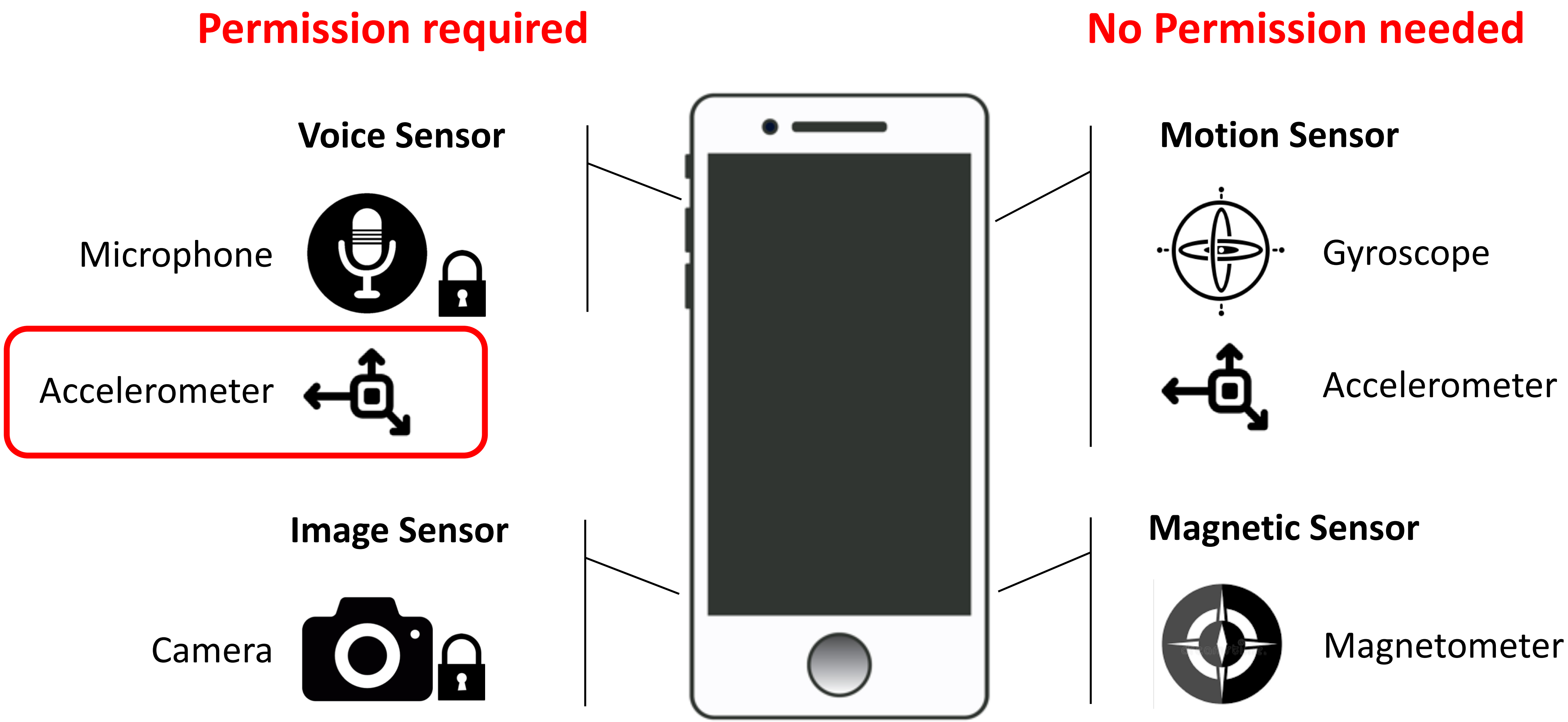
Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer

Authors: Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu and Kui Ren

Presenter: Shiqing Luo

Smartphone Sensors



Motion Sensor Threat to Speech Privacy

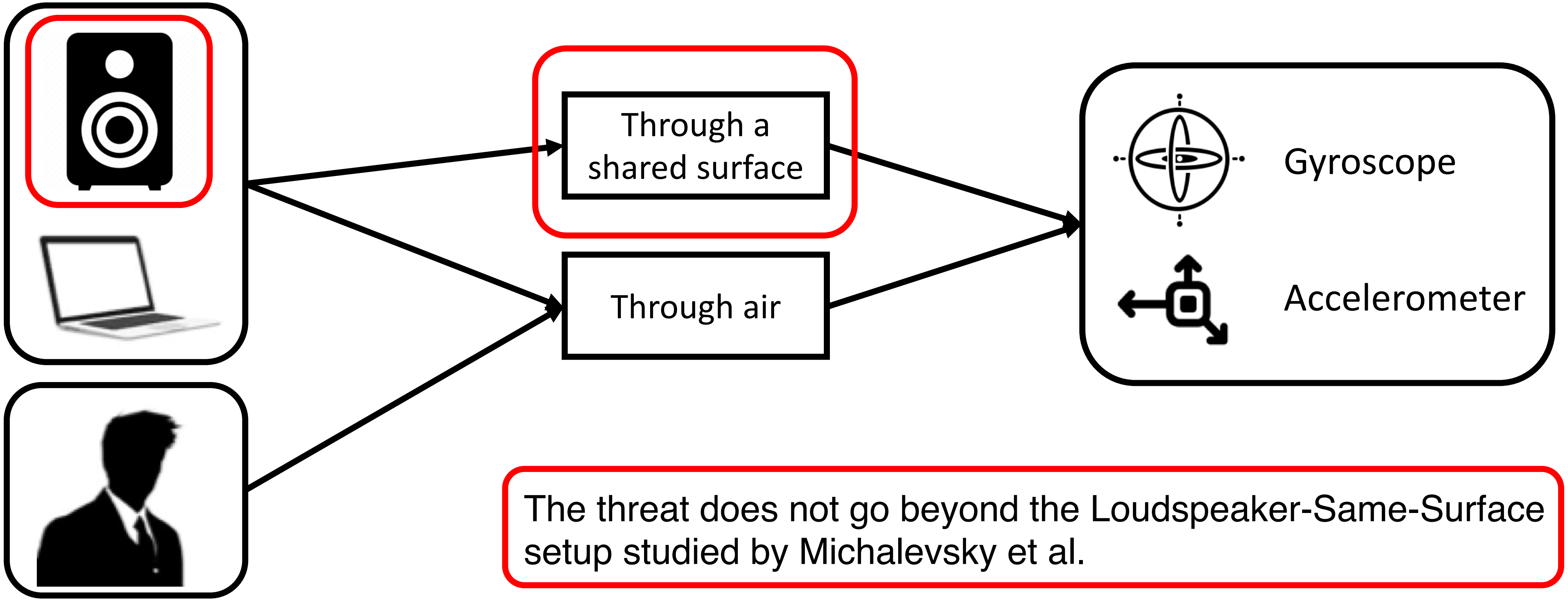
- A smartphone gyroscope can pick up surface vibrations incurred by an independent loudspeaker placed on the same table (Michalevsky et al., Usenix 2014).

- Gyroscopes are (lousy but still) microphones.
  - Very low signal to noise ratio
  - Low sampling frequency

Speaker	Speaker Identification	Digits Recognition
Mixed Female/Male	50%	17%
Female speakers	45%	26%
Male speakers	65%	23%

Motion Sensor Threat to Speech Privacy

- Only loudspeaker-rendered speech signals traveling through a solid surface can create noticeable impacts on motion sensors (Anand et al., S&P 2018).



