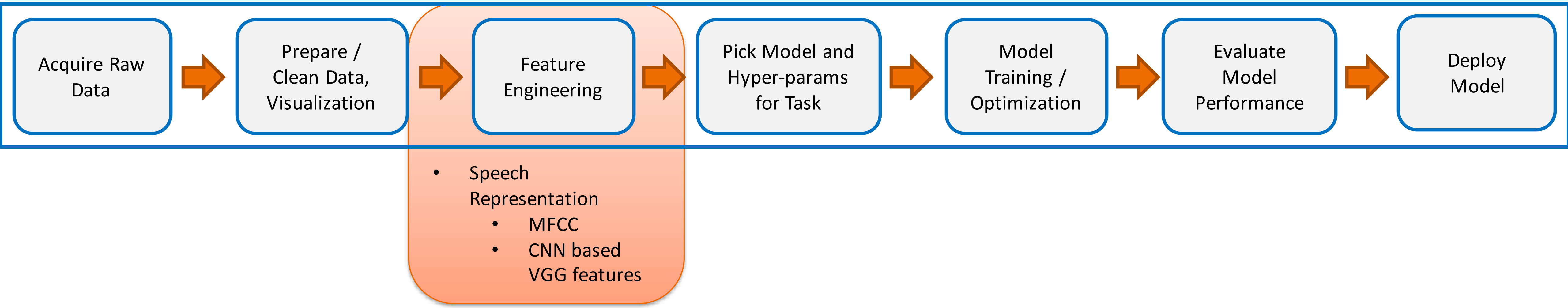
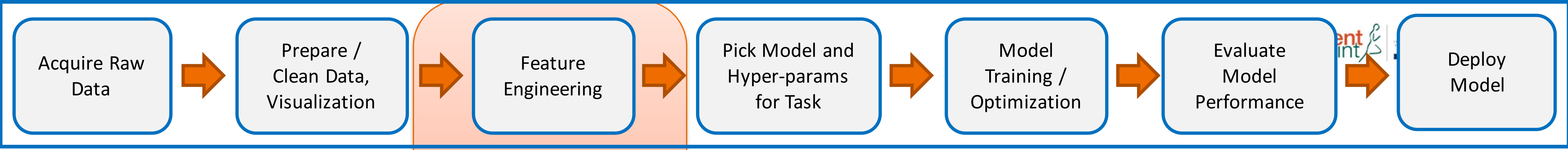
Focus for this lecture



Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- Speech Representation
  - MFCC
  - CNN based VGG features



Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- Speech

Speech

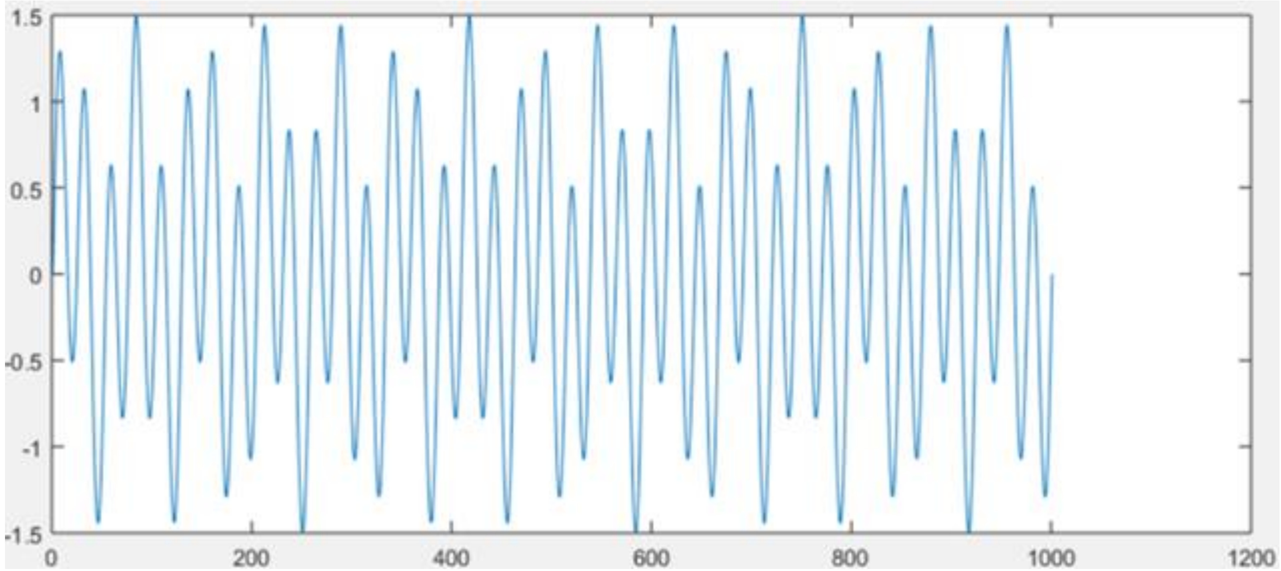
(Brief Explanation)

