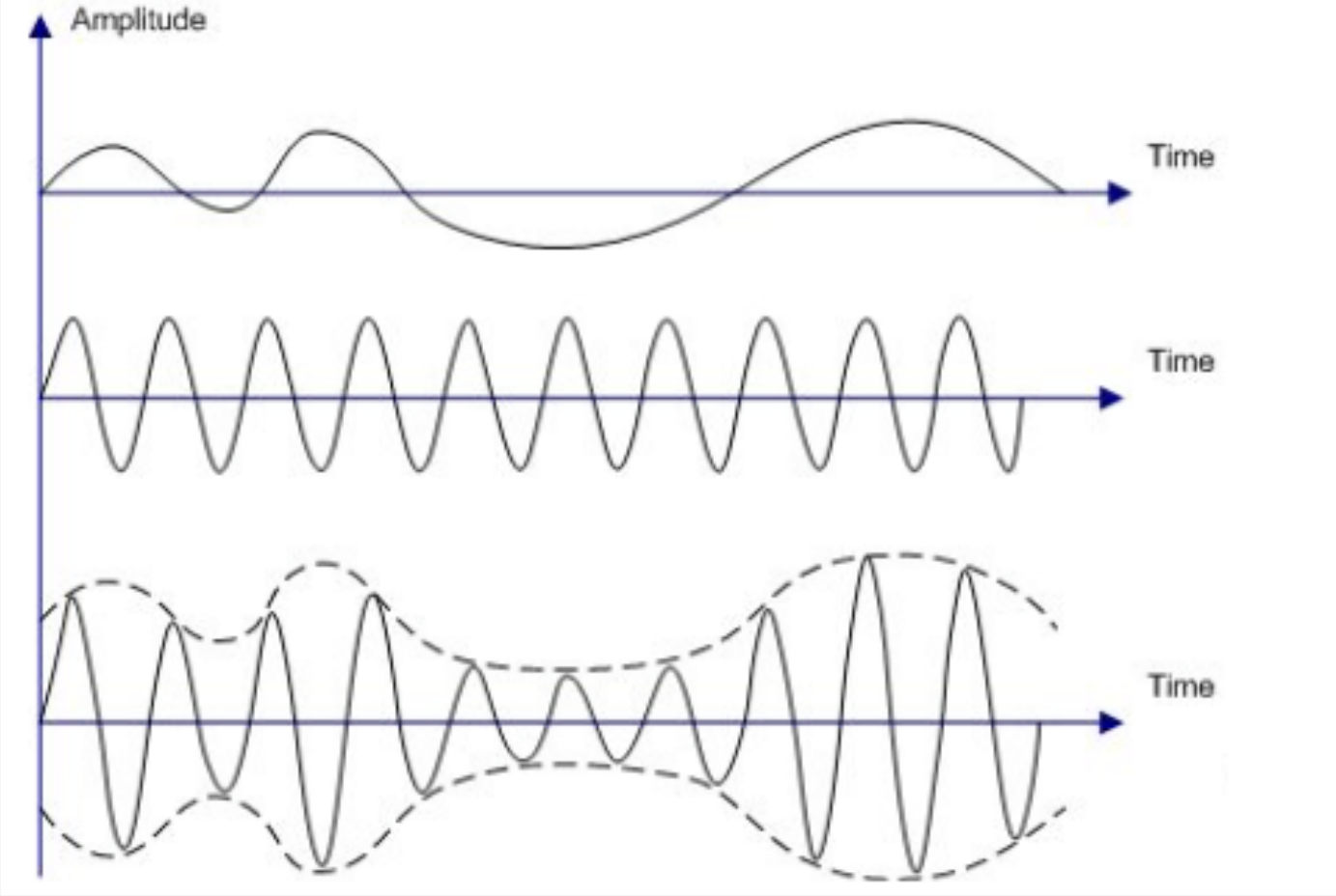
THREAT MODEL

- No target device access
- No owner interaction
  - In vicinity, but not in use and draw no attention
- Inaudible voice commands will be used
  - Ultrasounds
- Attacking equipment
  - Speaker to transmit ultrasound
  - Speaker is in the vicinity of target device