The threat does not go beyond the Loudspeaker-Same-Surface setup studied by Michalevsky et al.

Commonly Believed Limitations

- Can only pick up a narrow band of speech signals
  - Android has a sampling ceiling of 200 Hz
  - iOS has a sampling ceiling of 100 Hz

Fundamental frequency range of human speech	
	
85-180 Hz	165-255 Hz

- Does not go beyond the Loudspeaker-Same-Surface setup
  - Very low SNR (Signal-to-Noise Ratio)
  - Sensitive to sound angle of arrival

Our Observations: Sampling Frequency

- The actual sampling rates of motion sensors are determined by the performance of the smartphone.
- Accelerometers on recent smartphones can cover almost the entire fundamental frequency band (85-255Hz) of adult speech.

Model	Year	Sampling Rate
Moto G4	2016	100 Hz
Samsung J3	2016	100 Hz
LG G5	2016	200 Hz
Huawei Mate 9	2016	250 Hz
Samsung S8	2017	420 Hz
Google Pixel 3	2018	410 Hz
Huawei P20 Pro	2018	500 Hz
Huawei Mate 20	2018	500 Hz

Sampling frequencies supported by Android [1]

Delay Options	Delay	Sampling Rate
`DELAY_NORMAL`	200 ms	5 Hz
`DELAY_UI`	20 ms	50 Hz
`DELAY_GAME`	60 ms	16.7 Hz
`DELAY_FASTEST`	0 ms	AFAP

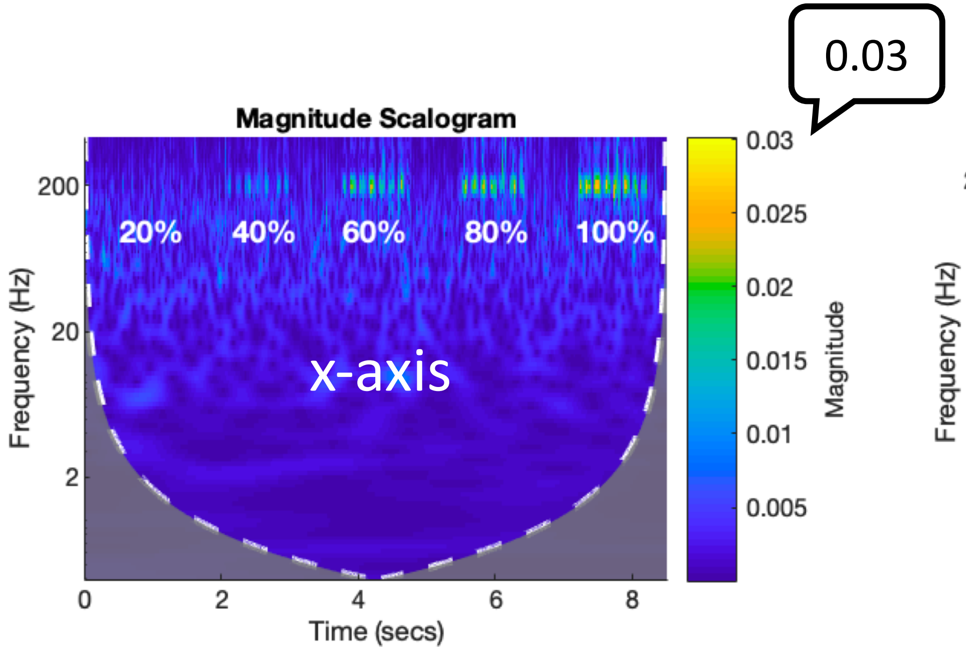
The 200 Hz sampling ceiling no longer exists

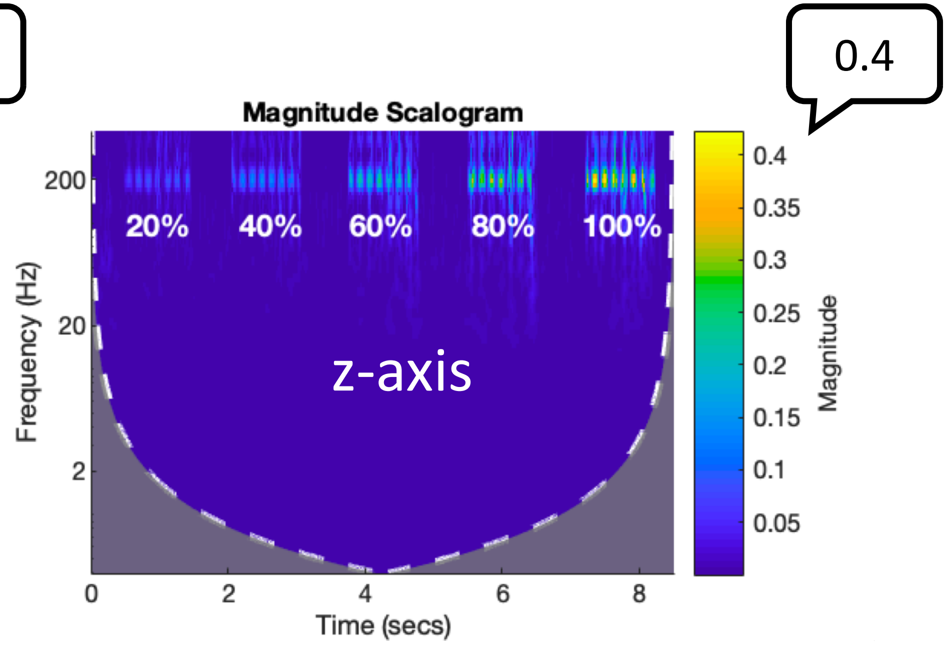
[1] "Sensor Overview," https://developer.android.com/guide/topics/sensors/sensors_overview.

Our Observations: New Setup

- Employs a smartphone’s accelerometer to eavesdrop on the speaker in the same smartphone.