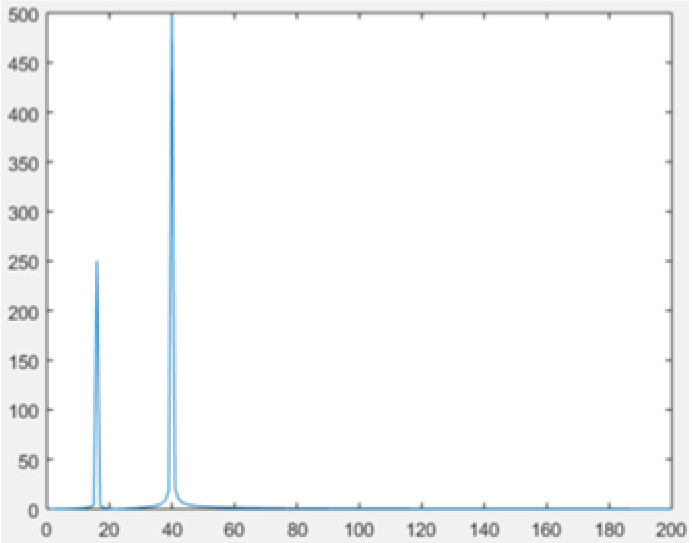
A quick primer on sound

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

Lets understand Sound Signal

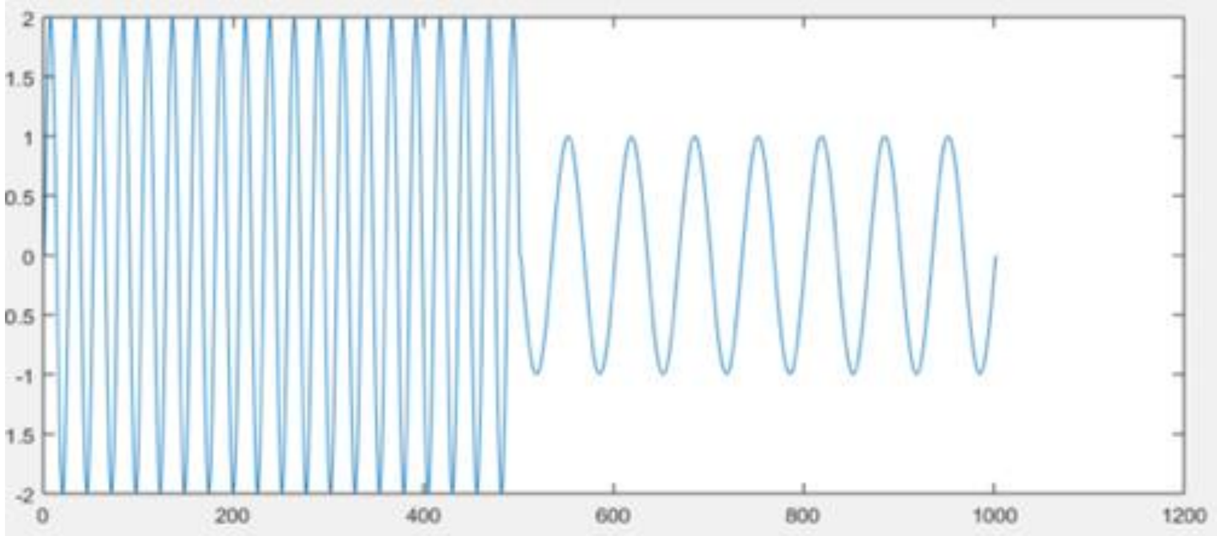
$$f(t) = \sin(2\pi \cdot 39t) + 0.5\sin(2\pi \cdot 15t)$$