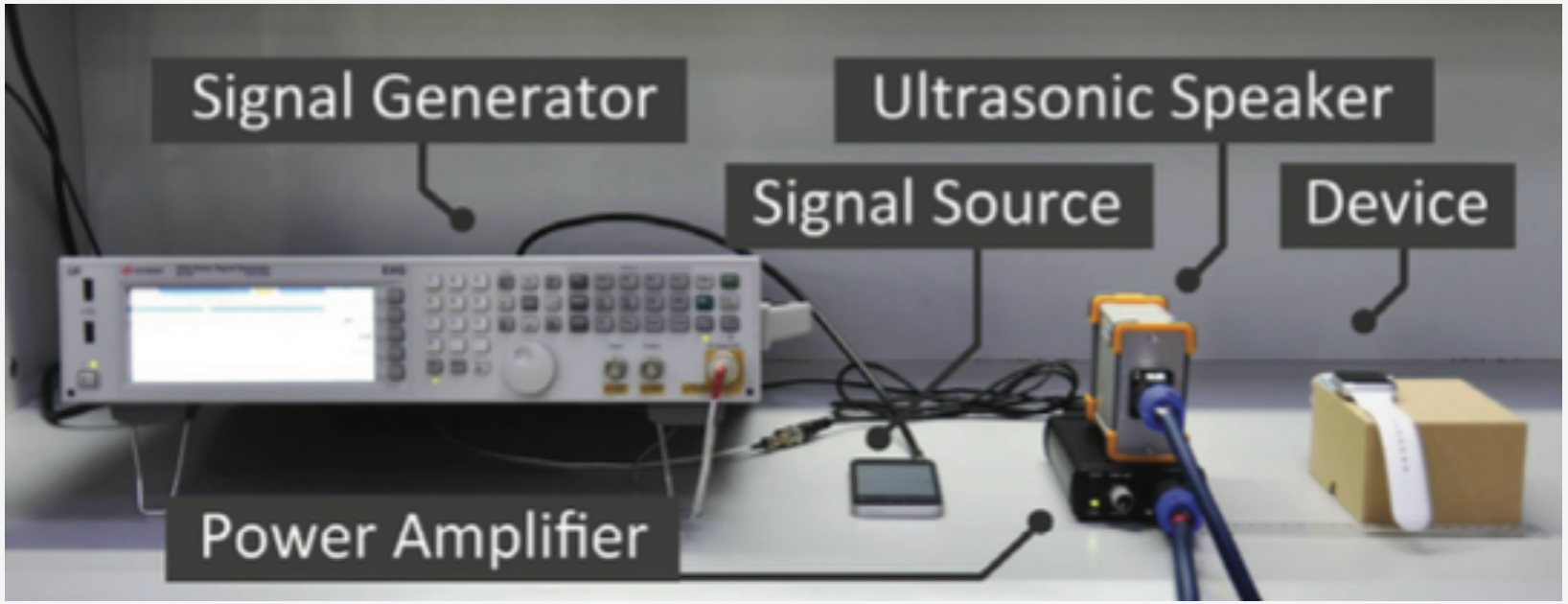
FEASIBILITY ANALYSIS

- The fundamental idea of DolphinAttack
  - To modulate the low-frequency voice signal (i.e., baseband) on an ultrasonic carrier before transmitting it over the air
  - To demodulate the modulated voice signals with the voice capture hardware (VCH) at the receiver
    - No control over VCH so modulated signals must be crafted so that it can be demodulated to the baseband signal using the VCH

FEASIBILITY ANALYSIS



FEASIBILITY ANALYSIS
EXPERIMENTAL SETUP



ATTACK DESIGN

- Case Study – Siri
- Siri Activation
  - "Hey Siri" – in the tone of the user it is trained for

- Generate Activation
  - Stolen phone (no owner)
  - Attacker can obtain a few recordings of the owner