- Much Higher SNR
- Sound always arrives from the same direction



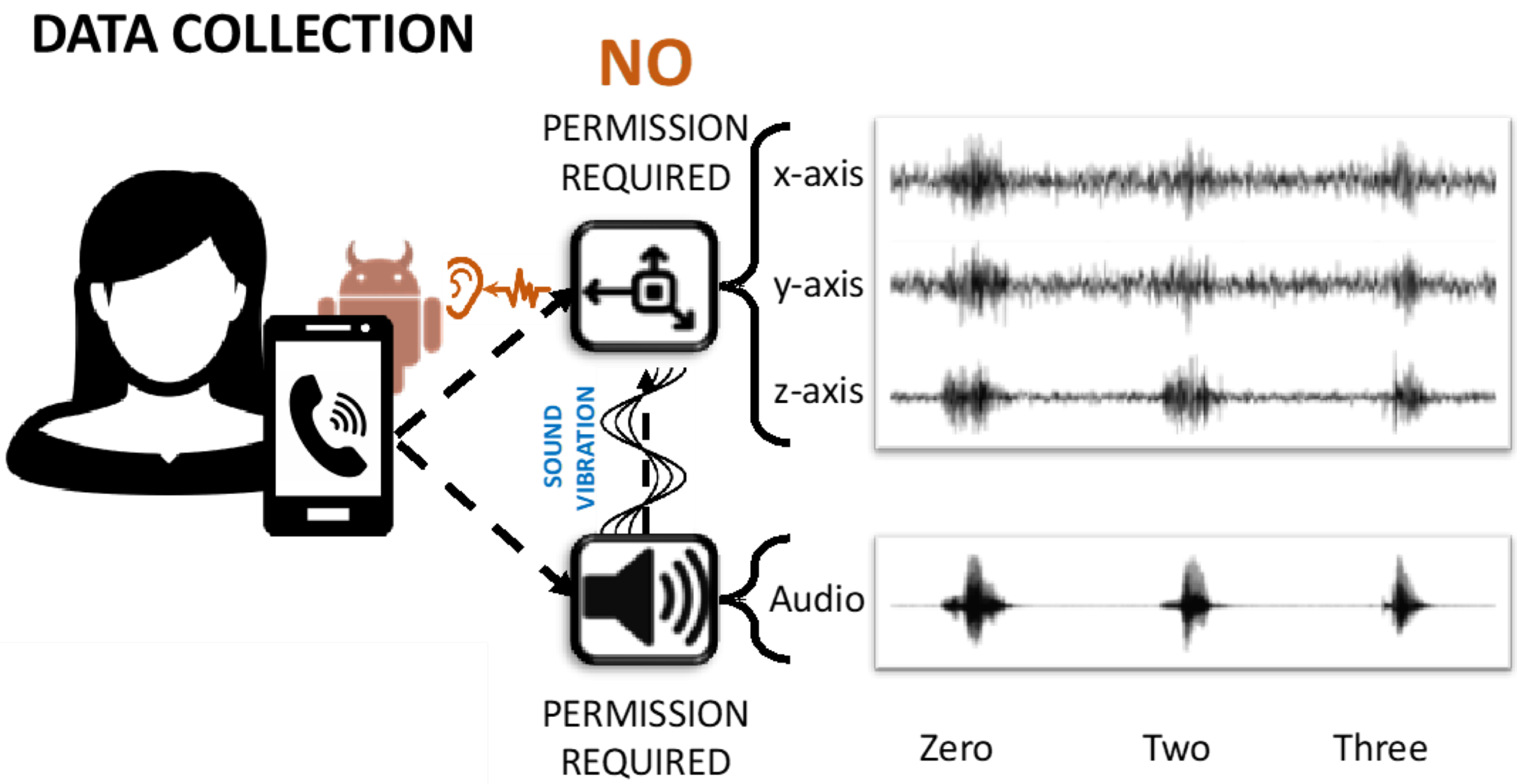


Our Observations: New Setup

- Employs a smartphone’s accelerometer to eavesdrop on the speaker in the same smartphone.

- Much Higher SNR
- Sound always arrives from the same direction
- A smartphone speaker is more likely to reveal sensitive information than an independent loudspeaker.

