DolphinAttack: Inaudible Voice Commands

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

Presenter: Huichen Li

This paper won the CCS 2017 Best Paper award

Speech Recognition Systems

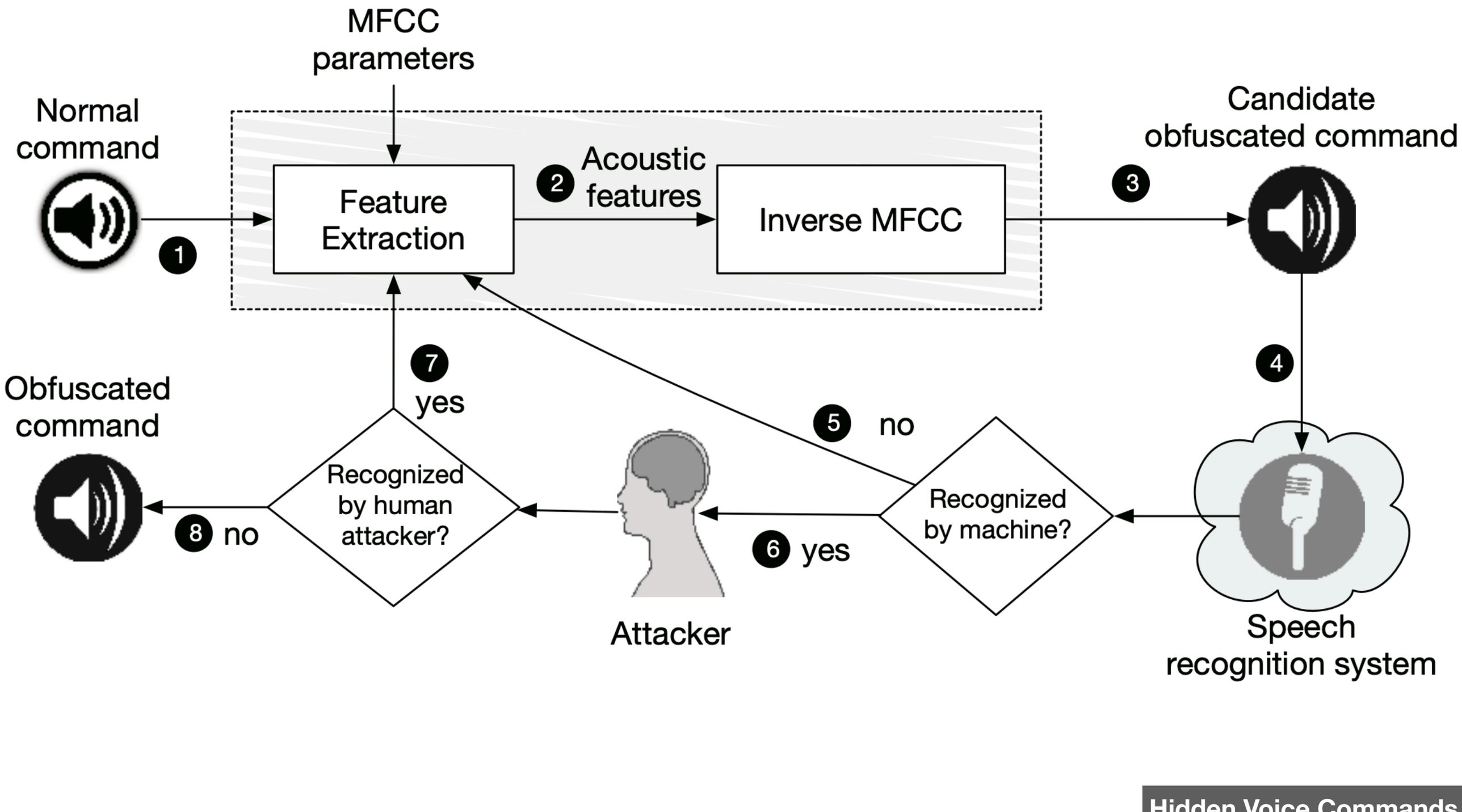
Apple Siri

Amazon Alexa

Google Now

Huawei HiVoice

Obfuscated Voice Commands



Hidden Voice Commands

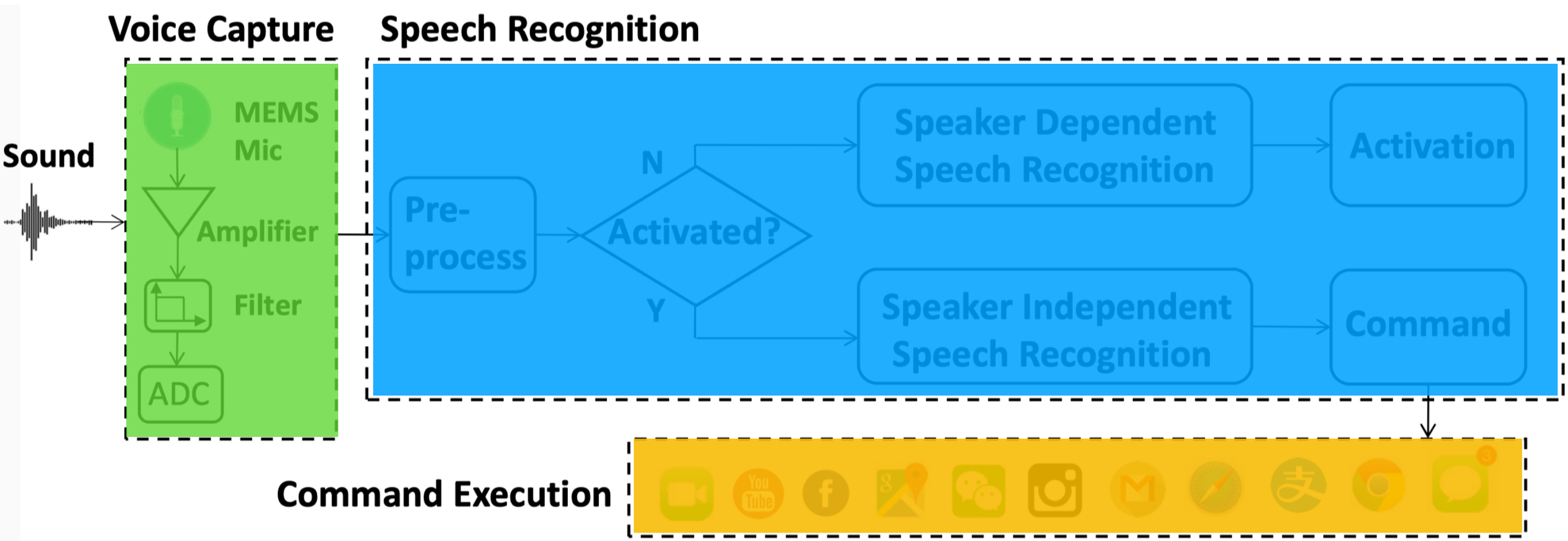
Threat Model

- **Inaudible** (with ultrasounds f > 20kHz)

Threat Model

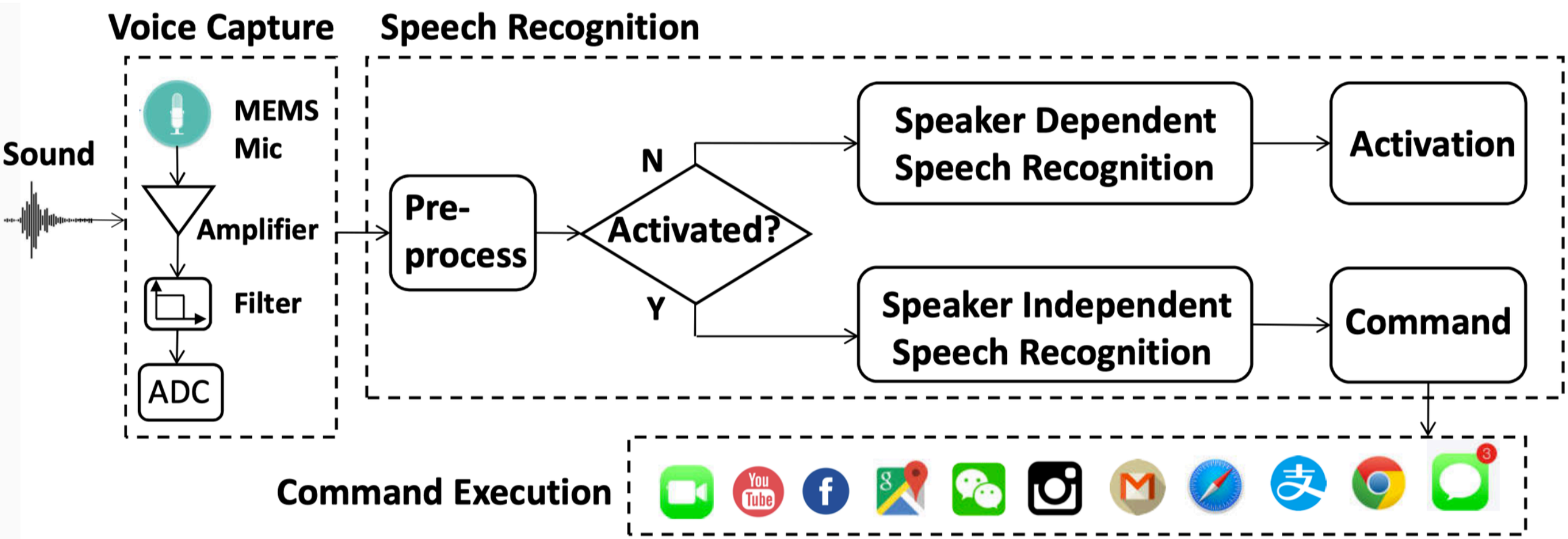
- **Inaudible** (with ultrasounds f > 20kHz)
- No owner interaction.
- **Whitebox**.
- **No (physical) target device access**.
- Attacker has required equipments (e.g. speakers for transmitting ultrasound near target devices).

Voice Controllable System

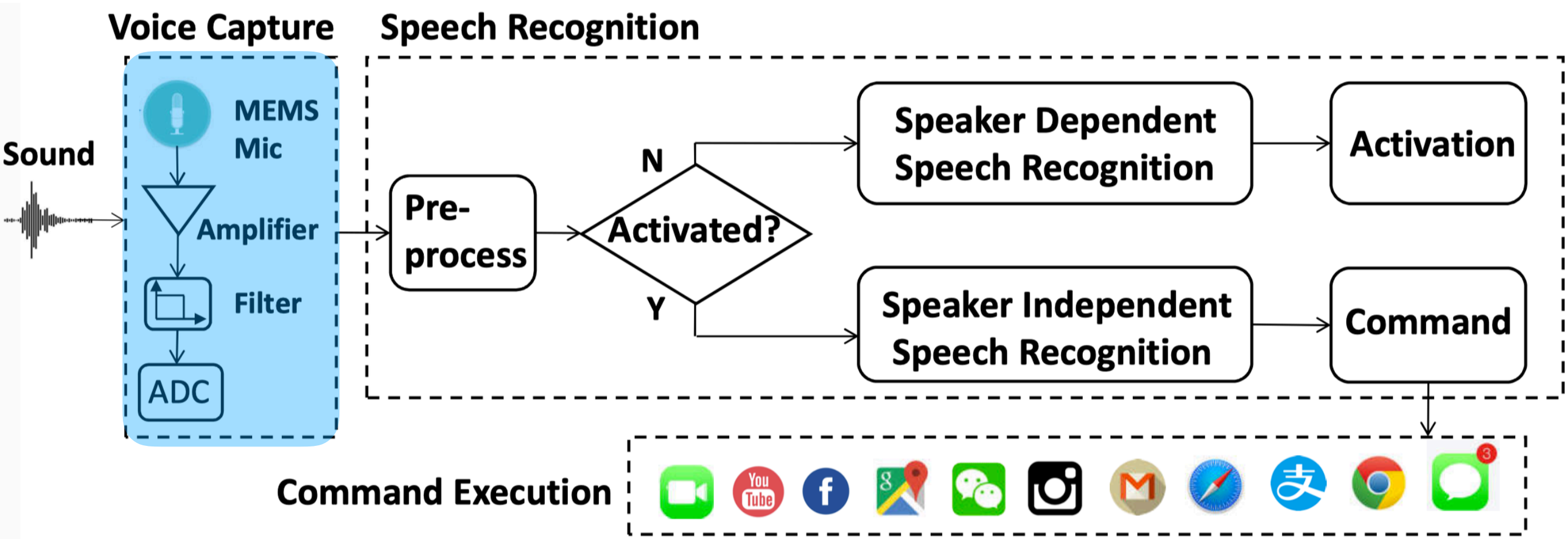


Q: Which parts of the VCS are most vulnerable?
(No known answer)