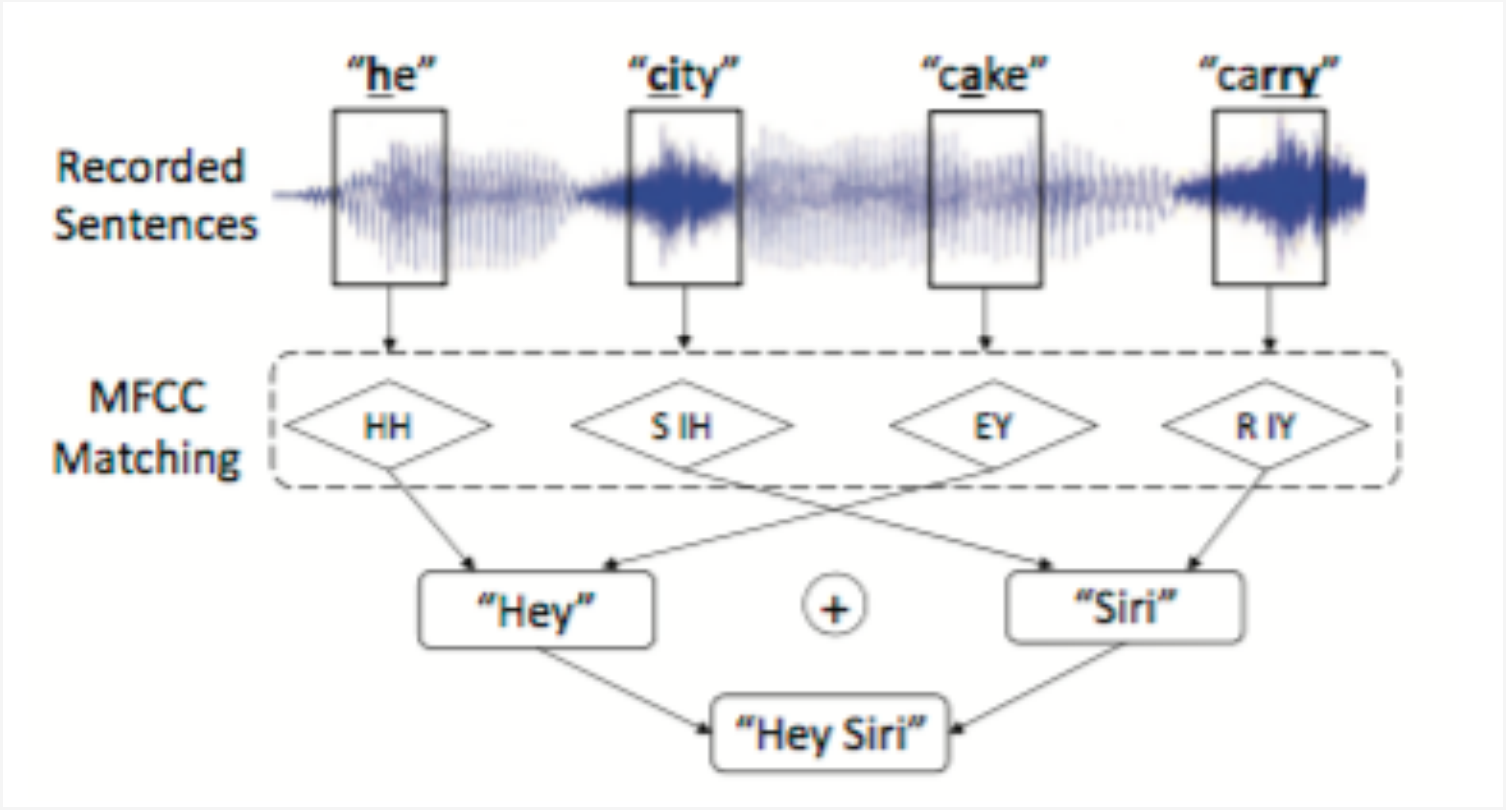
ATTACK DESIGN

- TTS-based Brute Force
  - Downloaded two voice commands from websites of these TTS systems
  - ”Hey Siri” from Google TTS was used to train Siri

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

35 of 89 types of activation commands activate Siri – 39%

ATTACK DESIGN



- 44 phonemes in English
  - 6 are used in "Hey Siri"
- "he", "cake", "city", "carry"
- "he is a boy", "eat a cake", "in the city", "read after me"

- Both able to activate Siri successfully

ATTACK DESIGN

- Voice commands are now generated
- Voice commands must be modulated onto ultrasonic carriers
- Lowest frequency of the modulated signal should be larger than 20 kHz to ensure inaudibility