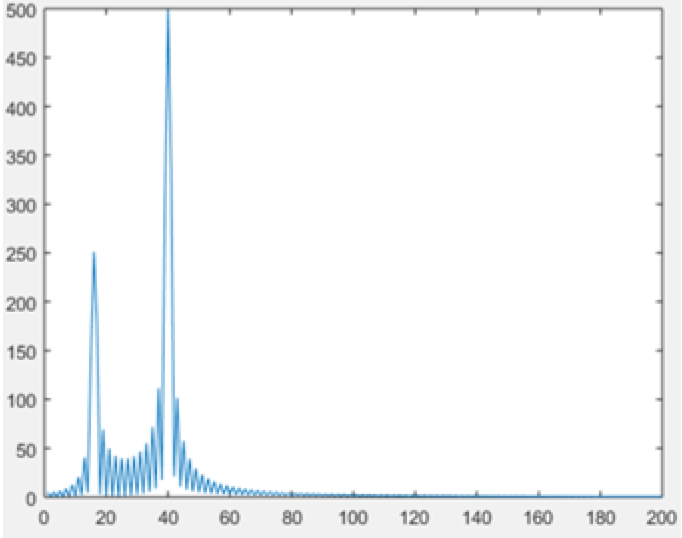




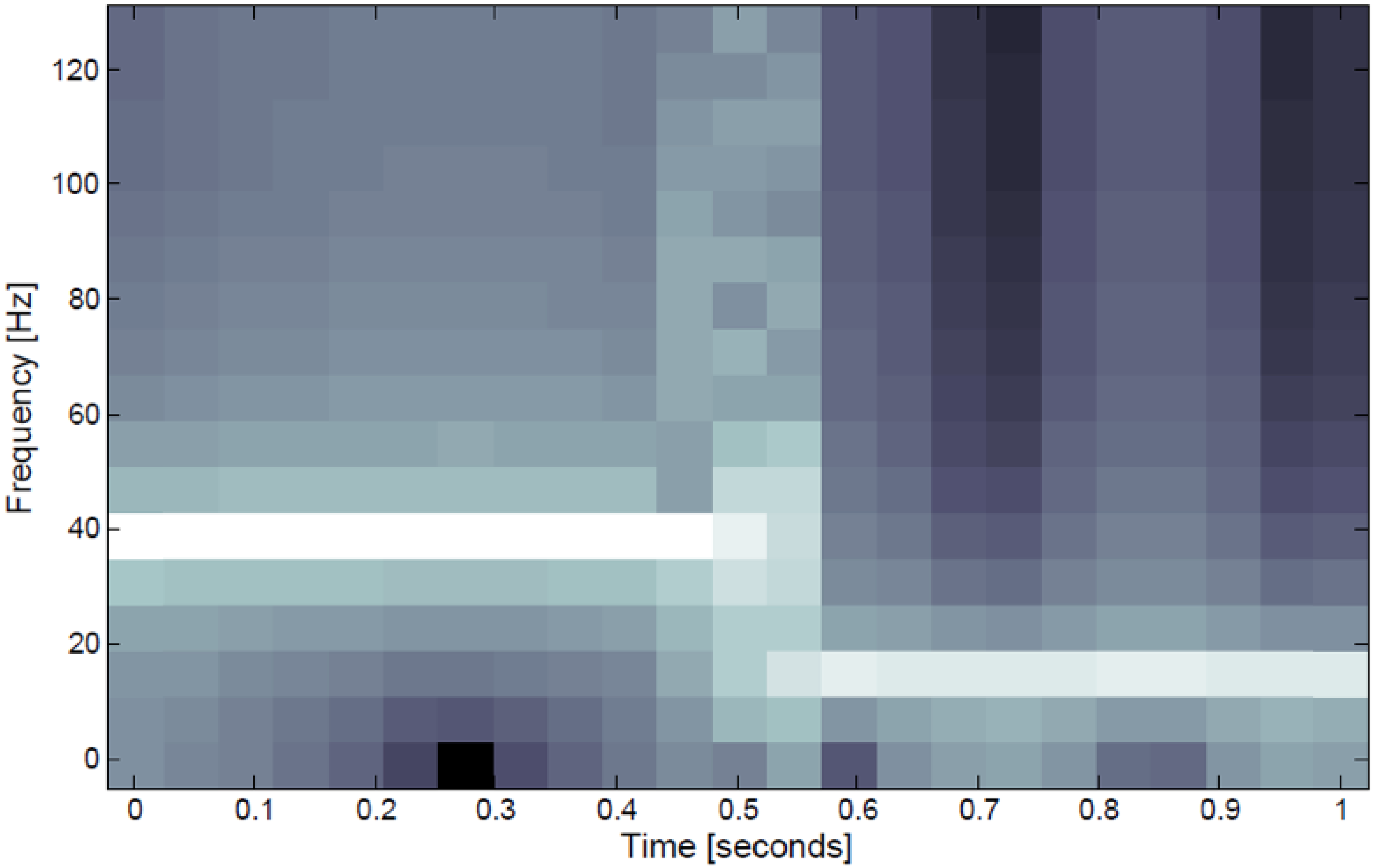
Example Sound Signal

$$g(t) = \begin{cases} 2 * \sin(2\pi \cdot 39t)\,, 0 \le t \le 1/2 \\ \sin(2\pi \cdot 15t)\,, 1/2 < t \le 1 \end{cases}$$