Voice Controllable System

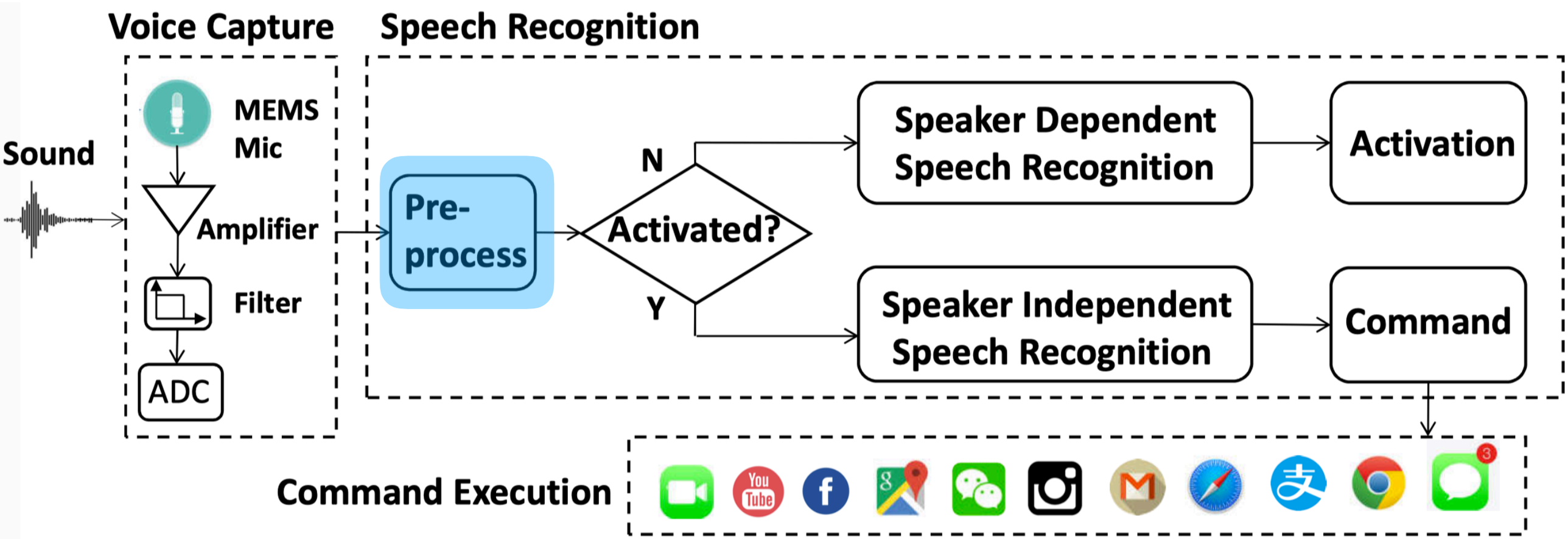


Voice Controllable System



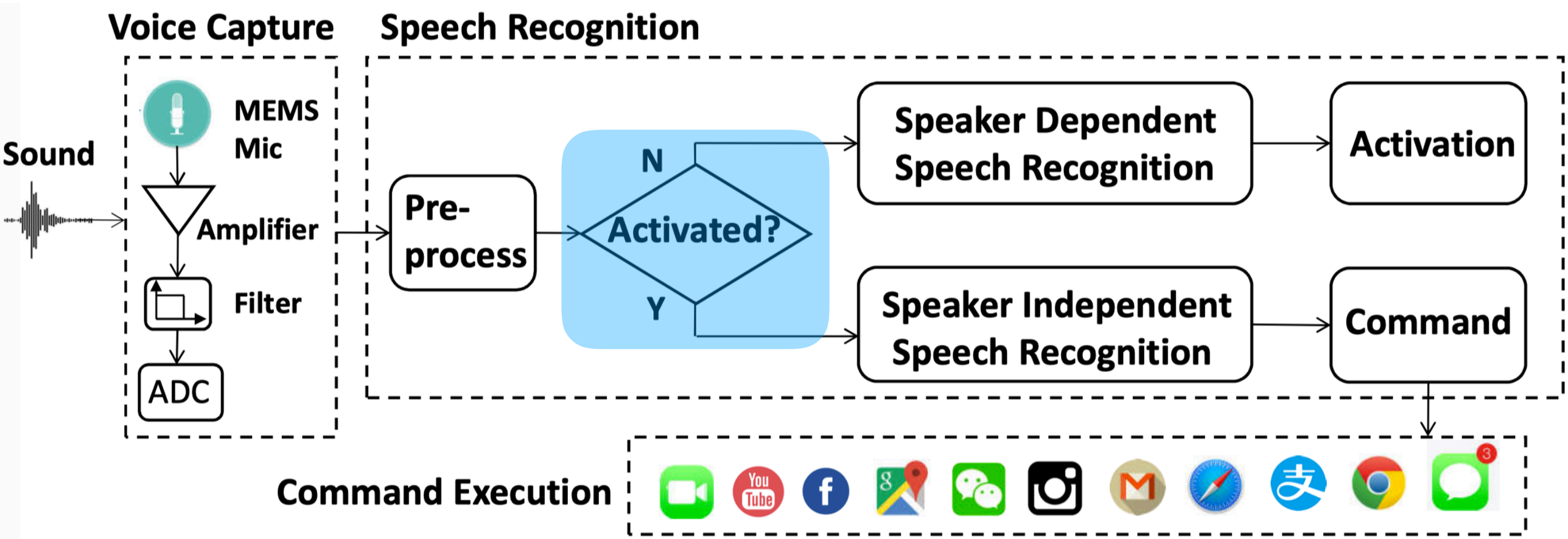
ambient voices:
recorded -> amplified -> filtered -> digitized

Voice Controllable System

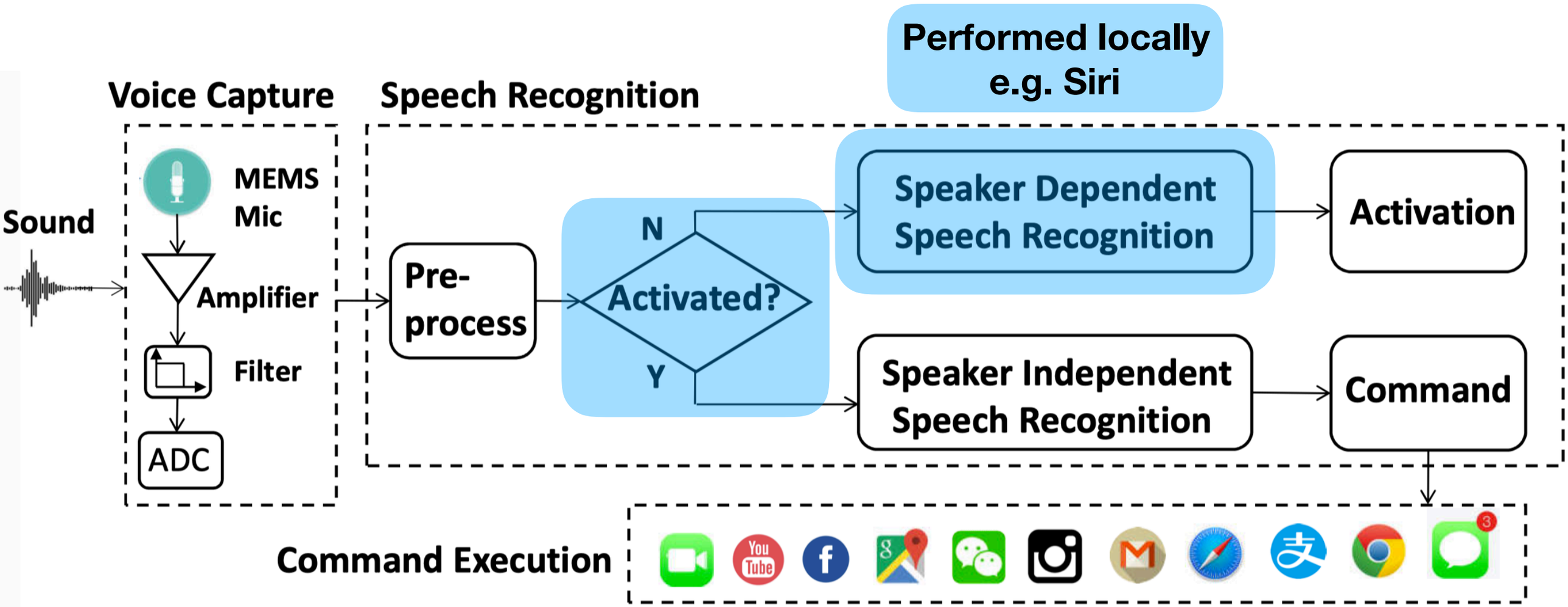


- remove frequencies that are beyond the audible sound range
- discard signal segments that contain sounds too weak to be identified

Voice Controllable System

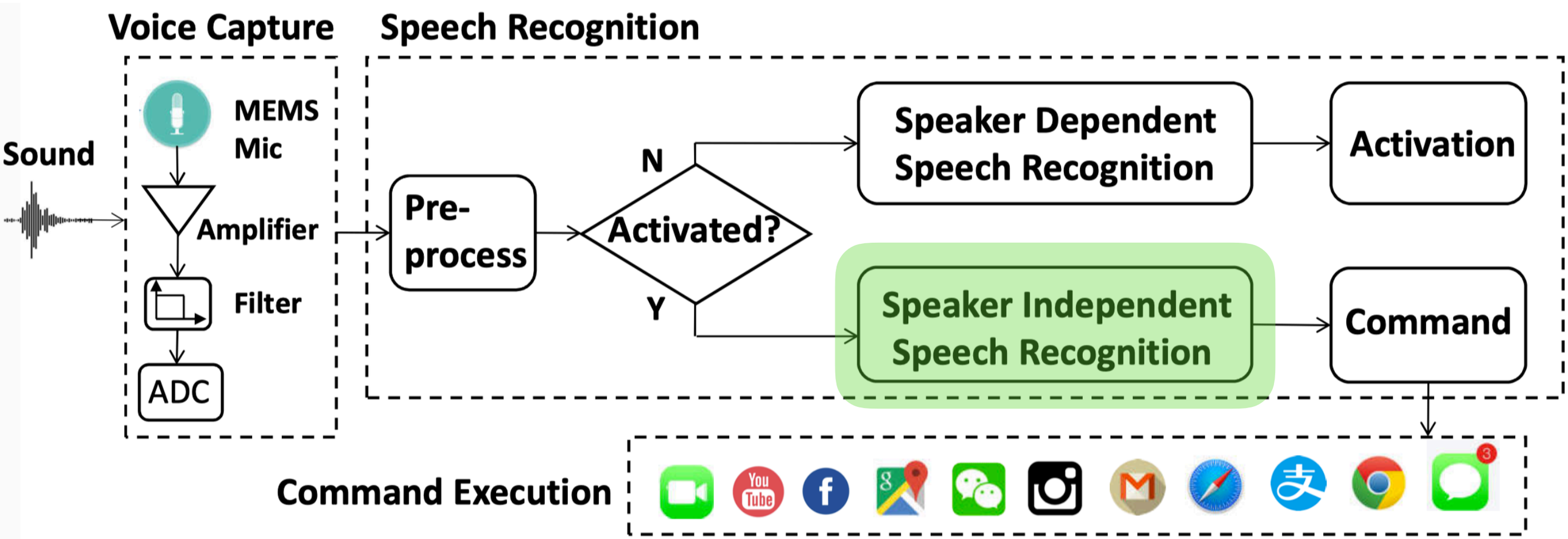


Voice Controllable System



- say pre-defined wake words
- press a special key

Voice Controllable System



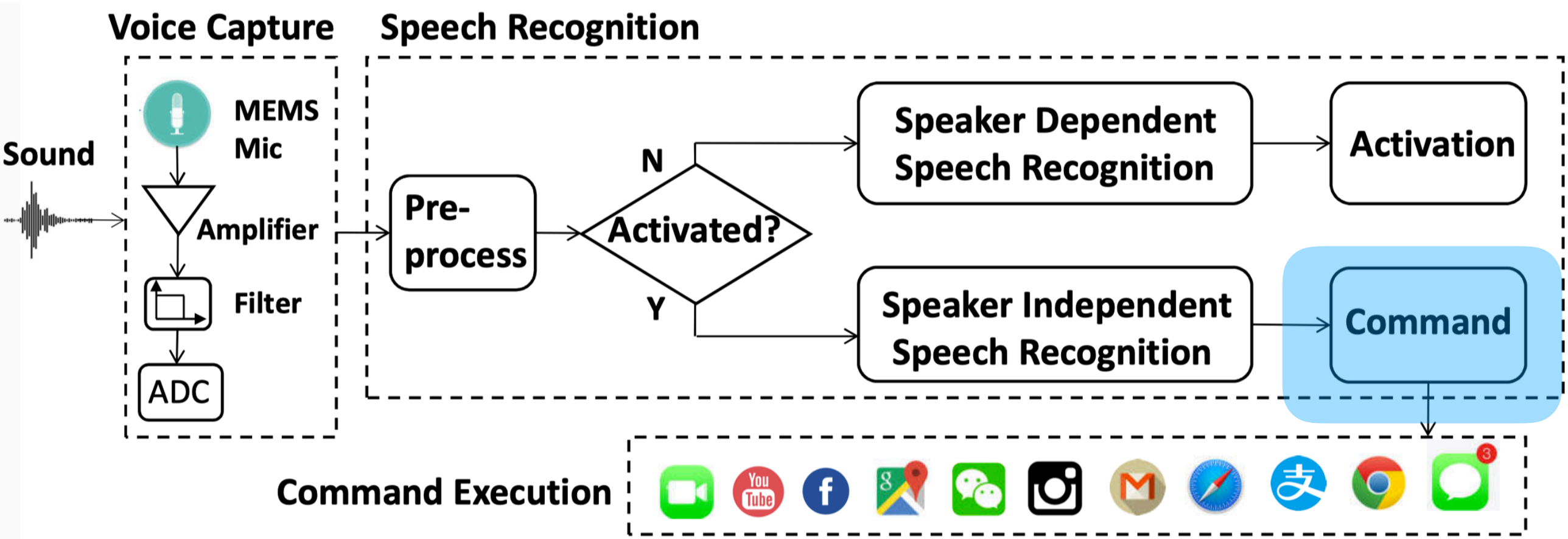
Via a cloud service

signals sent to servers -> extract features -> recognize commands

e.g. Mel-frequency cepstral coefficients(MFCC)