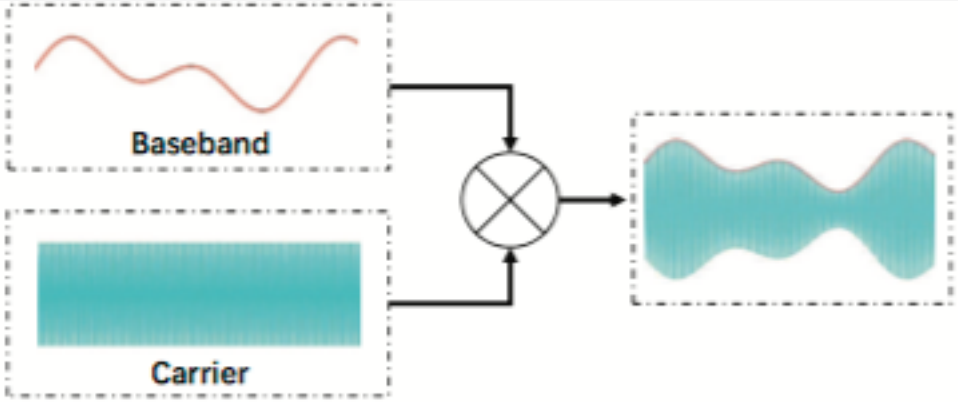
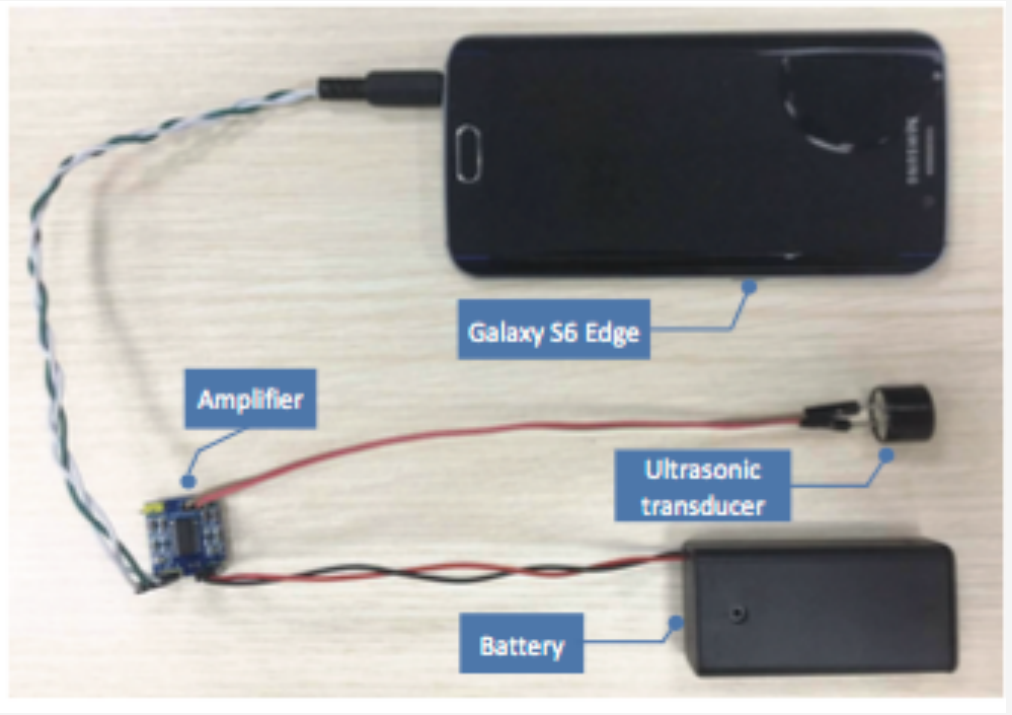


Figure 9: An illustration of modulating a voice command onto an ultrasonic carrier using AM modulation.

ATTACK DESIGN

- Voice Commands Transmitter
  - A powerful transmitter with signal generator
  - The portable transmitter with a smartphone