Threat Model

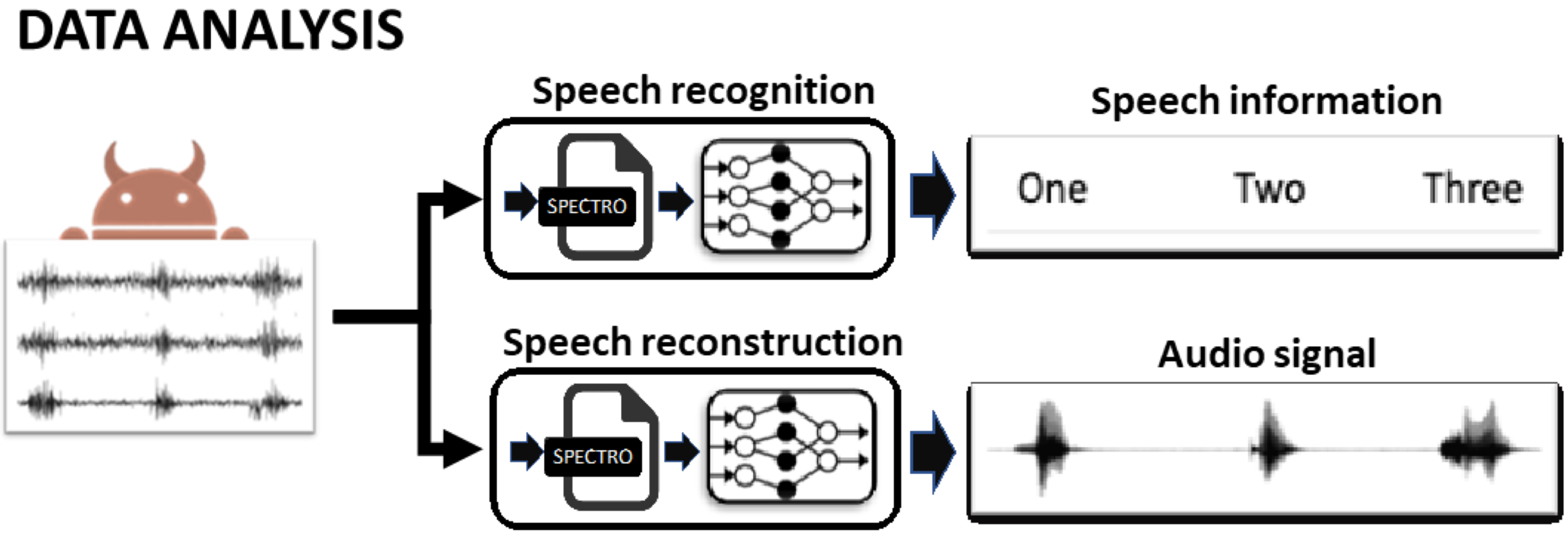


Handhold setting

Table setting

Accelerometer-based Smartphone Eavesdropping

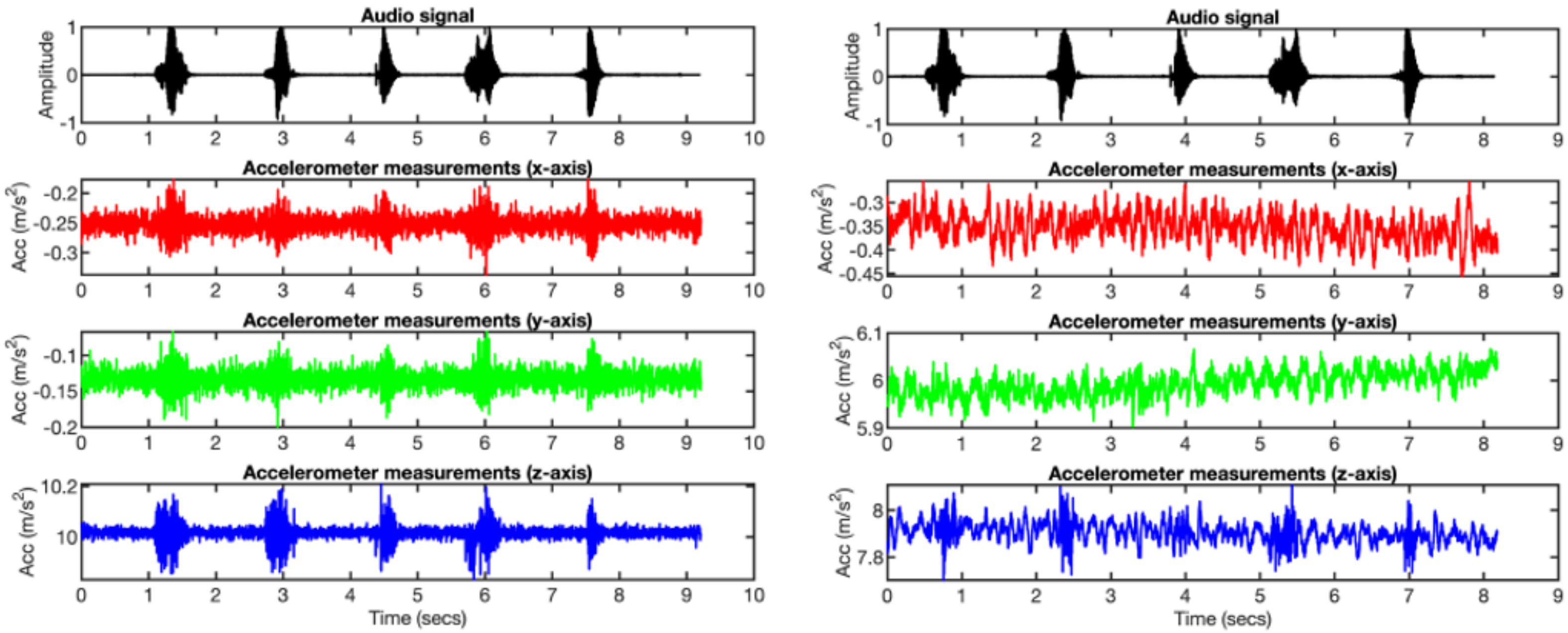
- Preprocessing: convert acceleration signals into spectrograms.
- Speech Recognition: convert spectrograms to text.
- Speech Reconstruction: reconstructs voice signals from spectrograms



Preprocessing

- Problems in Raw Acceleration Signals
  - Raw accelerometer measurements are not sampled at fixed interval.
  - Raw accelerometer measurements can be distorted by human movement.
  - Raw accelerometer measurements have captured multiple digits and needs to be segmented.

Time (ms)	x-axis (m/s^2)	y-axis (m/s^2)	z-axis (m/s^2)
1	-0.2130	-0.1410	10.0020
2	-0.1870	-0.1440	9.9970
3	-0.2110	-0.1510	9.9970
5	-0.2110	-0.1410	10.0070
8	-0.2080	-0.1340	10.0120
10	-0.2150	-0.1320	10.0070



(a) Table setting
(b) Handhold setting

Step 1: Generate Sanitized Single-word Signals

- Interpolation
  - Upsample accelerometer signals to 1000 Hz using linear interpolation.

Time (ms)	x-axis (m/s^2)	y-axis (m/s^2)	z-axis (m/s^2)
1	-0.2130	-0.1410	10.0020
2	-0.1870	-0.1440	9.9970
3	-0.2110	-0.1510	9.9970
5	-0.2110	-0.1410	10.0070
8	-0.2080	-0.1340	10.0120
10	-0.2150	-0.1320	10.0070

Step 1: Generate Sanitized Single-word Signals

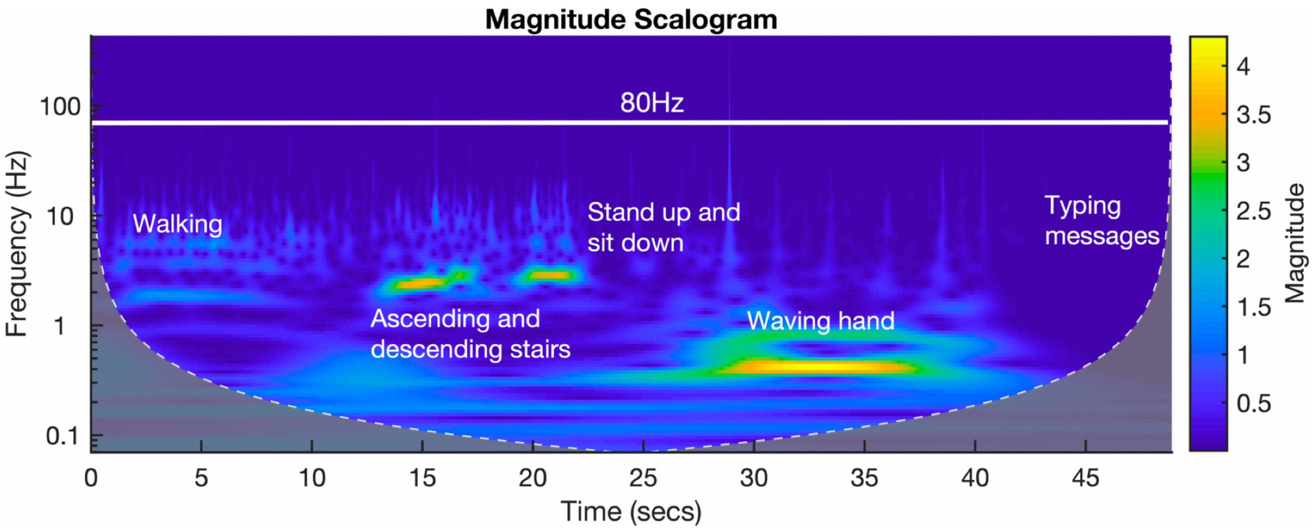
- Interpolation
  - Upsample accelerometer signals to 1000 Hz using linear interpolation.

Time (ms)	x-axis (m/s^2)	y-axis (m/s^2)	z-axis (m/s^2)
1	-0.2130	-0.1410	10.0020
2	-0.1870	-0.1440	9.9970
3	-0.2110	-0.1510	9.9970
4	-0.2110	-0.1460	10.0020
5	-0.2110	-0.1410	10.0070
6	-0.2100	-0.1387	10.0087
7	-0.2090	-0.1363	10.0103
8	-0.2080	-0.1340	10.0120
9	-0.2115	-0.1330	10.0095
10	-0.2150	-0.1320	10.0070

Step 1: Generate Sanitized Single-word Signals

- Interpolation
  - Upsample accelerometer signals to 1000 Hz using linear interpolation.
- High-pass filter
  - Convert the acceleration signal along each axis to the frequency domain and eliminate frequency components below 80 Hz.