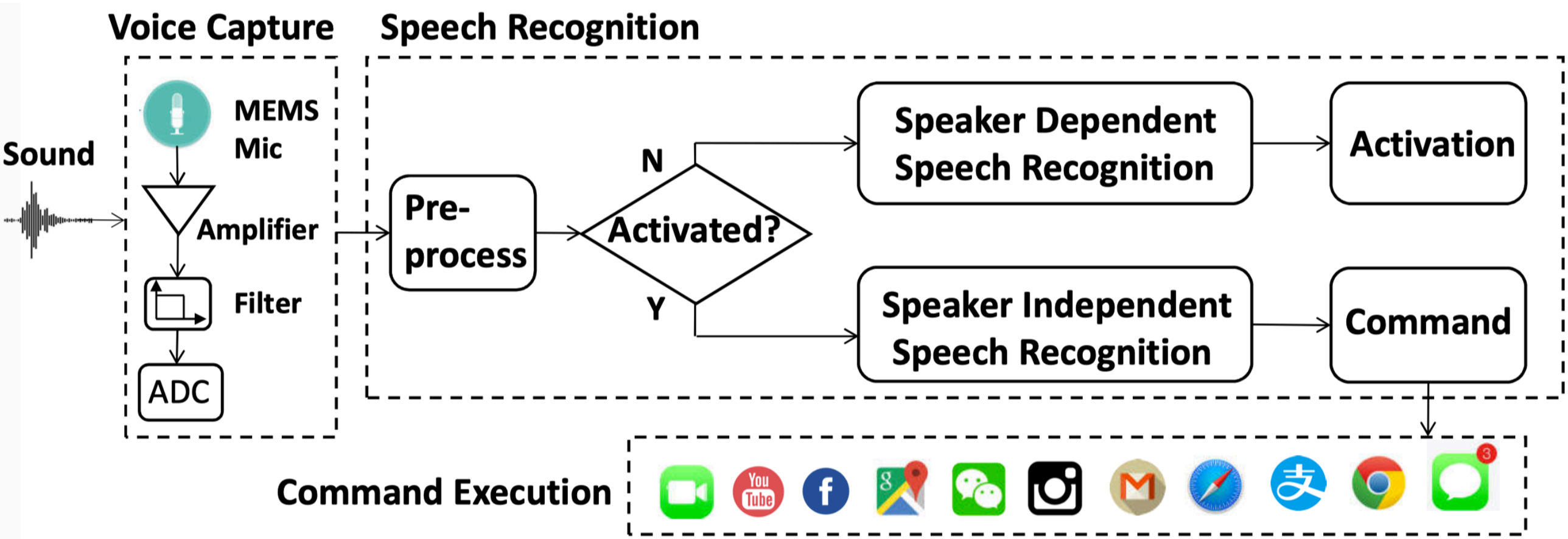
e.g. machine learning

Voice Controllable System



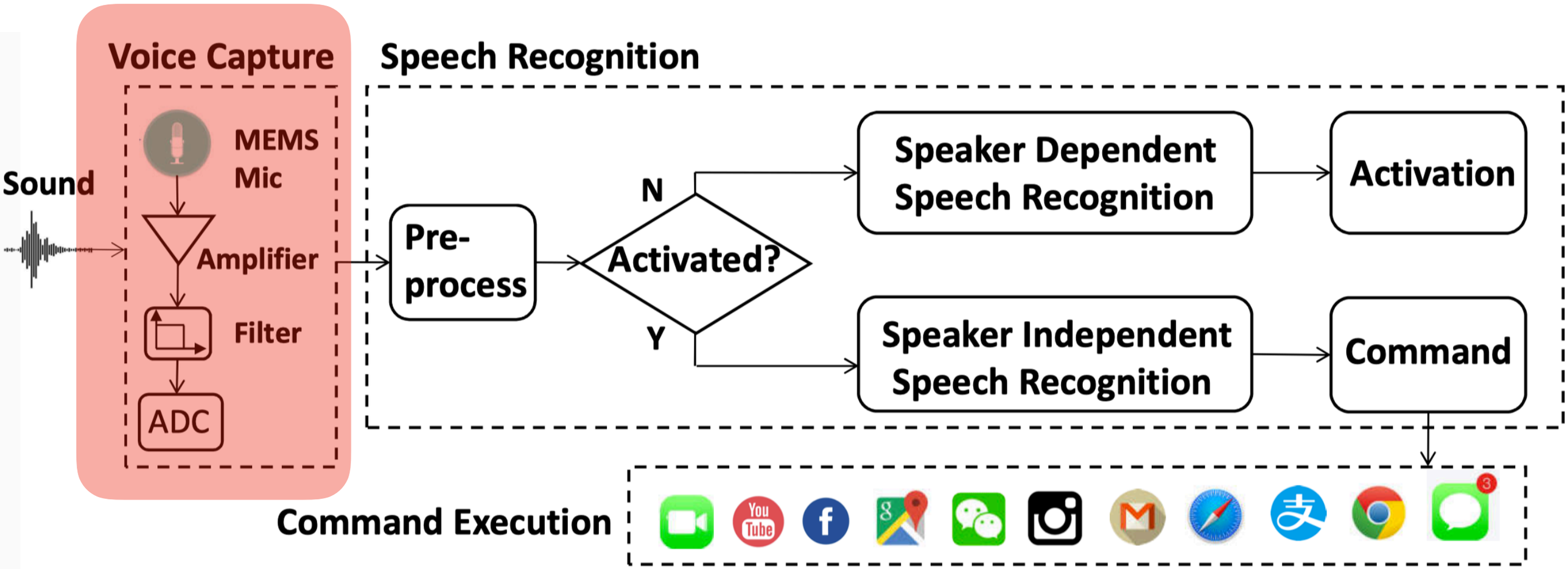
launch the corresponding application or execute an operation

Voice Controllable System



Q: Which parts of the VCS are most vulnerable?
(No known answer)
Take a guess!

Focus of Attack



Inaudible!

Doubts on Inaudible Voice Commands

- How can inaudible sounds be **audible** to devices?

  low-pass filters? low audio sampling rates?

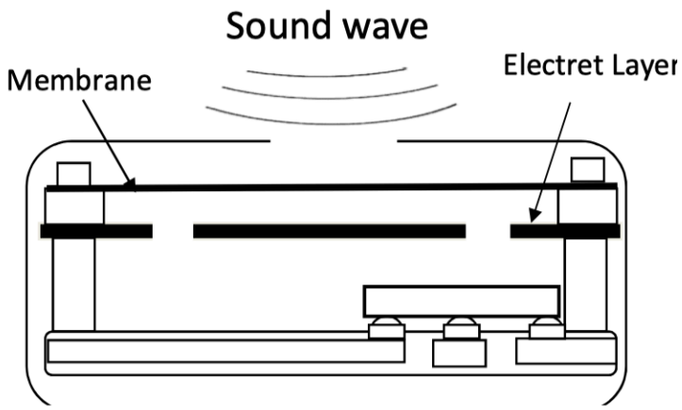
- How can inaudible sounds be **intelligible** to SR systems?

  SR systems do not recognize signals that do not match human tonal features?

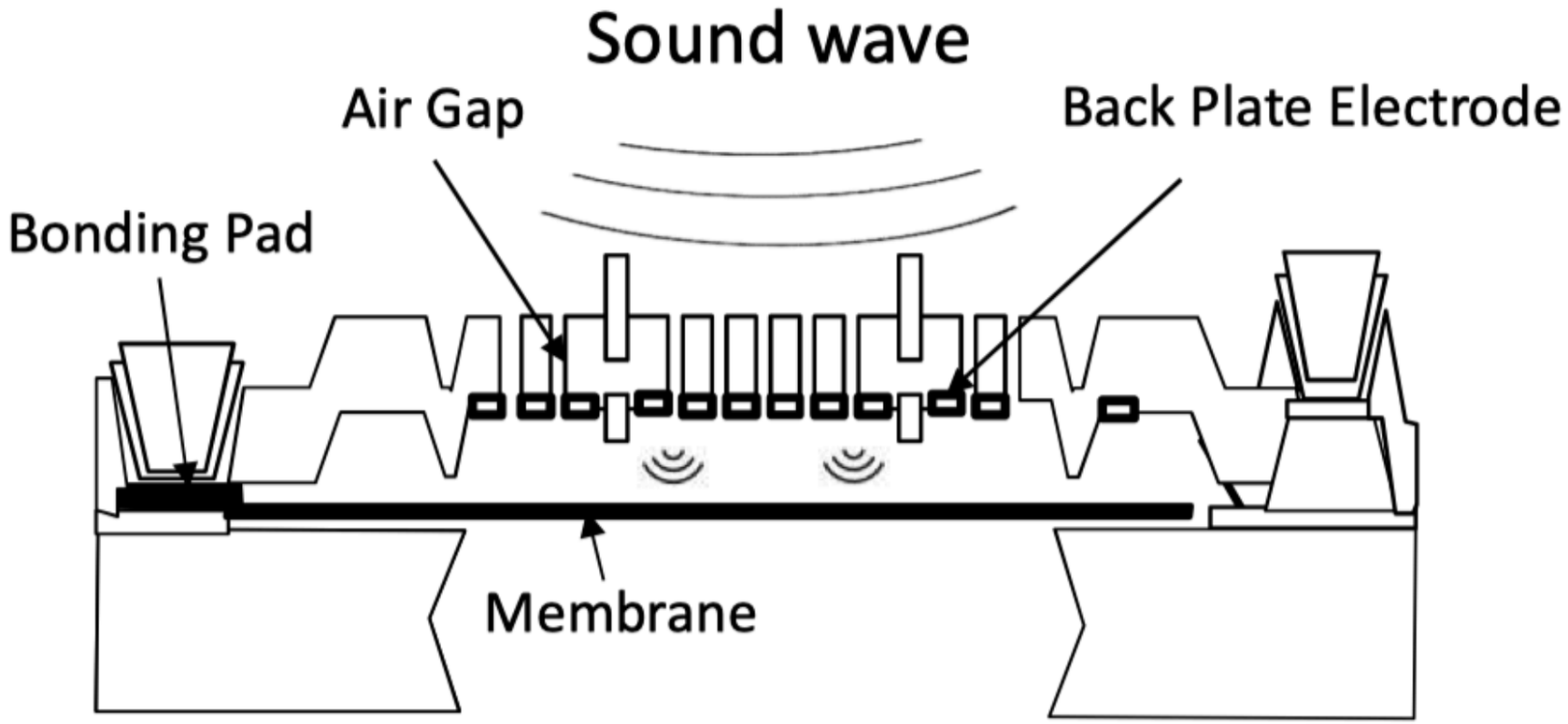
- How can inaudible sounds cause **unnoticed** security breach to VCS?

  speaker-dependent wake words?

Microphone



(a) Structure of ECM



(b) Structure of MEMS Microphone

Pros: - miniature package sizes
- low power consumption

air pressure change -> capacitive change -> AC signal

Nonlinearity of Microphone

in ultrasound bands f > 20kHz

$$s_{out}(t) = \sum_{i=1}^{\infty} A_i s^i(t) = A_1 s(t) + A_2 s^2(t) + A_3 s^3(t) + \ldots$$

$$\approx A_1 s(t) + A_2 s^2(t)$$

m(t): target voice signal $m(t) = \cos(2\pi f_m t)$

$$s_{in}(t) = m(t)\cos(2\pi f_c t) + \cos(2\pi f_c t)$$

LPF

Fourier Transformation

f_m $f_c - f_m, f_c + f_m,$ and f_c

$2(f_c - f_m), 2(f_c + f_m), 2f_c, 2f_c + f_m,$ and $2f_c - f_m$

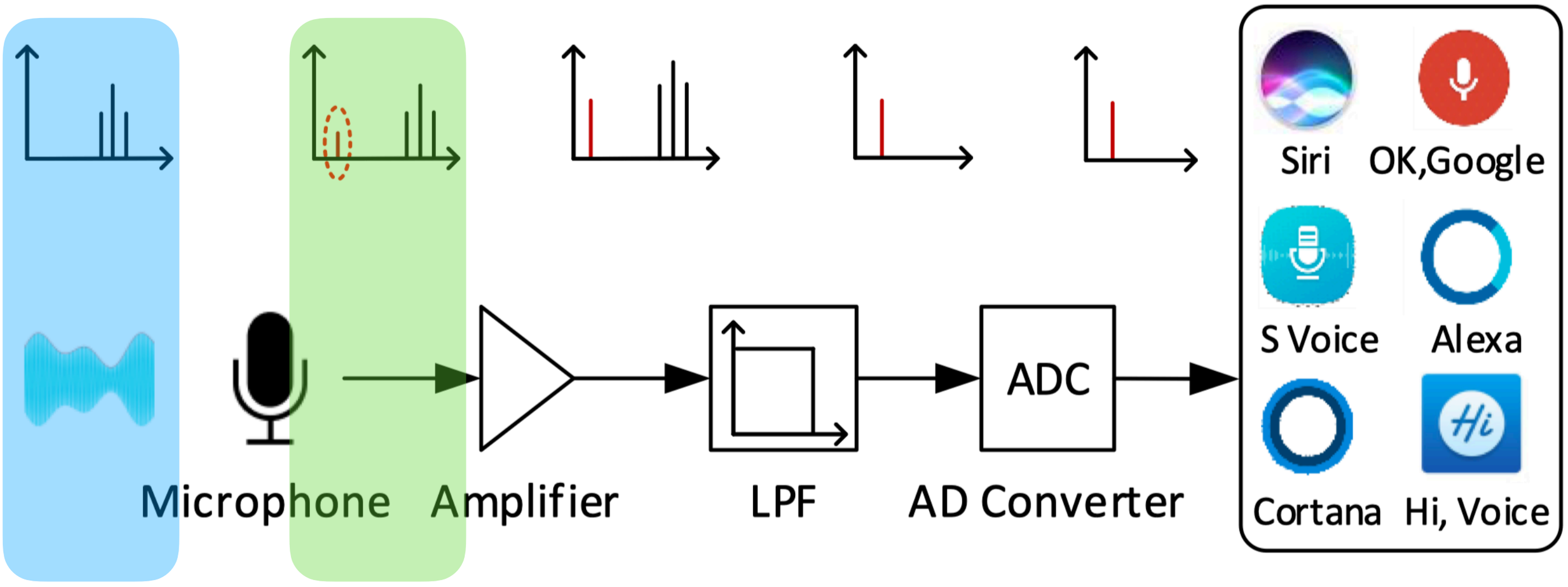
s1(t) = cos(2π f1 t) at frequency f1=38kHz
s2(t) = cos(2π f2 t) at frequency f2=40kHz
s_hi (t) = s1(t) + s2(t)