ATTACK EXPERIMENT

- https://www.youtube.com/watch?v=21HjF4A3WE4

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*

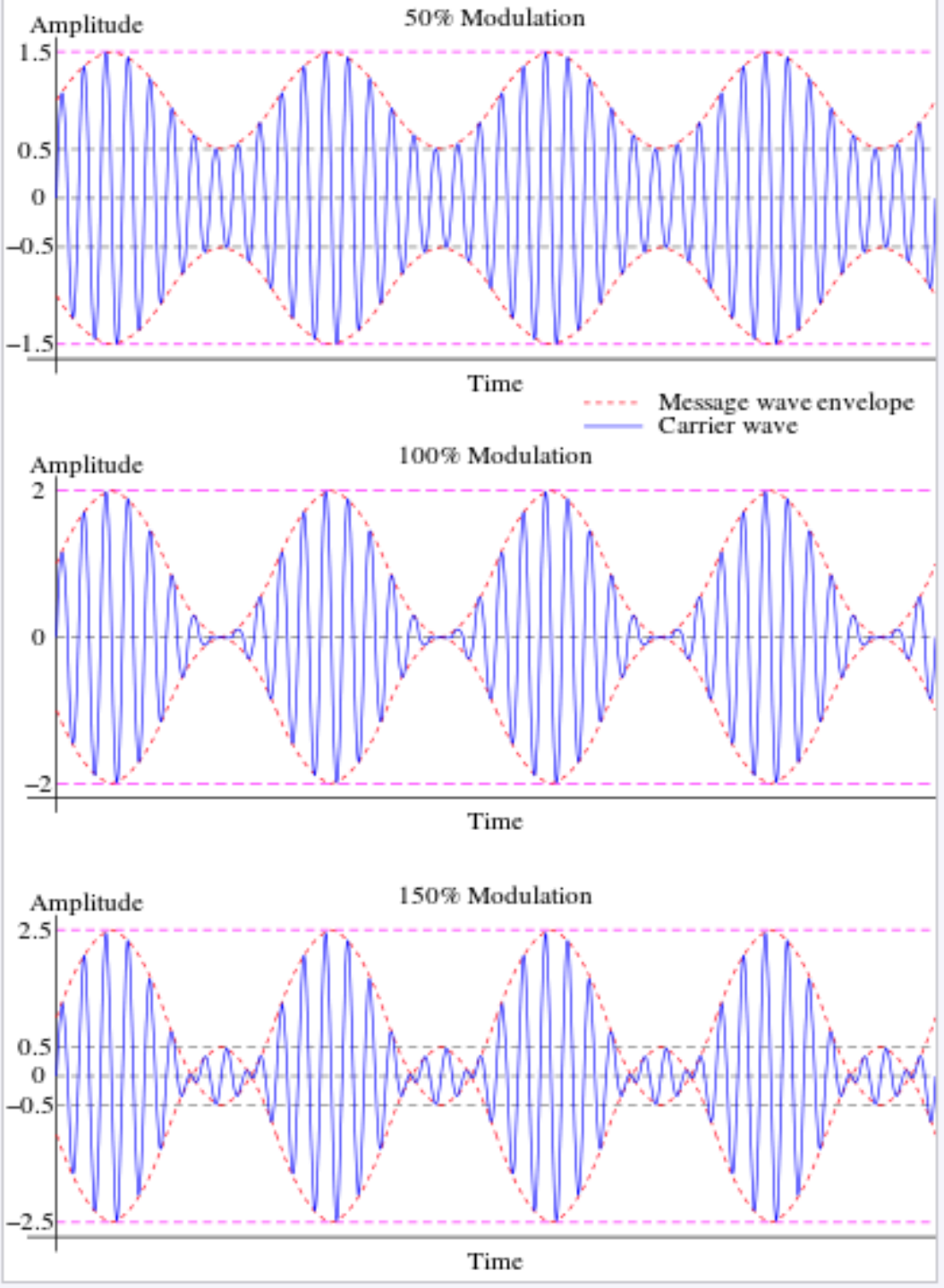
List of system and voice commands set to be tested