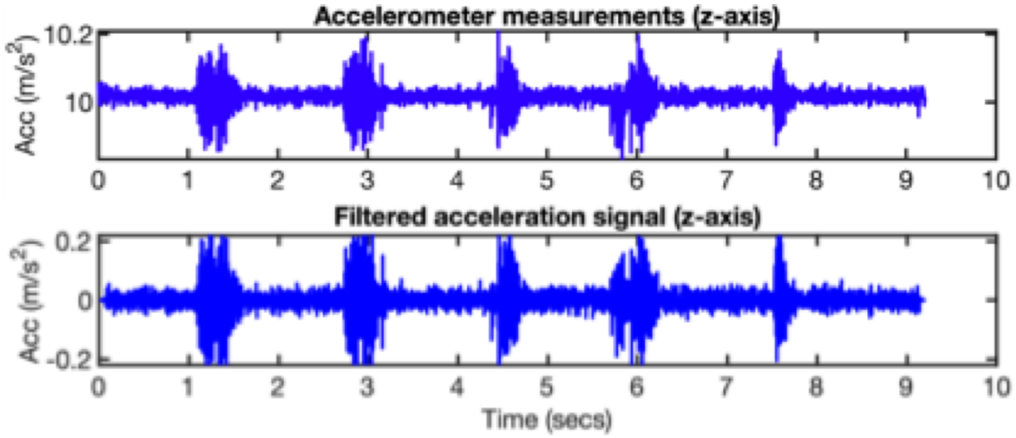
Fundamental frequency range of human speech	
85-180 Hz	165-255 Hz



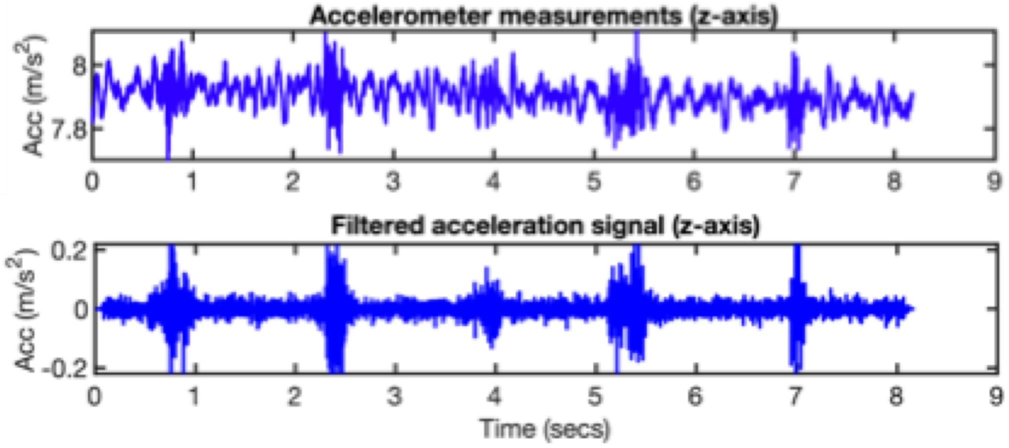
Step 1: Generate Sanitized Single-word Signals

- Interpolation
  - Upsample accelerometer signals to 1000 Hz using linear interpolation.
- High-pass filter
  - Convert the acceleration signal along each axis to the frequency domain and eliminate frequency components below 80 Hz.

Table setting



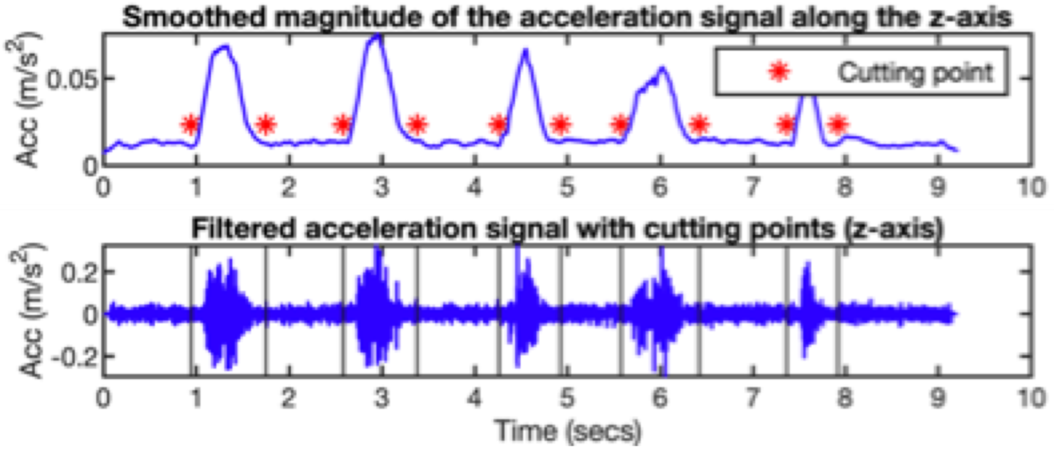
Handhold setting



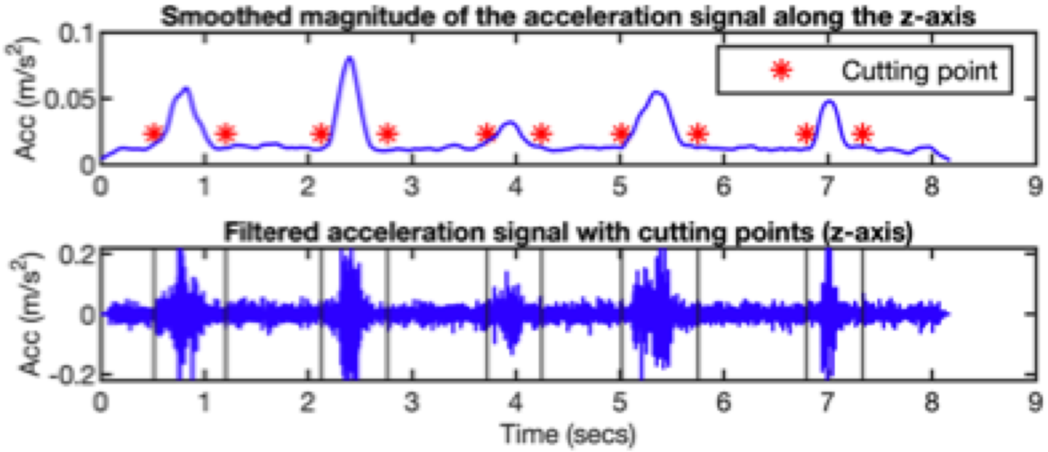
Step 1: Generate Sanitized Single-word Signals

- Interpolation
  - Upsample accelerometer signals to 1000 Hz using linear interpolation.
- High-pass filter
  - Convert the acceleration signal along each axis to the frequency domain and eliminate frequency components below 80 Hz.
- Segmentation
  - Calculate the magnitude of the acceleration signal and smooth the obtained magnitude sequence with moving average.
  - Locate all regions with magnitudes higher than a threshold.

Table setting



Handhold setting



Step 2: Generate Spectrogram Images

- Signal-to-spectrogram conversion
  - Divide the signal into multiple short segments with a fixed overlap.
  - Window each segment with a Hamming window and calculate its spectrum through STFT (Short-Time Fourier Transform).
  - Three spectrograms can be obtained for each single-word signal.