$$
\begin{aligned}
s_{out}(t) = &\quad A_1 s_{hi}(t) + A_2 s_{hi}^2(t) \\
= &\quad A_1(s_1(t)+s_2(t)) + A_2(s_1(t)+s_2(t))^2 \\
= &\quad A_1\cos(2\pi f_1 t) + A_1\cos(2\pi f_2 t) \\
+ &\quad A_2\cos^2(2\pi f_1 t) + A_2\cos^2(2\pi f_2 t) \\
+ &\quad 2A_2\cos(2\pi f_1 t)\cos(2\pi f_2 t)
\end{aligned}
$$

$$
\begin{aligned}
s_{out}(t) = &\quad A_1\cos(2\pi f_1 t) + A_1\cos(2\pi f_2 t) \\
+ &\quad A_2 + 0.5A_2\cos(2\pi 2 f_1 t) + 0.5A_2\cos(2\pi 2 f_2 t) \\
+ &\quad A_2\cos(2\pi(f_1+f_2)t) + A_2\cos(2\pi(f_2-f_1)t)
\end{aligned}
$$

$$
s_{low}(t) = A_2 + A_2\cos(2\pi(f_2-f_1)t)
$$

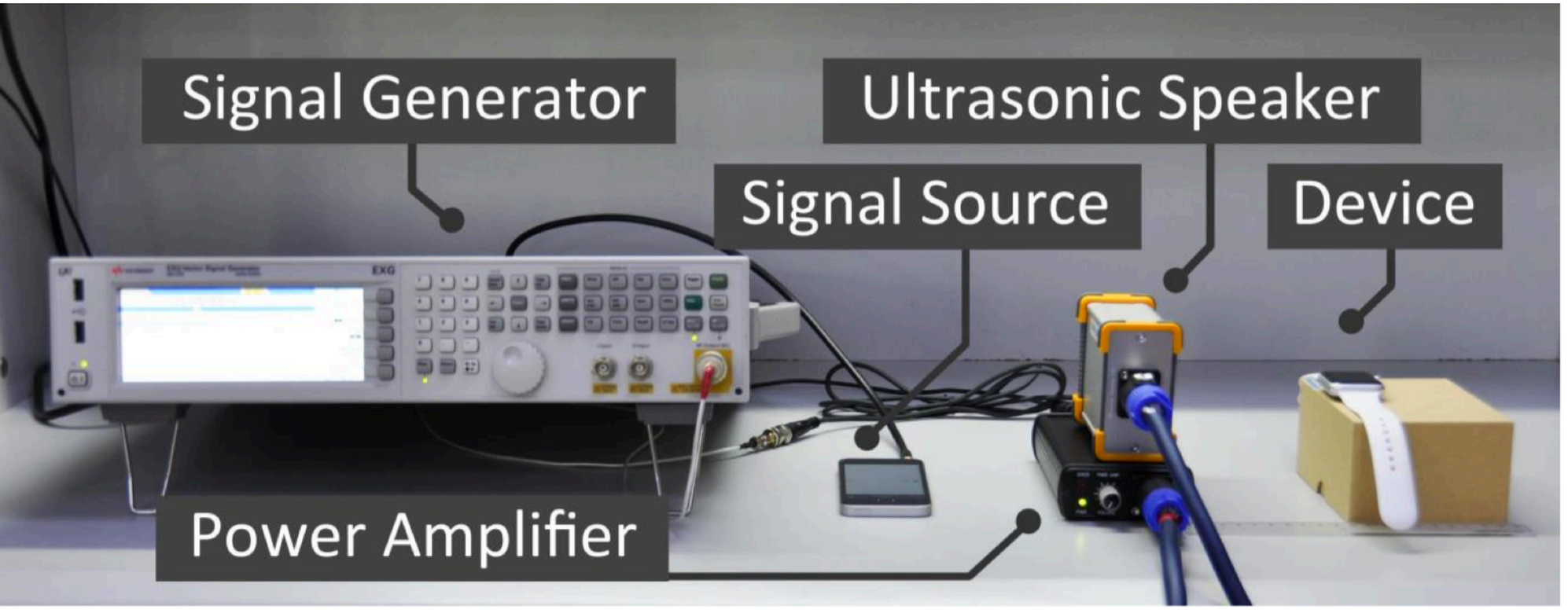
Modulated Tone Traversing Voice Capture Device

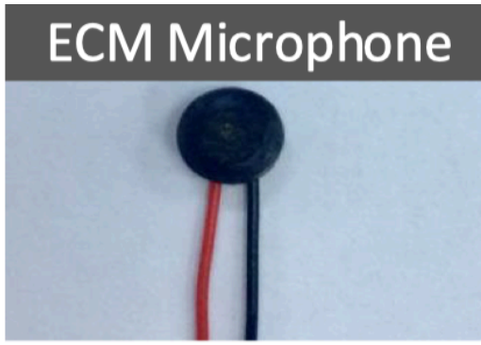


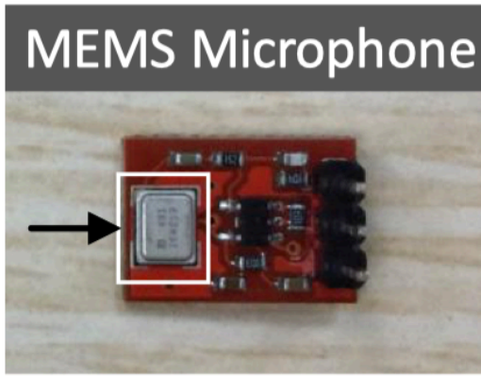
Nonlinearity Evaluation: Questions

- Will the demodulation work well in practice?
- Will the demodulated voice signal remain similar to the original one?

Nonlinearity Evaluation: Experimental Setup



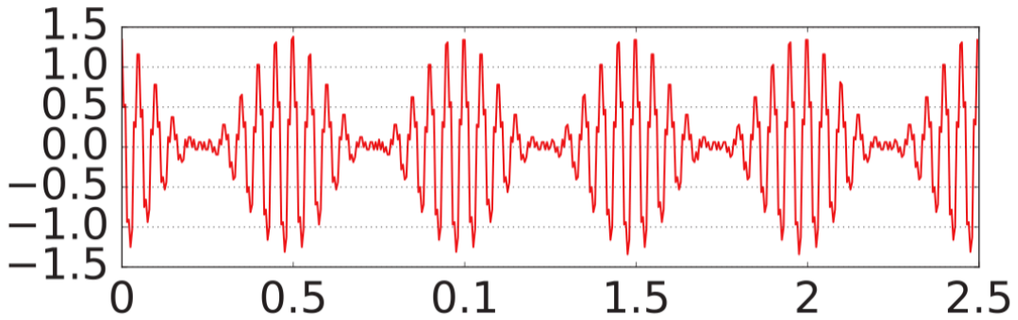


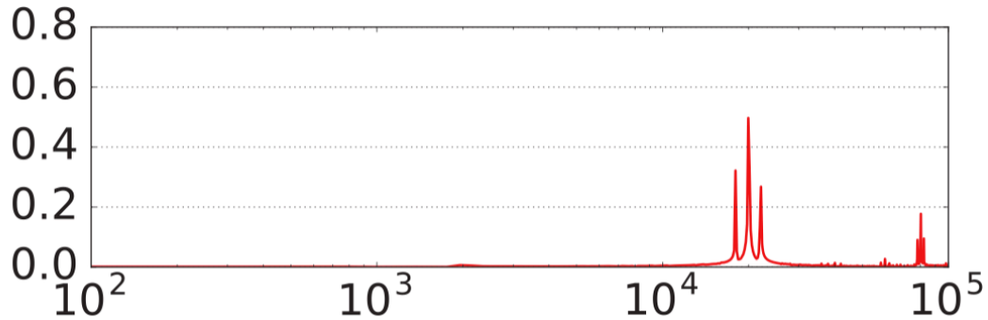


iPhone SE -> vector signal generator -> power amplifier -> ultrasonic speaker

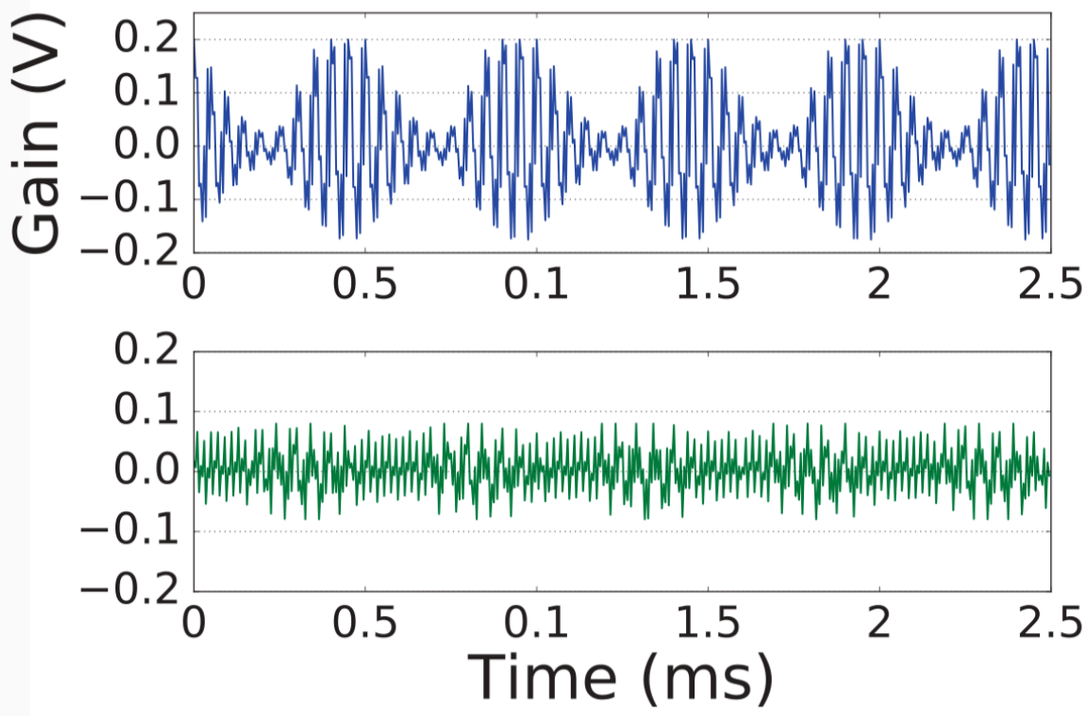
baseband signal -> modulated onto a carrier -> amplified -> transmitted

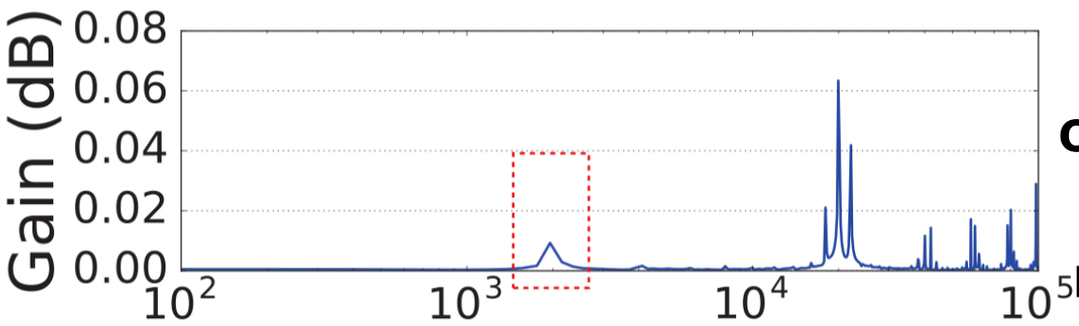
Nonlinearity Evaluation: Single Tone Results