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

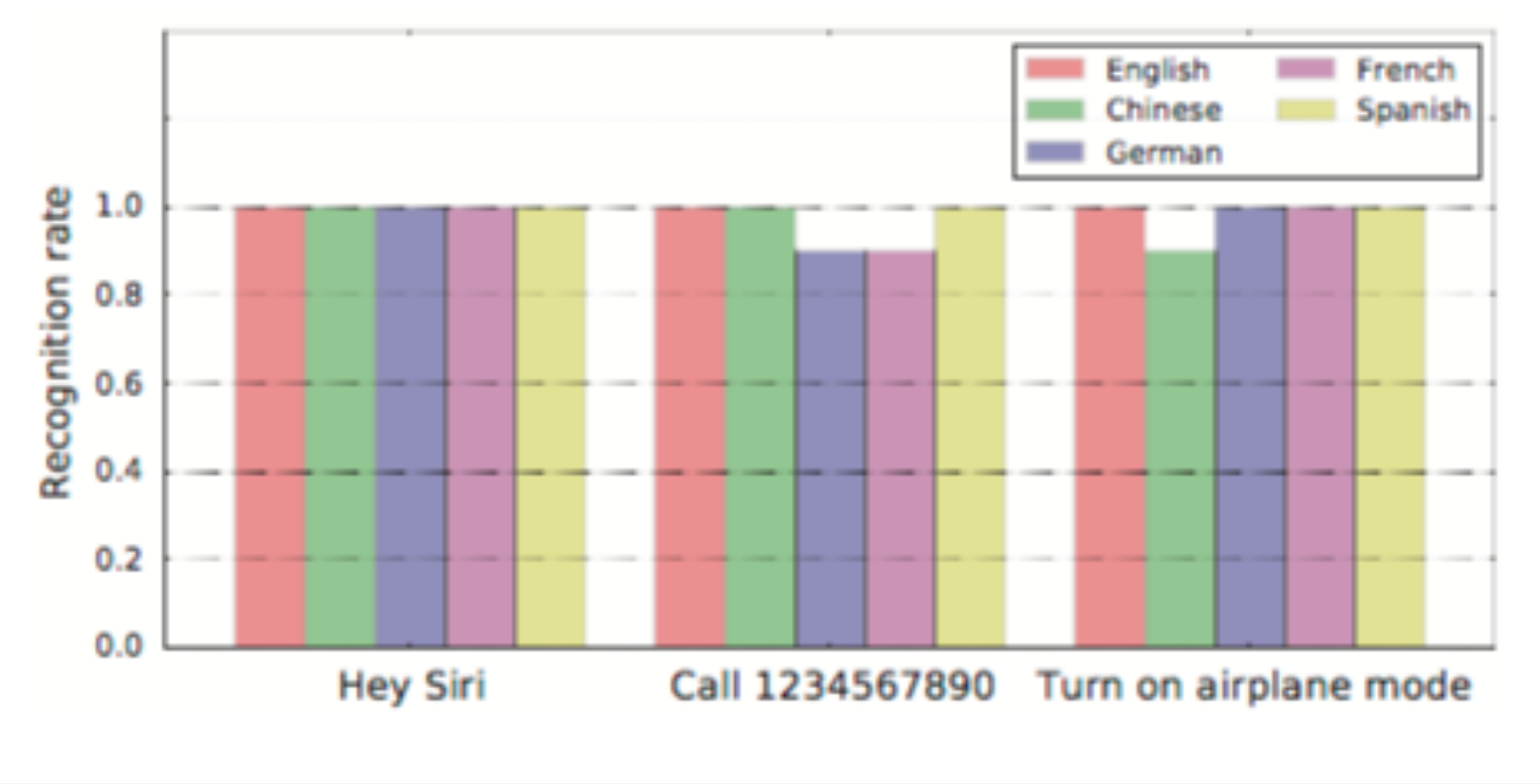
† Another iPhone SE with identical technical spec.

• Experimented with the front/top microphones on devices.



- Experiments of researchers show that the modulation depth is hardware dependent
- The modulation depth at the prime fc is when recognition attacks are successful and 100% accurate
- The minimum depth for successful recognition attacks on each device is shown on table
- Modulation depth m is defined as m = M /A where A is the carrier amplitude, and M is the modulation amplitude
  - If m = 0.5, the carrier amplitude varies by 50% above (and below) its unmodulated level