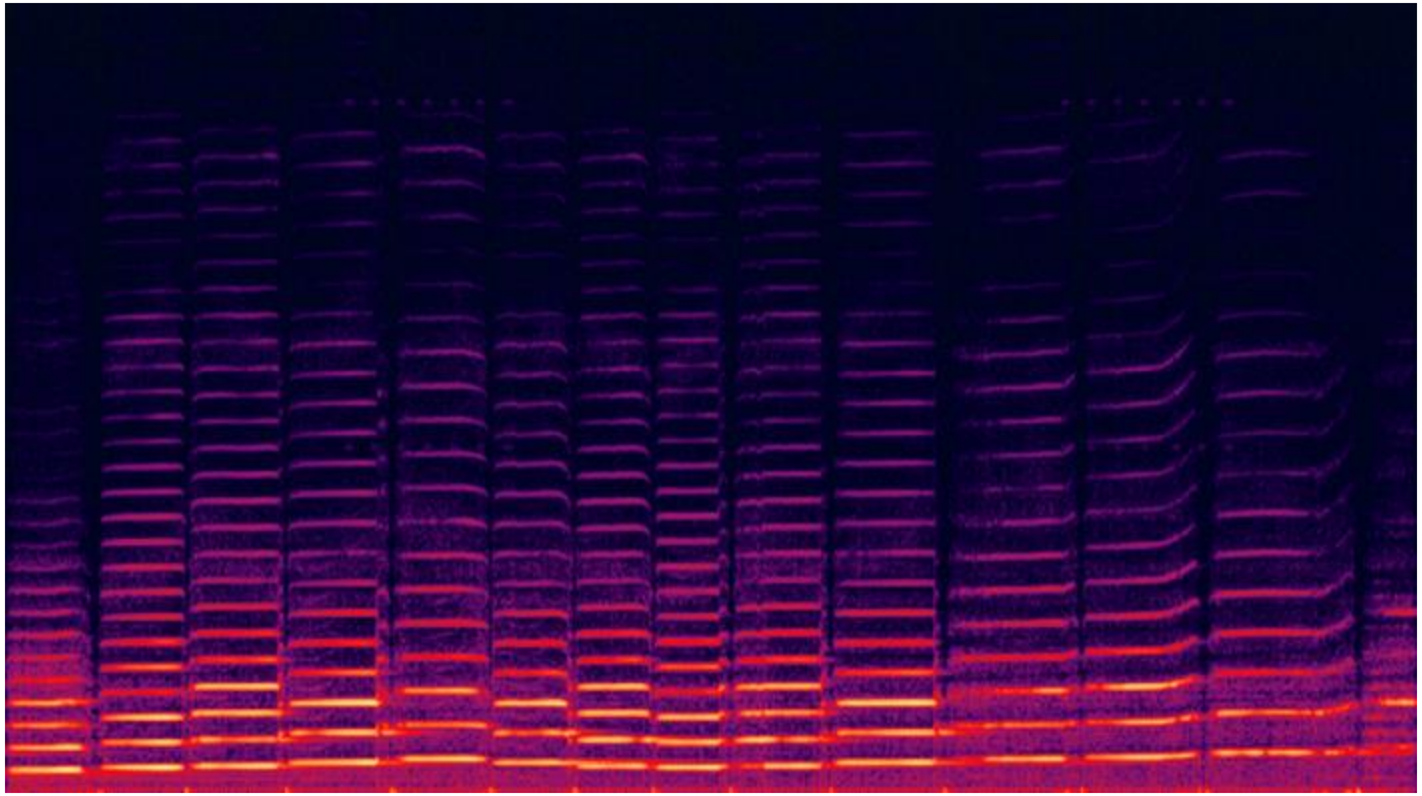
Spectrogram



Spectrogram of a piecewise monochromatic signal.

Lighter color ⍰ greater DFT magnitude

Spectrogram



Example Problem

- Sound waves (.wav files)
- 10 short commands (“zero”, “one”, “two”)