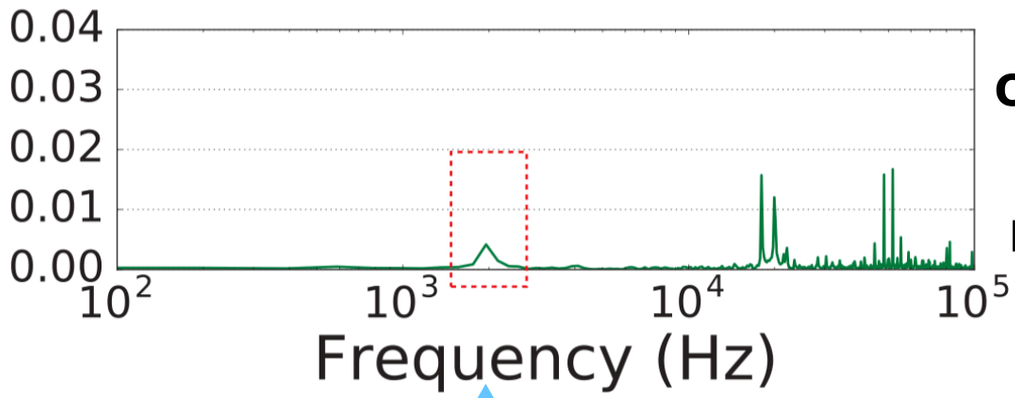


original





output signal of MEMS microphone



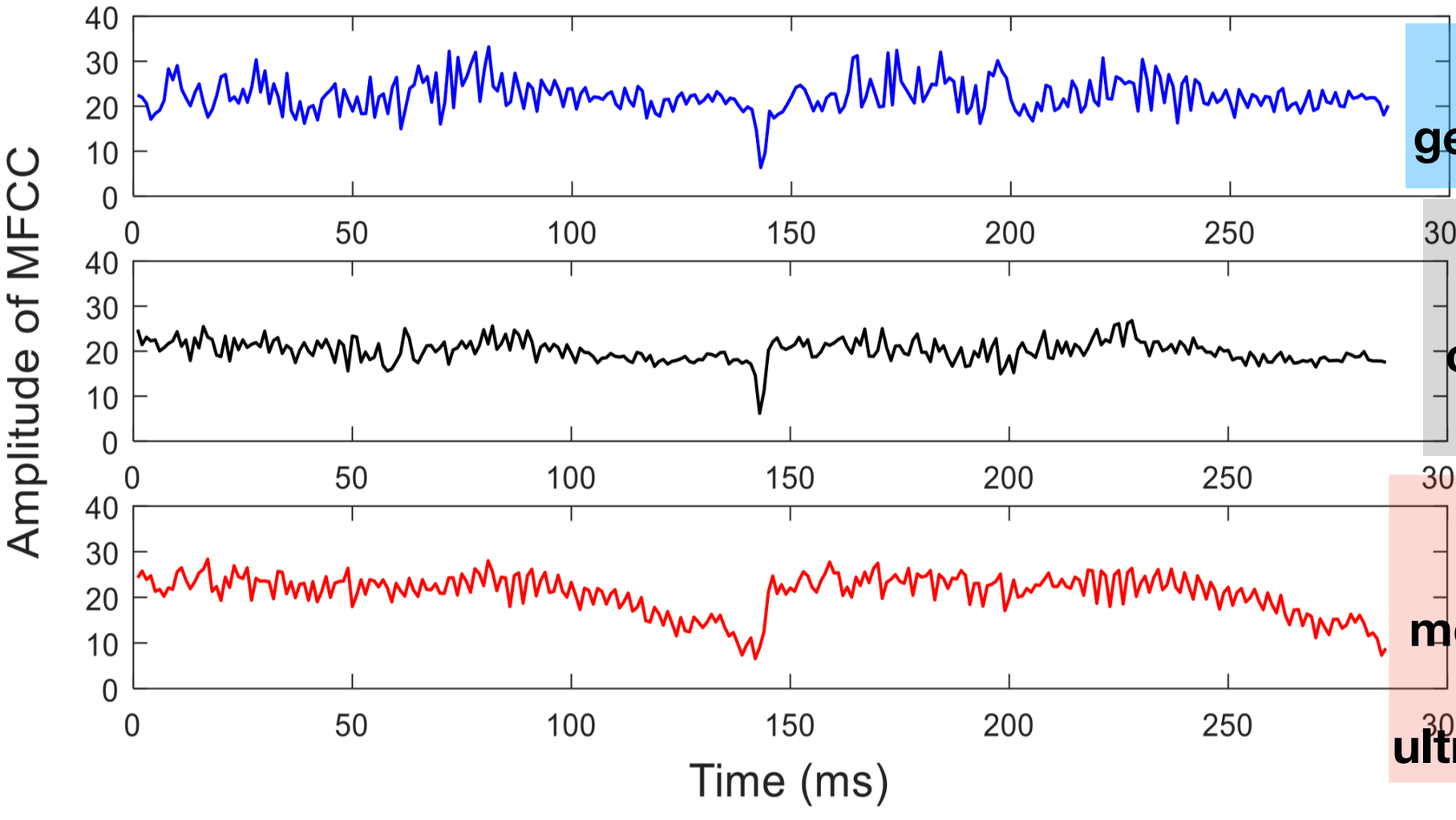
output signal of ECM microphone

20 kHz carrier

2 kHz baseband

Demodulation successful!

Nonlinearity Evaluation: Voices Results



MCD between original and recorded

original TTS generated voice

recorded as the original voice is played

3.1

recorded as the modulated voice is played by ultrasonic speaker

7.6

Similar!

Mel-Cepstral Distortion (MCD) quantifies distortion between two MFCCs two voices are considered to be acceptable to voice recognition systems if their MCD values are smaller than 8

Attack Design

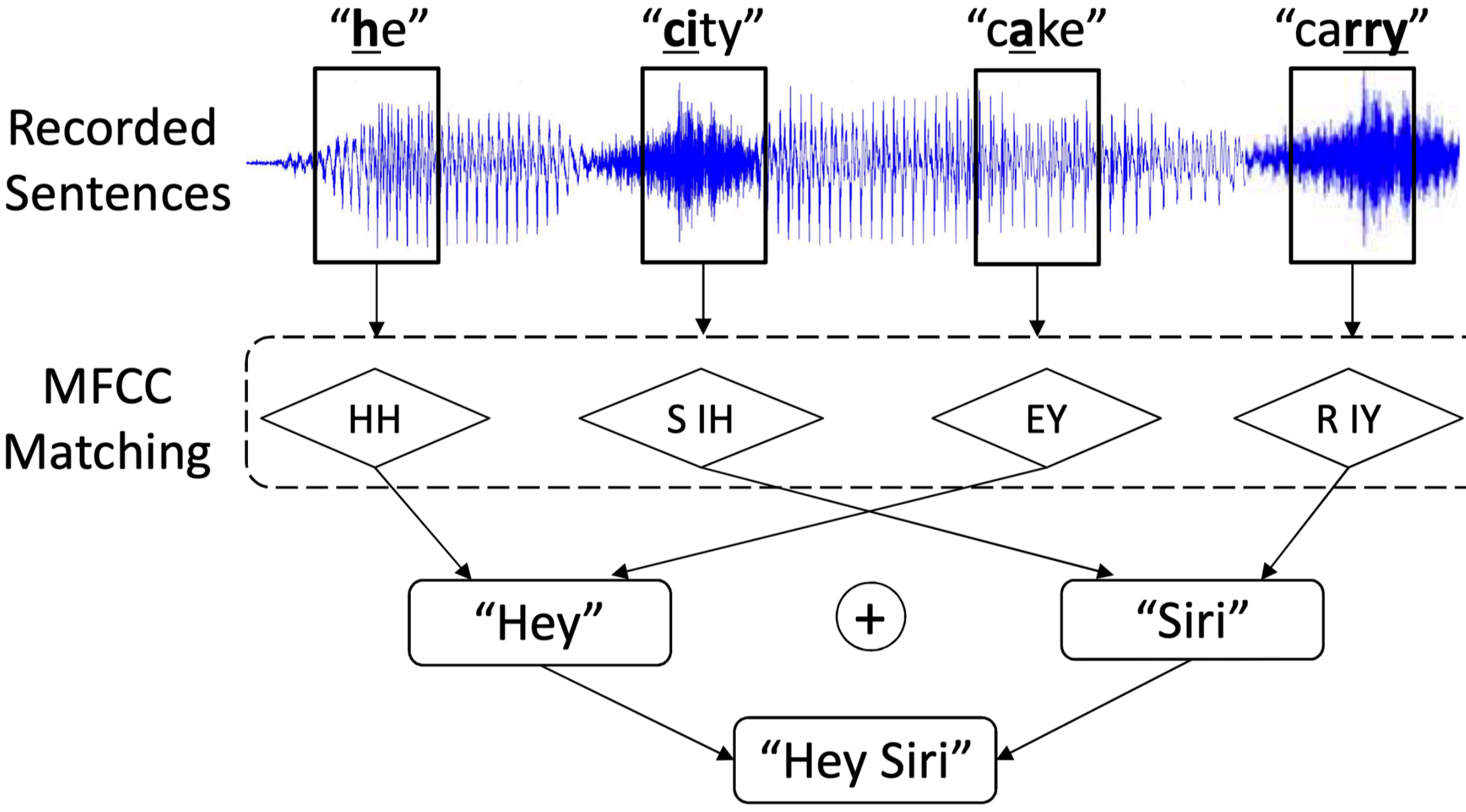
- Generate voice commands
- Modulate baseband signals
- Launch attack with a portable transmitter

Activation Voice Commands Generation: 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

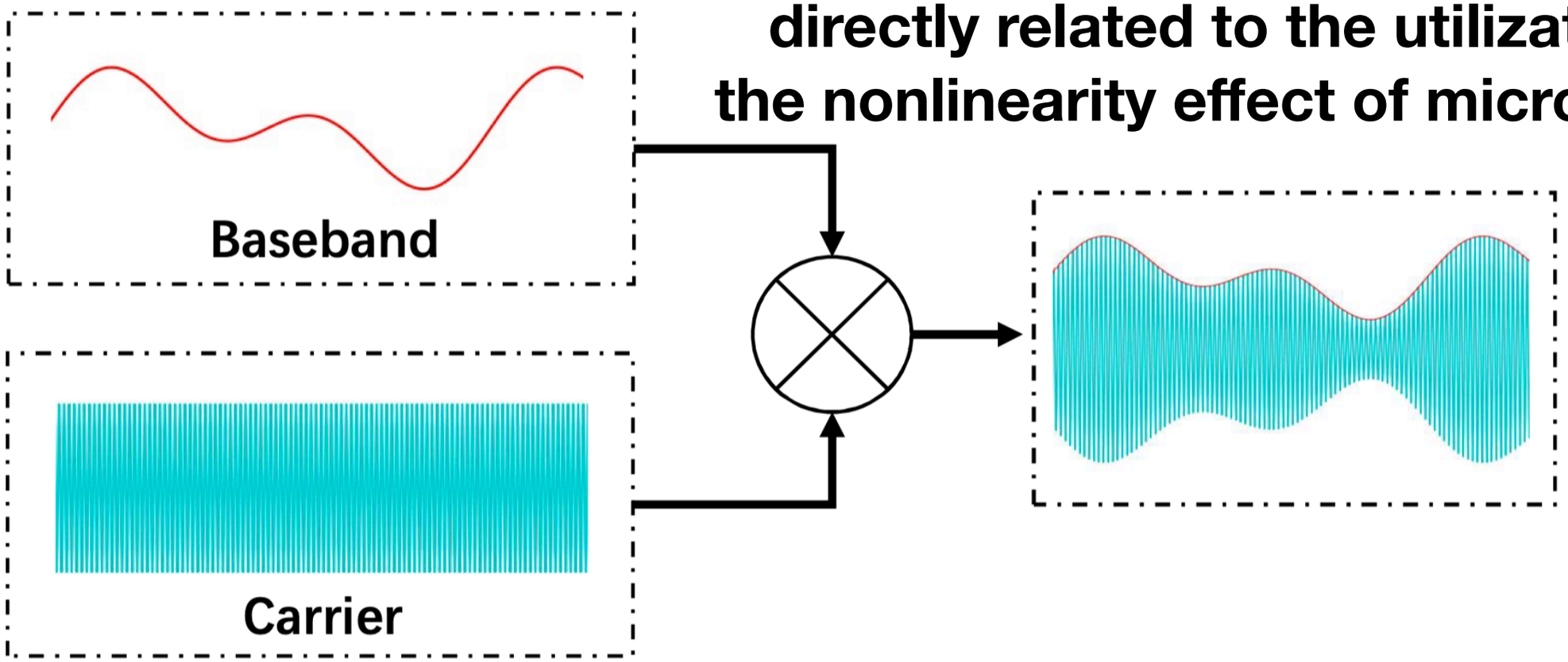
Siri is trained with Google TTS

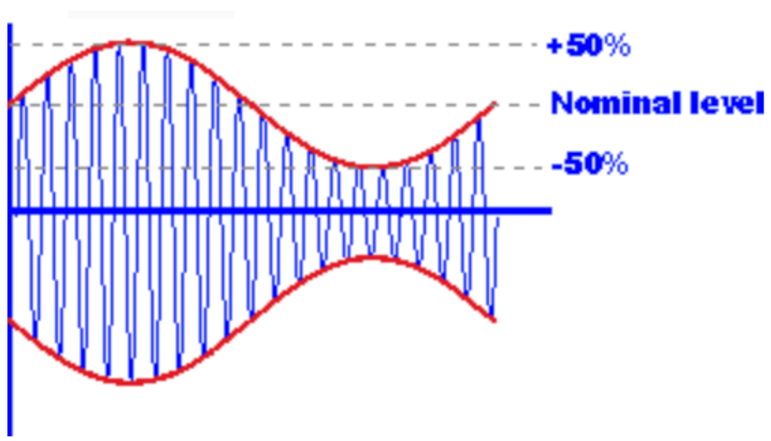
Activation Voice Commands Generation: Concatenative



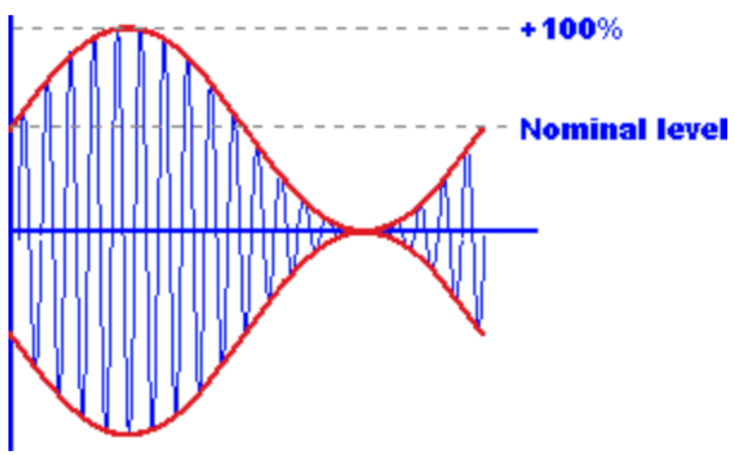
Amplitude Modulation (AM): Depth (index)