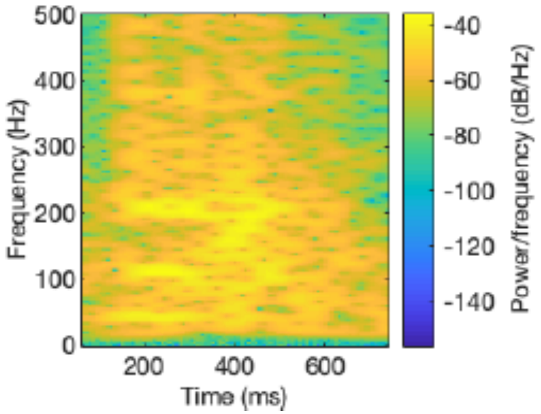
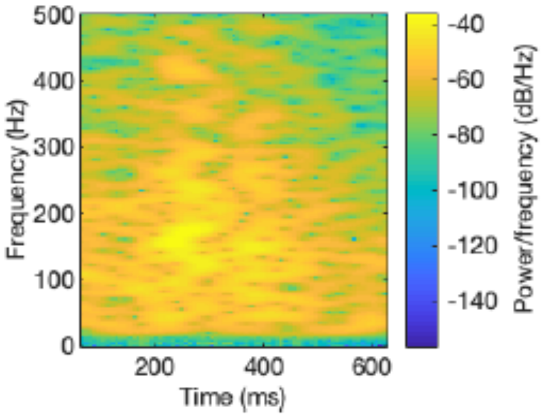


Table setting



Handhold setting

Step 2: Generate Spectrogram Images

- Signal-to-spectrogram conversion
  - Divide the signal into multiple short segments with a fixed overlap.
  - Window each segment with a Hamming window and calculate its spectrum through STFT (Short-Time Fourier Transform).
  - Three spectrograms can be obtained for each single-word signal.

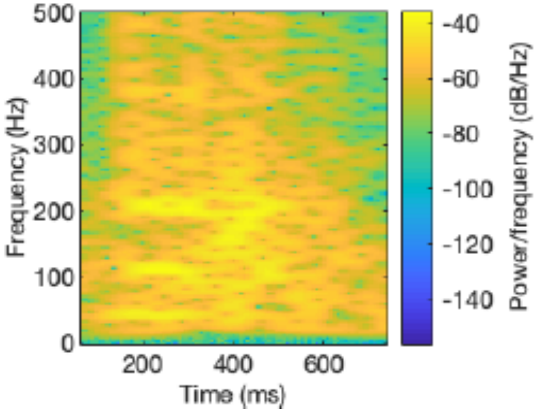
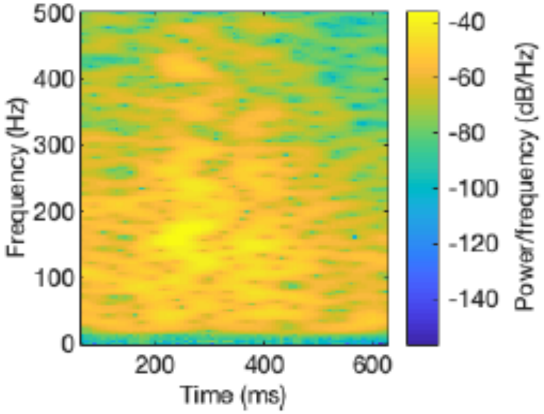


Table setting



Handhold setting

- Generate Spectrogram-Images
  - Fit the three m x n spectrograms into one m x n x 3 tensor.
  - Take the square root of all the elements in the tensor and map the obtained values to integers between 0 and 255.
  - Export the m x n x 3 tensor as an image in PNG format

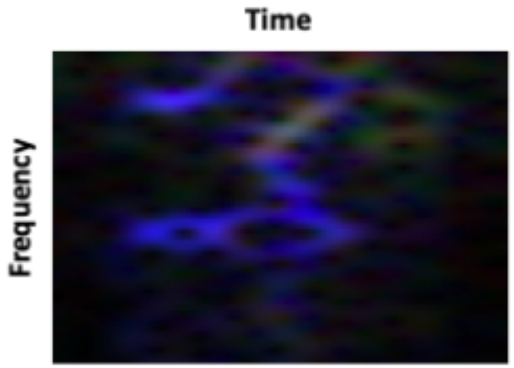
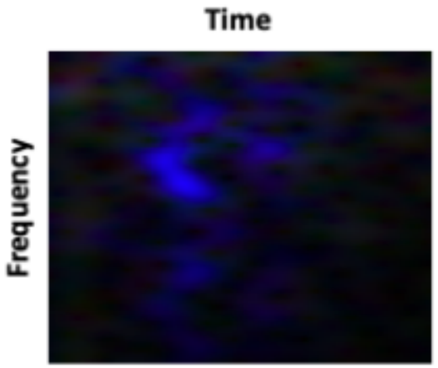


Table setting

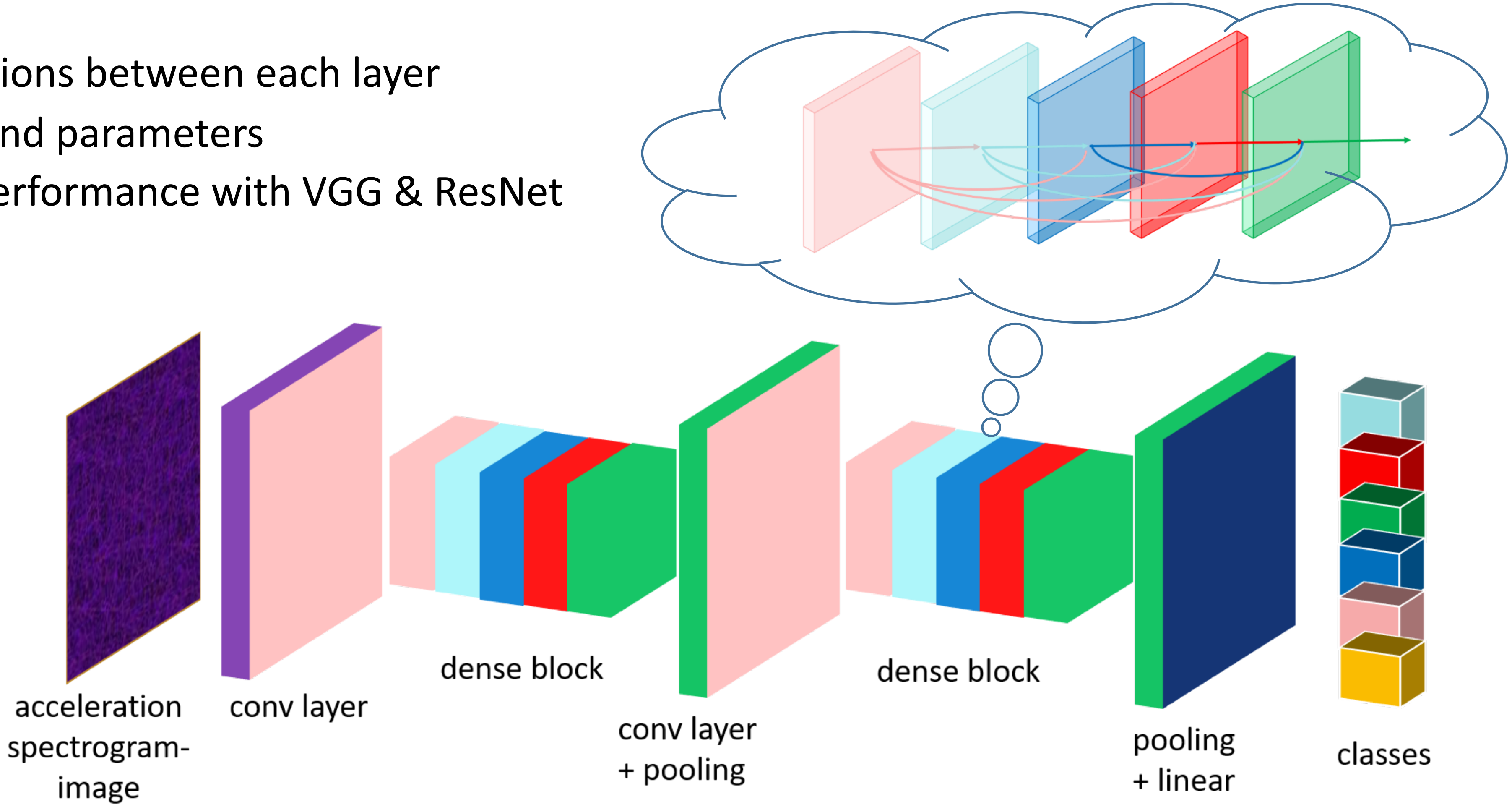


Handhold setting

Speech Recognition

- DenseNet:
  - Direct connections between each layer
  - Fewer nodes and parameters
  - Comparable performance with VGG & ResNet

$$\mathbf{x}_l = \mathcal{H}_l([\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{l-1}])$$



Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Recognition Results

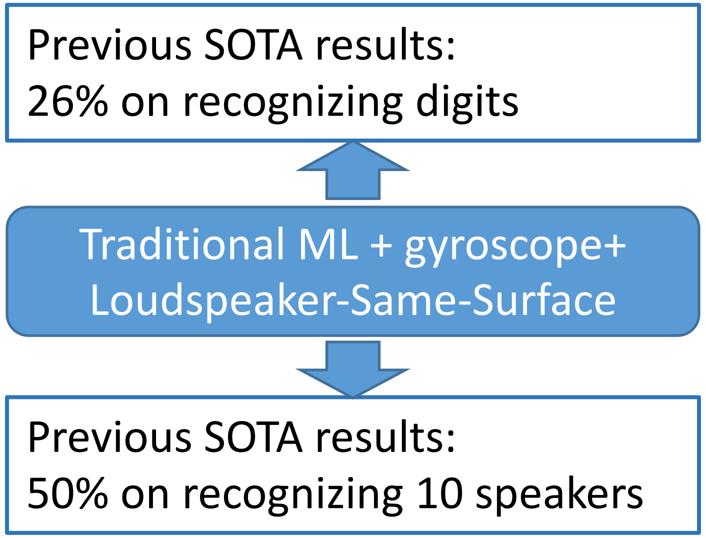
- Dataset (80% training data and 20% testing data) :
  - Digits: 10k single-digit signals from 20 speakers
  - Digits + Letter: 36*260 single-word signals from 10 speakers.

- Recognizing Digits & Letters (common elements in password)

Tasks	Top1 Acc	Top3 Acc	SOTA
Digits	78%	96%	26%
D + L	55%	78%	-

- Recognizing 20 Speakers (connect multiple attack results)

Top1 Acc	Top3 Acc	SOTA
70%	88%	50% (10)



Hot Word Search

- Locate and identify pre-trained hot (sensitive) words from sentences.

Insensitive words — Hot words

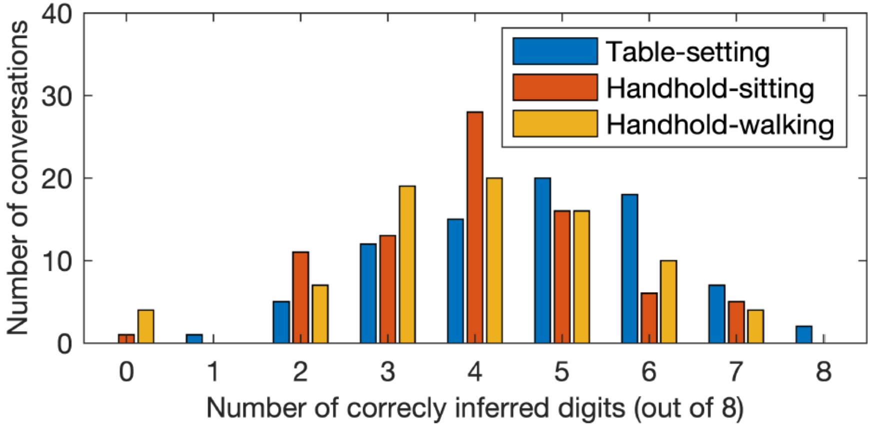
Here is my social security number

- Speakers
  - Two males and two females
- Training data
  - 128*8 hot words
  - 2176 insensitive words (negative samples)
- Testing data
  - 200 short sentences, each of which contains several insensitive words and one to three hot words.

Hot word	TPR	FPR
Password	94%	0.4%
Username	97%	0.4%
Social	100%	0.3%
Security	91%	0.0%
Number	88%	0.1%
Email	88%	1.4%
Credit	88%	0.3%
Card	97%	1.4%

Case Study: End-to-End Attack

- Attack scenario:
  - The victim makes a phone call to a remote caller and requests a password during the conversation.
  - The password is eight digits in length and is preceded by the hot word “password (is)”.
- Attach process:
  - 1) Hotword search: Locate password.
  - 2) Digits recognition: Recognize eight-digit password.



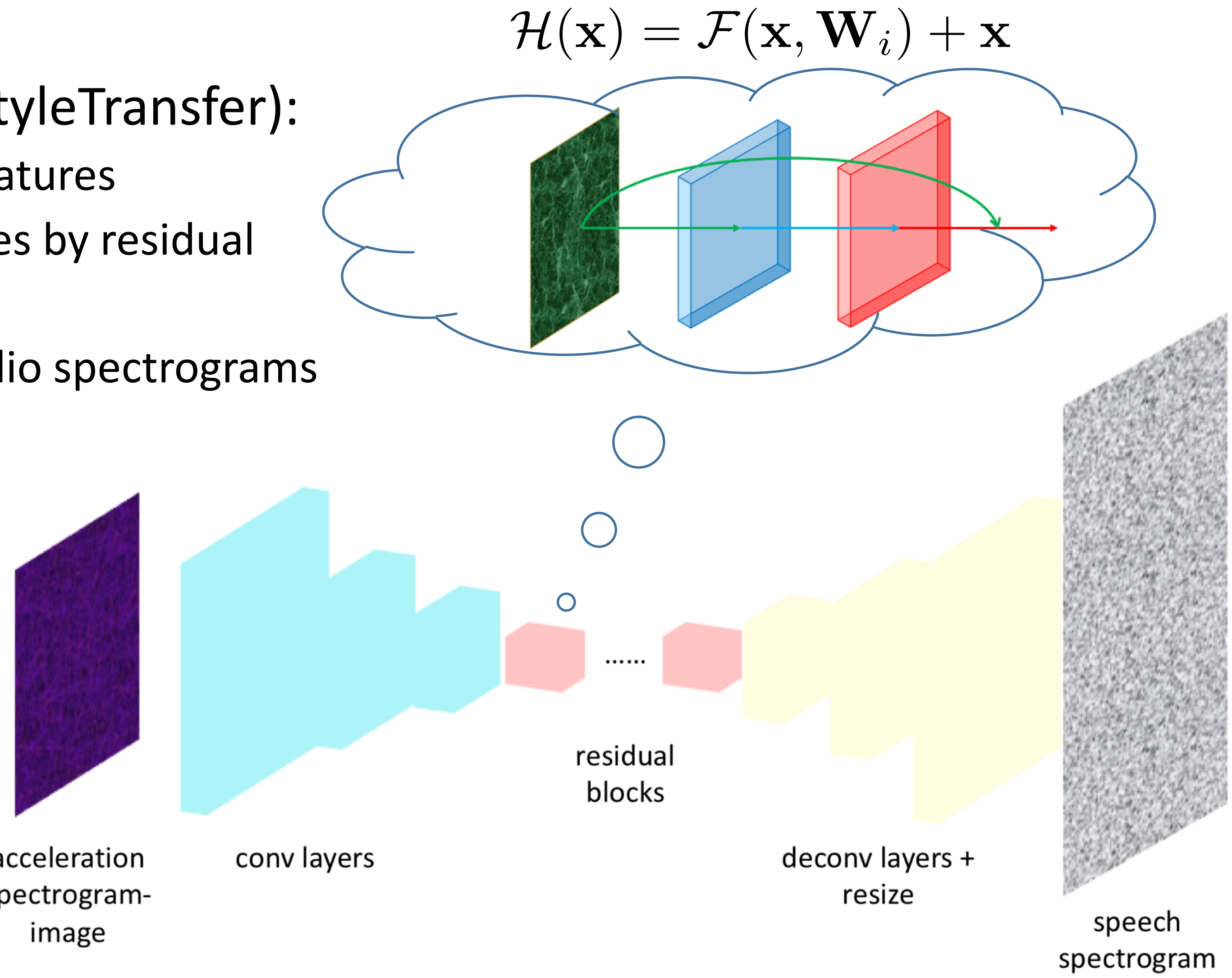
- Training data
  - 200 “password”s (Hotword search)
  - 2200 other word (Hotword search)
  - 280*10 digits (Digits recognition)
- Testing data
  - 80 conversations for each setting.