IMPACT OF LANGUAGE

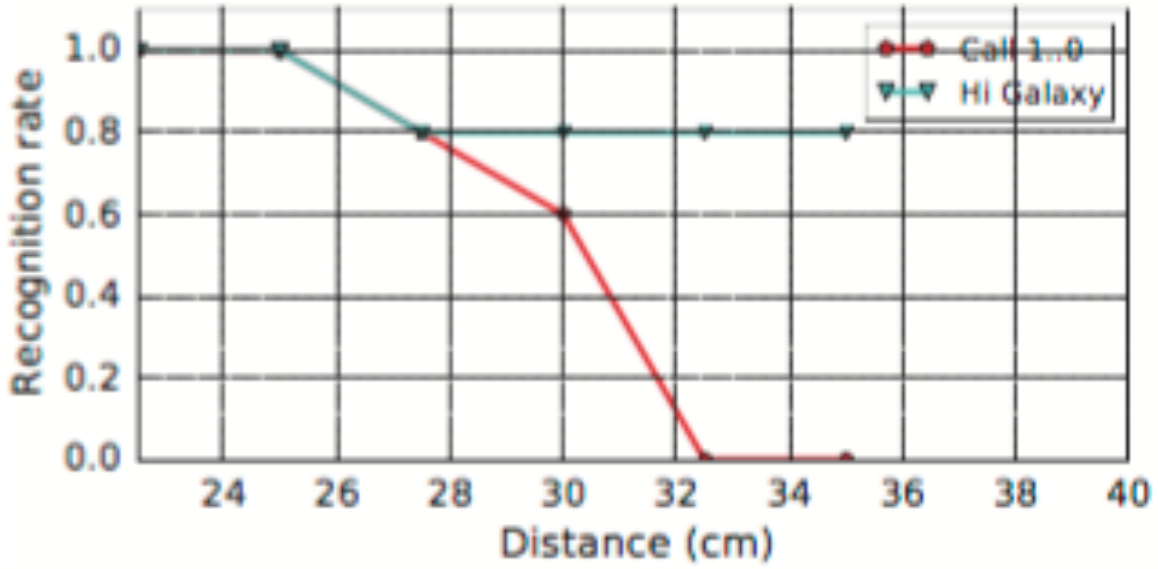


activating SR systems -- initiating to spy on the user -- denial of service

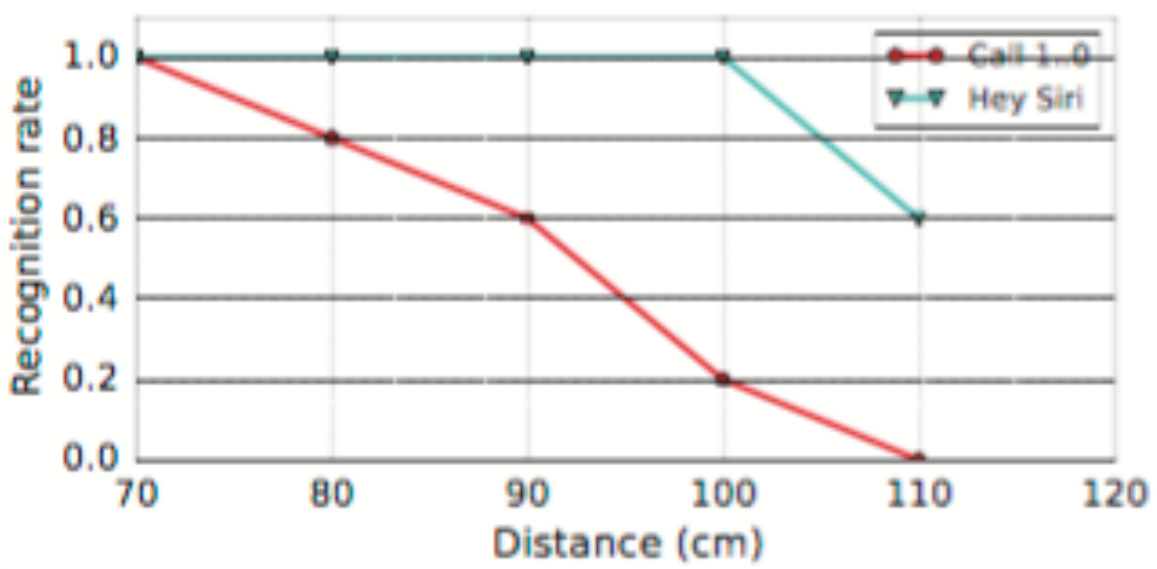
IMPACT OF BACKGROUND NOISE

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 DISTANCE



(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

- Hardware based
  - Microphone Enhancement
    - “a microphone shall be enhanced and designed to suppress any acoustic signals whose frequencies are in the ultrasound range.“
  - Inaudible Voice Command Cancellation
    - add a module prior to LPF to detect the modulated voice commands and cancel baseband

DEFENSES

- Software based
  - Use Supported Vector Machine to detect DolphinAttack
    - A supervised learning model using an algorithm to analyze data for classification

REALISTIC????

CONCLUSION

- Inaudible attacks to SR systems
- Dolphin Attack leverages amplitude modulation
- Hardware and software based defenses