directly related to the utilization of the nonlinearity effect of microphones

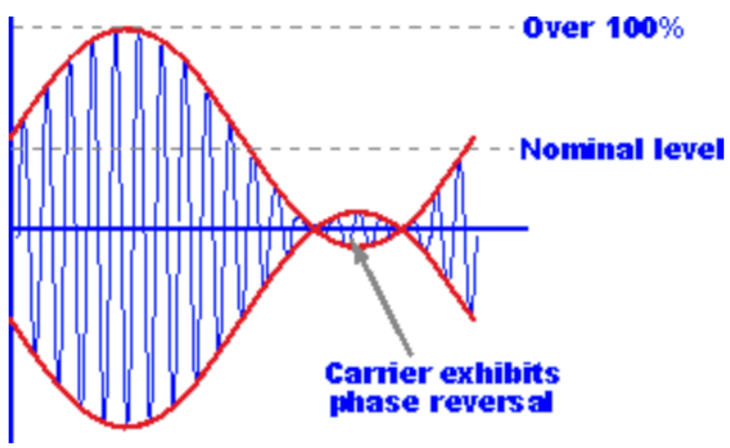




Amplitude modulated index of 0.5

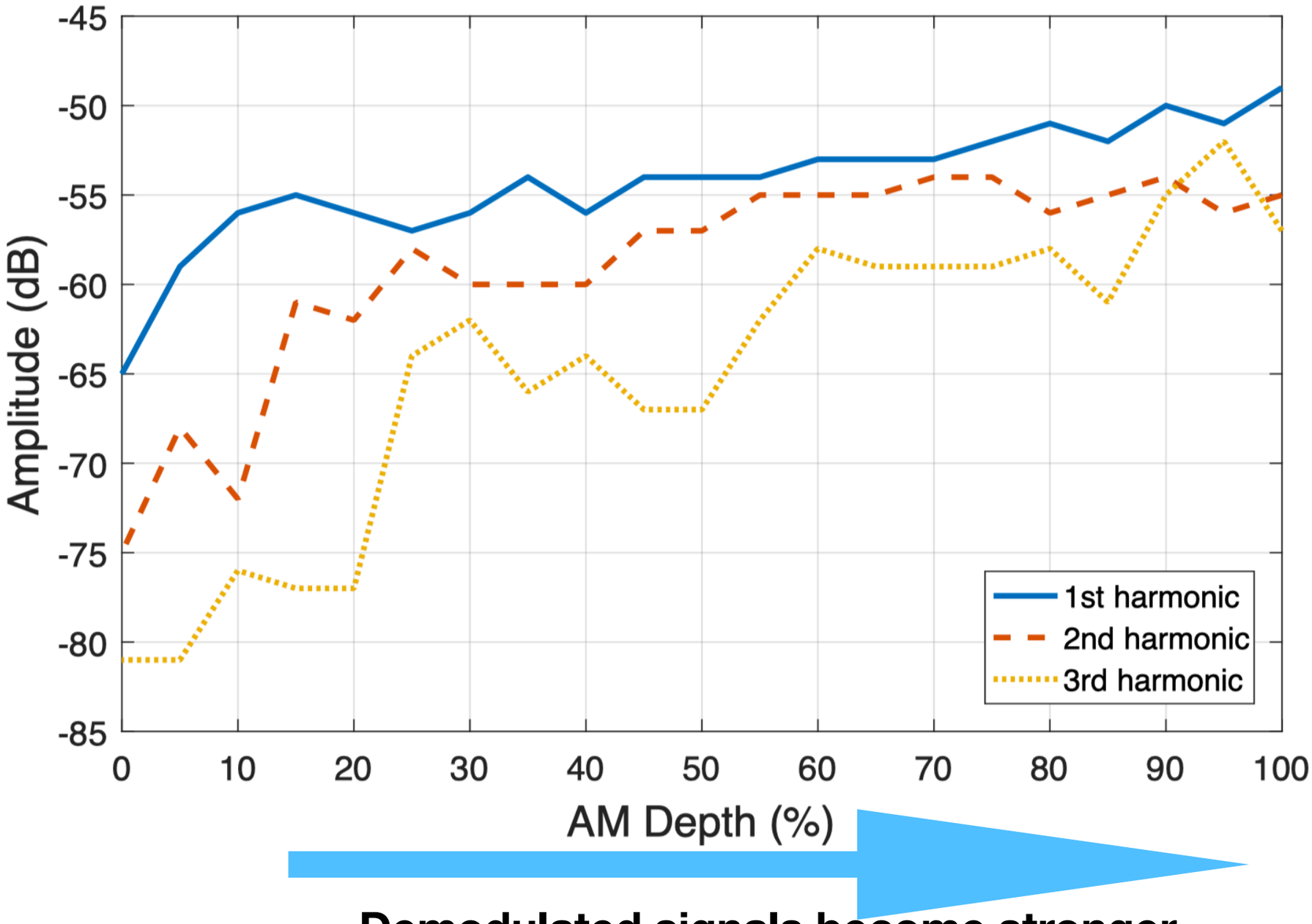


Amplitude modulated index of 1.0



Amplitude modulated index of more than 1.0
i.e. over-modulated

Analysis: Modulation Depth



Demodulated signals become stronger

Signal-to-noise ratio and the attack success rate get higher

Amplitude Modulation (AM): Carrier Frequency f

- Factors for choosing f:
    - frequency range of ultrasounds
    - bandwidth of the baseband signal
    - cut-off frequency of the low pass filter
    - frequency response of the microphone on the VCS
    - frequency response of the attacking speaker

Amplitude Modulation (AM): Carrier Frequency f

- Factors for choosing f:
  - frequency range of ultrasounds
  - bandwidth of the baseband signal
  - cut-off frequency of the low pass filter
  - frequency response of the microphone on the VCS
  - frequency response of the attacking speaker

Inaudibility: lowest frequency > 20 kHz

Amplitude Modulation (AM): Carrier Frequency f

- Factors for choosing f:
  - frequency range of ultrasounds
  - bandwidth of the baseband signal
  - cut-off frequency of the low pass filter
  - frequency response of the microphone on the VCS
  - frequency response of the attacking speaker

Inaudibility: lowest frequency > 20 kHz

w: frequency range of voice command

Amplitude Modulation (AM): Carrier Frequency f

- Factors for choosing f:
  - frequency range of ultrasounds
  - bandwidth of the baseband signal
  - cut-off frequency of the low pass filter
  - frequency response of the microphone on the VCS
  - frequency response of the attacking speaker

Inaudibility: lowest frequency > 20 kHz

w: frequency range of voice command

f - w > 20 kHz

Amplitude Modulation (AM): Carrier Frequency f

- Factors for choosing f:
  - frequency range of ultrasounds
  - bandwidth of the baseband signal
  - cut-off frequency of the low pass filter
  - frequency response of the microphone on the VCS
  - frequency response of the attacking speaker

Inaudibility: lowest frequency > 20 kHz

w: frequency range of voice command

f - w > 20 kHz

otherwise carrier will not be filtered.

Amplitude Modulation (AM): Carrier Frequency f

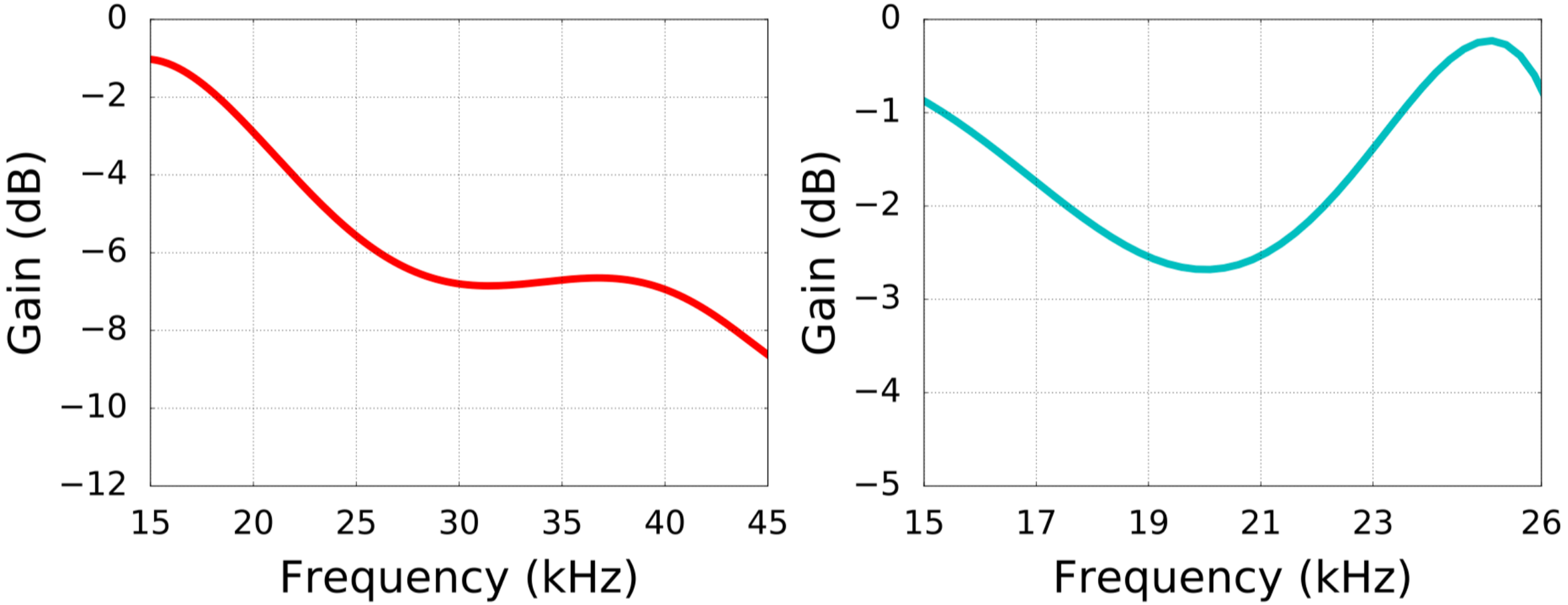
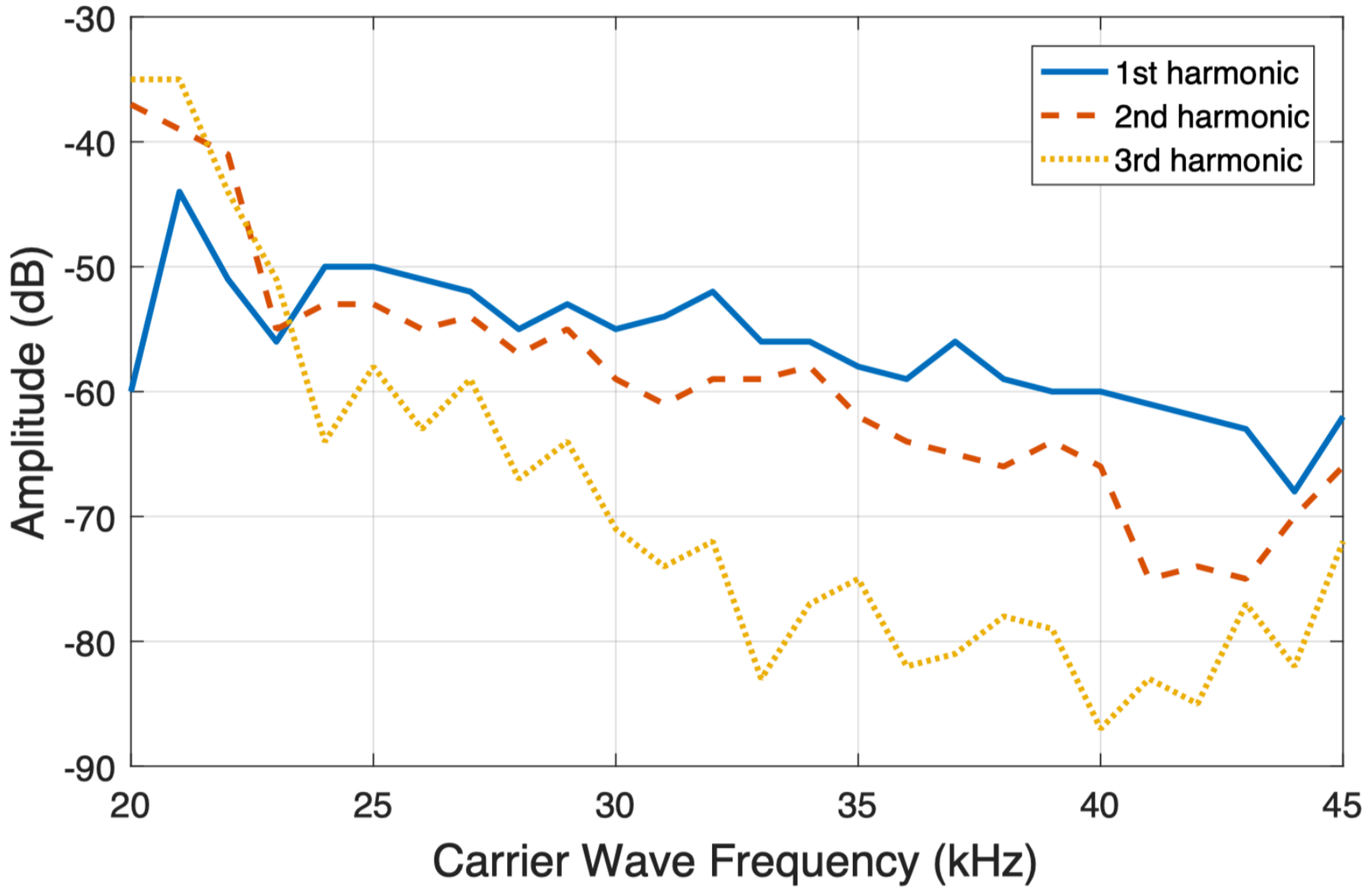