Setting	Password search	Digit recognition		
		top1	top3	top5
Table-setting	92%	59%	84%	92%
Handhold-sitting	85%	51%	83%	94%
Handhold-walking	91%	50%	81%	91%

Speech Reconstruction

- Reconstruction Network (Refer to StyleTransfer):
  - Encoder: encode spectrograms into features
  - Residue Blocks: refine encoded features by residual mappings (inspired by ResNet)
  - Decoder: decode the features into audio spectrograms

- GL algorithm:
  - Recover the phase from spectrogram
  - Recover audio signals



Johnson, Justin, et al. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

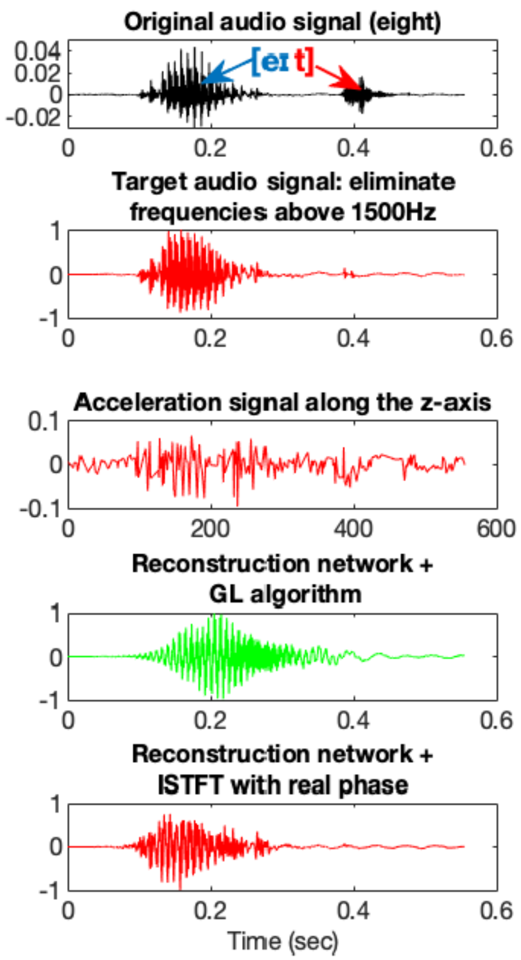
Reconstruction Results

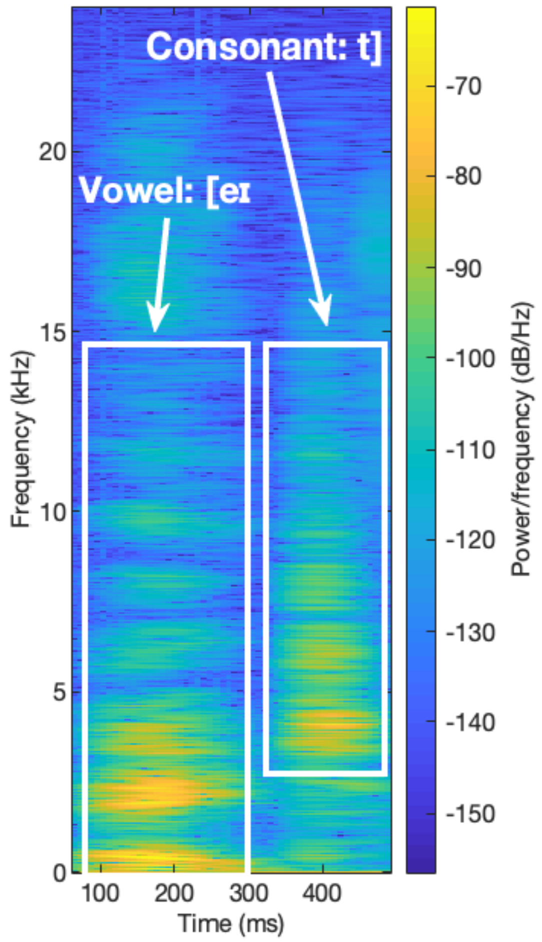
- Listen to some reconstructed examples

Password is one two three

Angry bird is my username

Here is my social security number





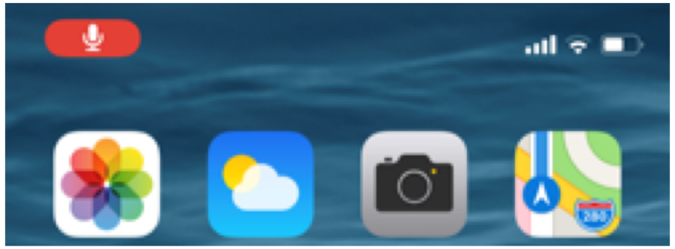
Defense

- Limit the sampling rate of the accelerometer.
  - According to Android Developer, the recommended sampling rates for the user interface and mobile games are 16.7 Hz and 50 Hz respectively.
  - Applications requiring sampling rates above 50 Hz should request a permission through <user-permission >

Recognition accuracy on the digits dataset

Sampling rate	300 Hz	200 Hz	160 Hz	100 Hz	50 Hz
Recognition accuracy	73%	64%	56%	47%	30%

- Notify the user when some applications are collecting accelerometer readings in the background.



Conclusion

- Sound signals emitted by smartphone speakers can significantly affect the accelerometer on the same smartphone.
- Accelerometers on recent smartphones almost cover the entire fundamental frequency band of speech voice.
- Using deep learning techniques, it is possible to recognize and reconstruct the speech signals from the accelerometer measurements.

Thank you!

Zhongjie Ba, Email: zhongjie.ba@mcgill.ca