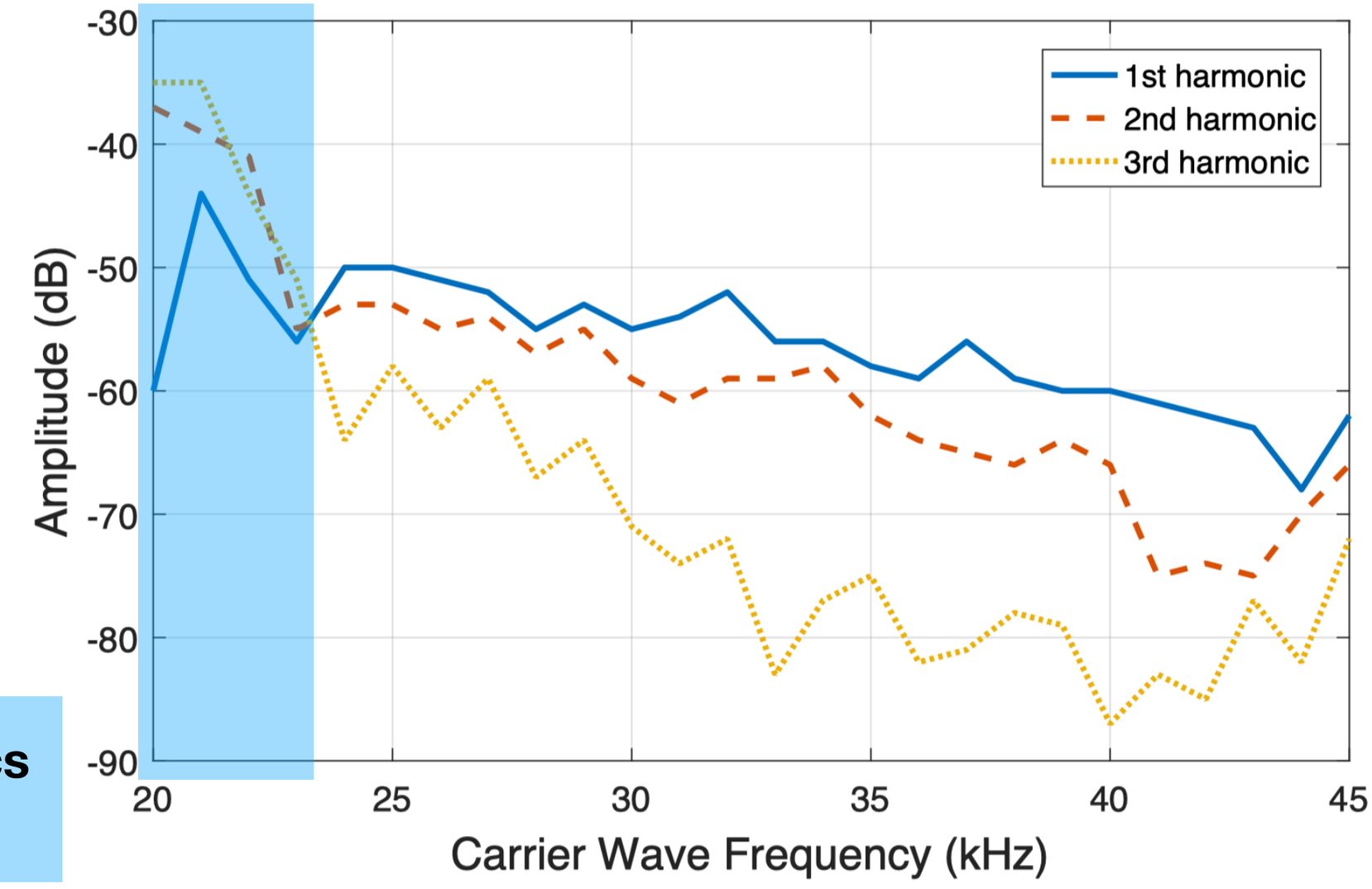
Figure 10: The frequency responses of the ADMP401 MEMS microphone (left) and the Samsung Galaxy S6 Edge speaker (right).

Analysis: Carrier Wave Frequency



400 Hz baseband and higher order harmonics

Analysis: Carrier Wave Frequency



amplitude of the harmonics larger than baseband

Unacceptable to SR systems!

400 Hz baseband and higher order harmonics

Amplitude Modulation (AM): Voice Selection

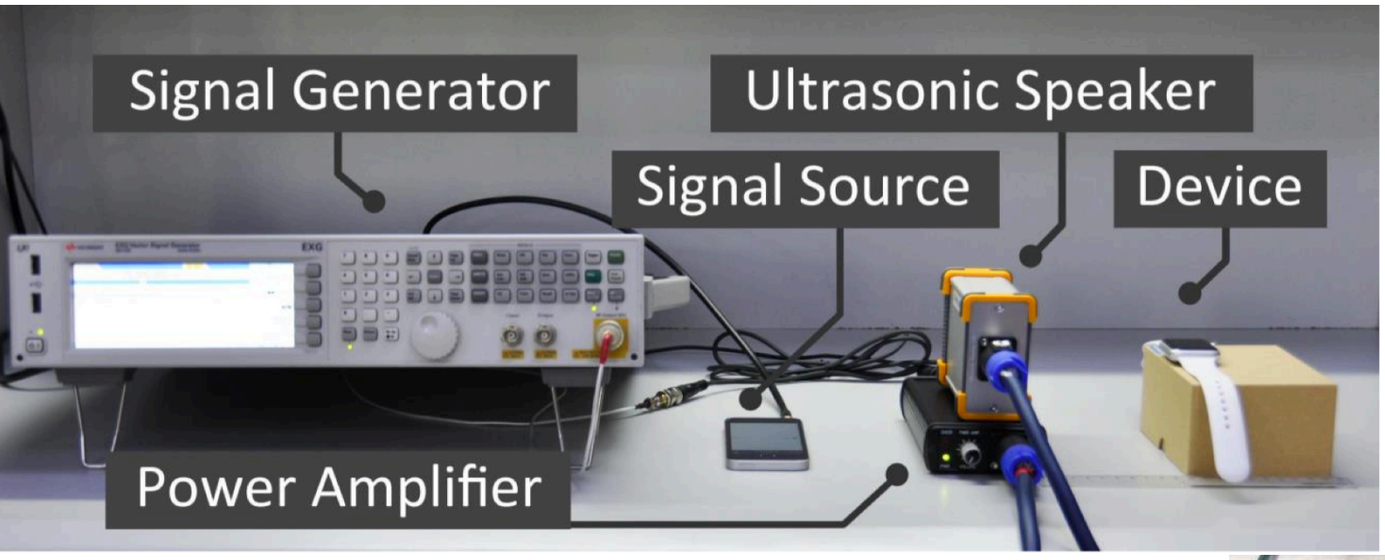
f - w > 20 kHz

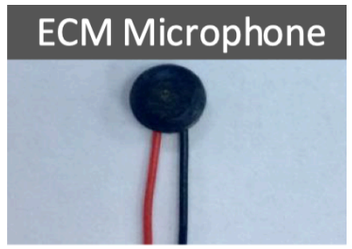
- Various voices map to various baseband frequency ranges.

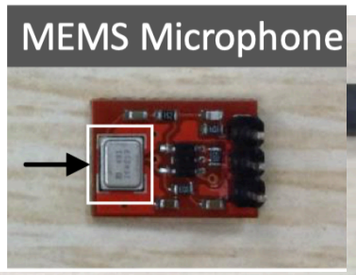
- A voice with a **small** bandwidth shall be selected to create baseband voice signals

Voice Commands Transmitter

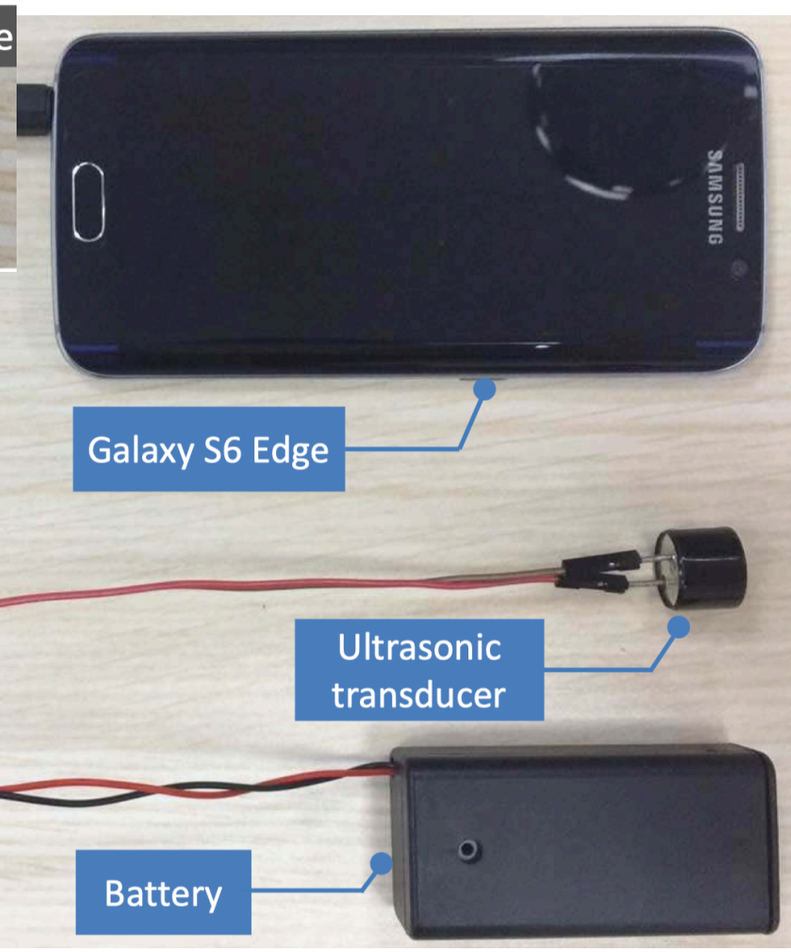
Powerful transmitter:
driven by a dedicated signal generator







Portable transmitter:
driven by a smartphone



Experimental Goal

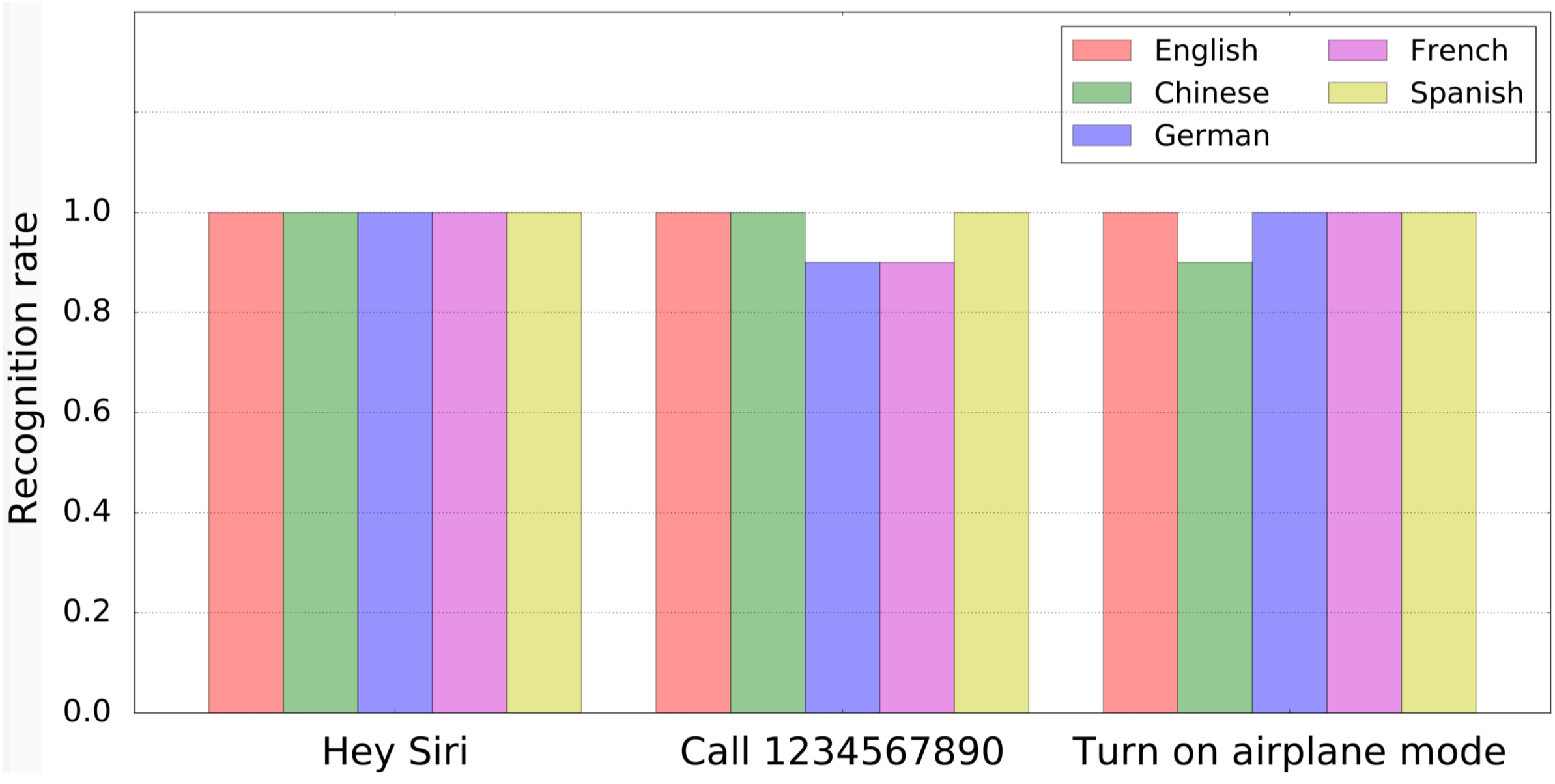
- Examining the feasibility of attacks.
- Quantifying the parameters in tuning a successfully attack.
- Measuring the attack performance.

Feasibility Experiments: Device/System & Commands

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.

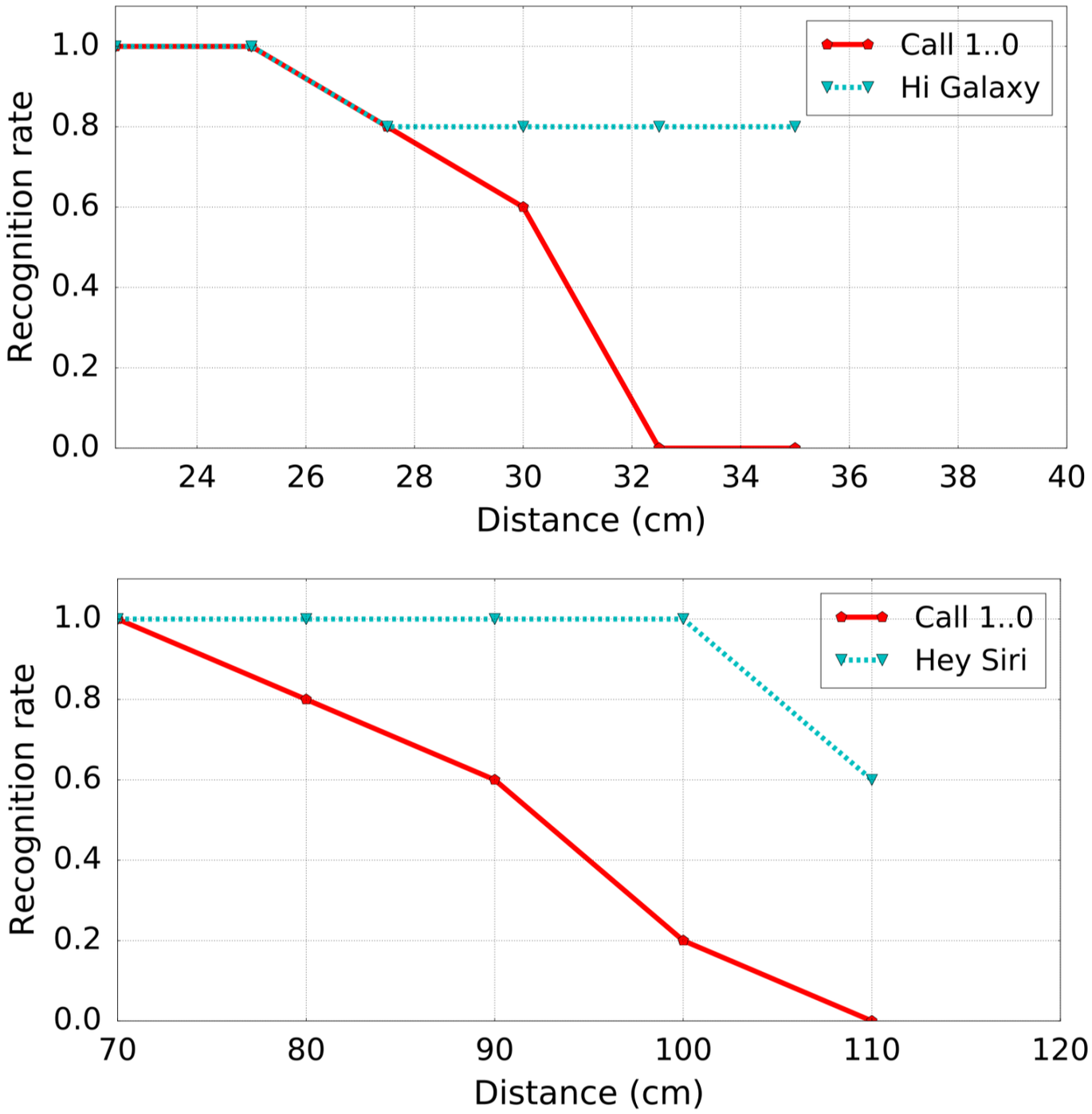
Impact: Languages



Impact: Background Noise

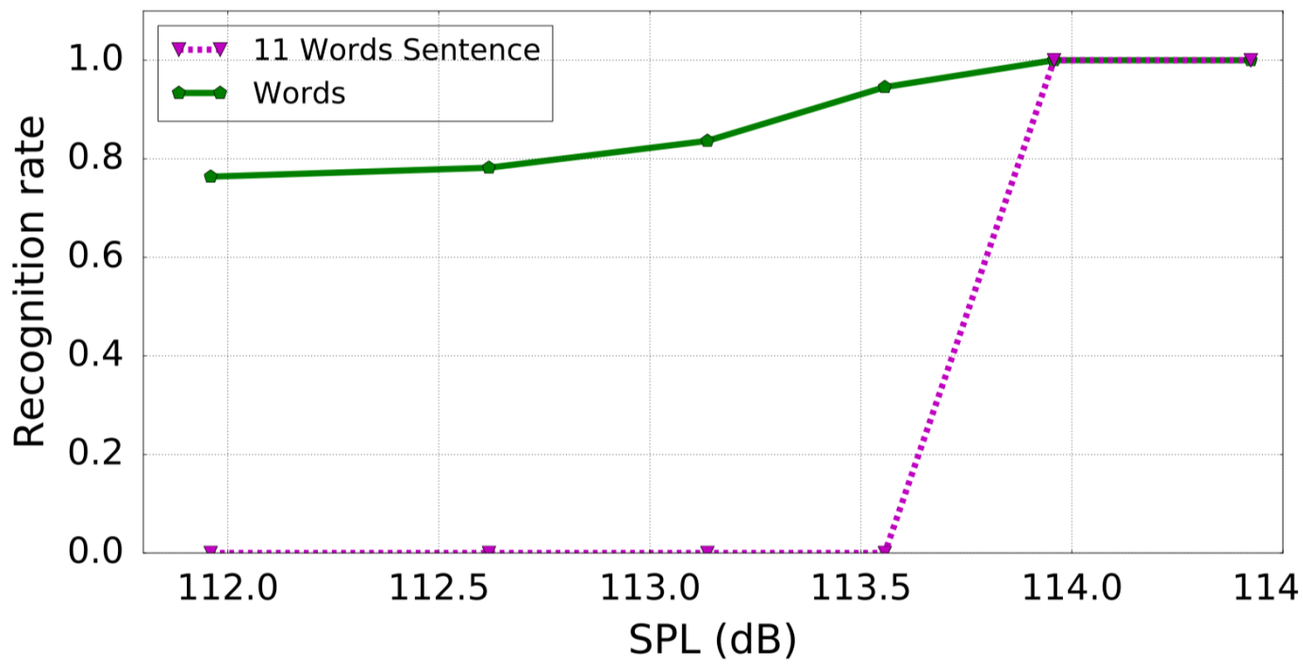
Scene	Noises (dB)	Recognition rates	
		Hey Siri	Turn on airplane mode
Office	55–65	100%	100%
Cafe	65–75	100%	80%
Street	75–85	90%	30%

Impact: Distance

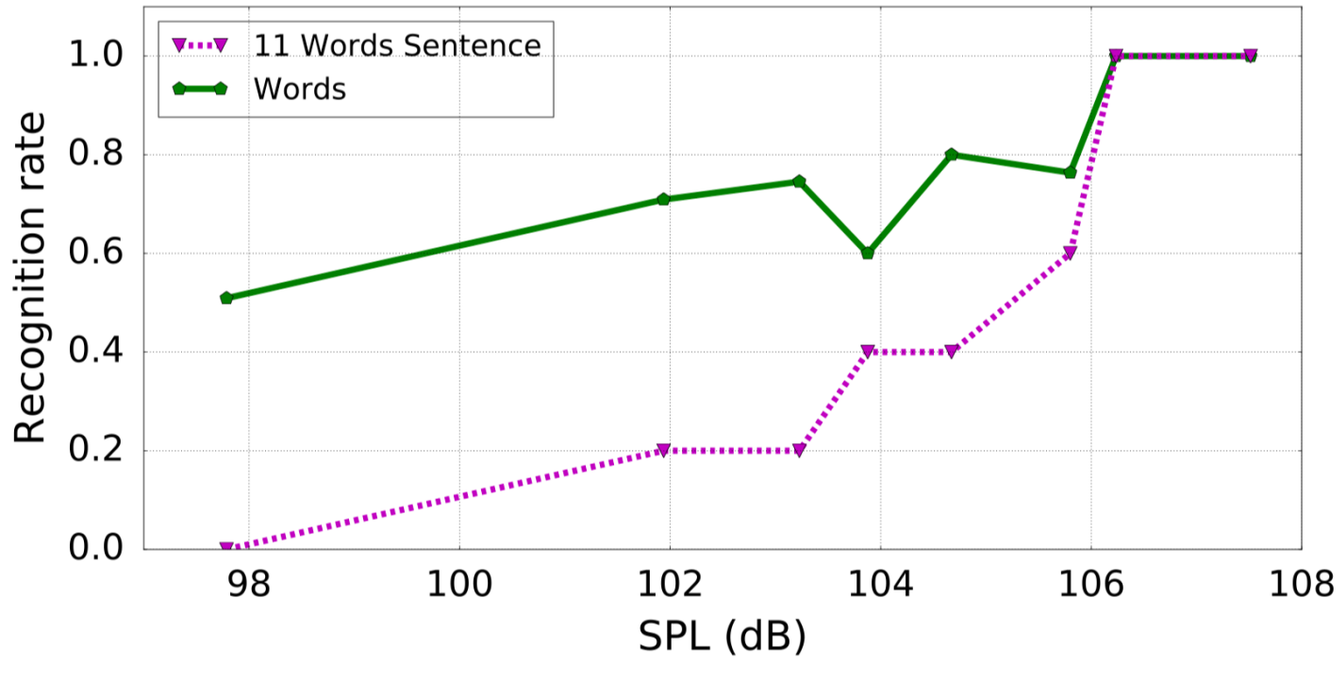


(b) The recognition rates of the Apple watch

Impact: Sound Pressure Levels



(a) The recognition rates of the Galaxy S6 Edge



(b) The recognition rates of the Apple watch

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.

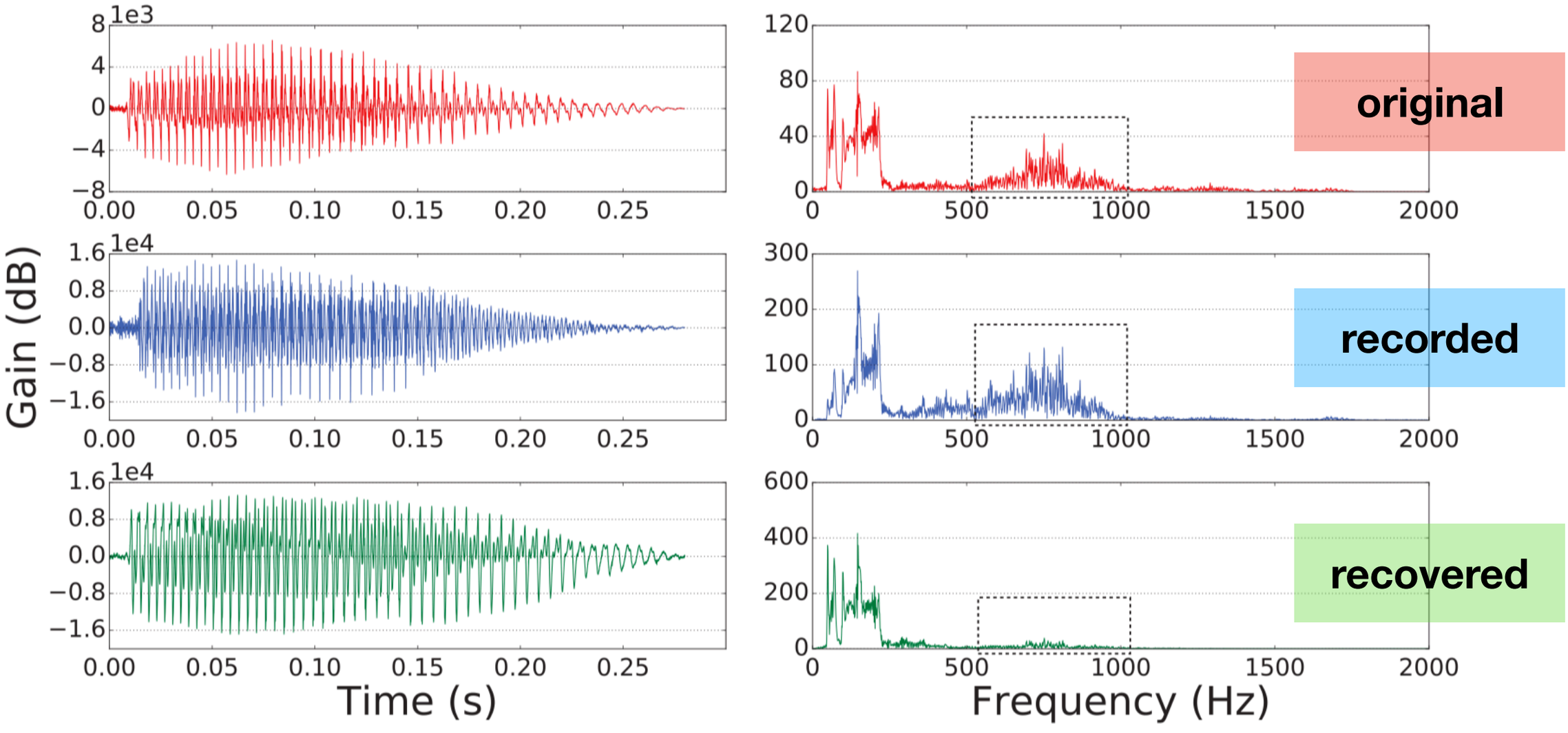
* Experimented with the front/top microphones on devices.

Almost all the systems can be attacked!

Defense: Hardware-based

- Microphone Enhancement.
    - Suppress any acoustic signals whose frequencies are in the ultrasound range.
- Inaudible Voice Command Cancellation.
    - Demodulate the signals to obtain the baseband and subtract it.

Defense: Software-based



support vector machine (SVM)
-> 10 training sample (5 positive, 5 negative)
-> 14 testing samples
100% true positive and false positive rates

Q: rigorous?

Remote attack?

Related Work

CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition

Xuejing Yuan[1,2], Yuxuan Chen[3], Yue Zhao[1,2], Yunhui Long[4], Xiaokang Liu[1,2], Kai Chen*[1,2], Shengzhi Zhang[3,5], Heqing Huang , XiaoFeng Wang[6], and Carl A. Gunter[4]

- Embed commands into songs -> distribute through the internet

Inaudible Voice Commands: The Long-Range Attack and Defense

Nirupam Roy, Sheng Shen, Haitham Hassanieh, Romit Roy Choudhury
University of Illinois at Urbana-Champaign

- Use multiple speakers to mitigate leakage

Thanks!