- 1 sec duration
- 5000 samples (many people)

Representations

A: MFCC (Signal processing based; Classical)

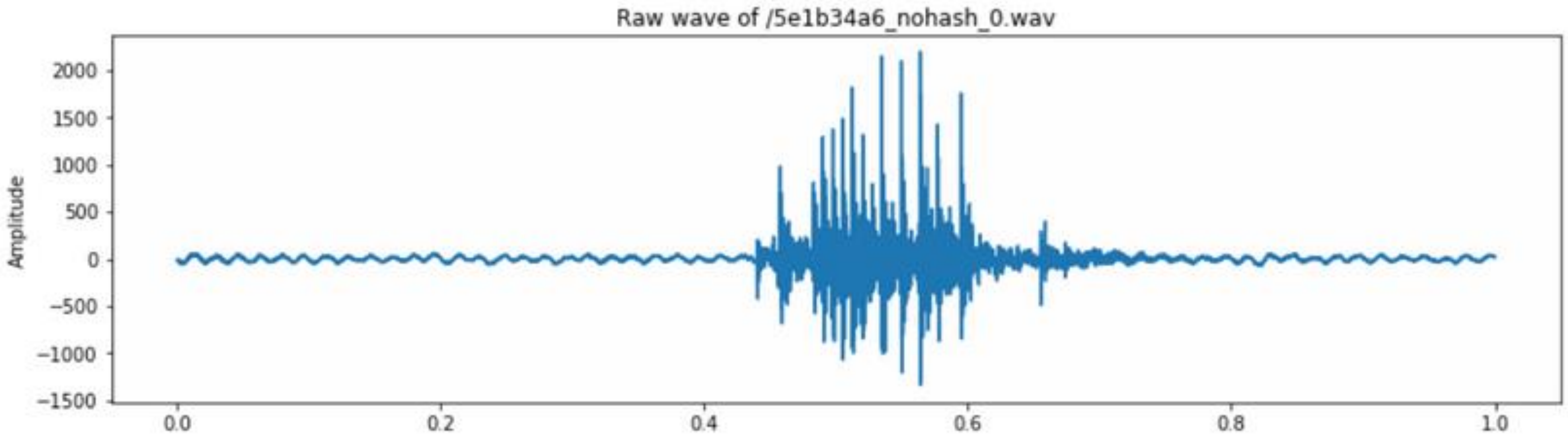
- Mel Frequency Cepstral Coefficients

B: CNN Based (Modern)

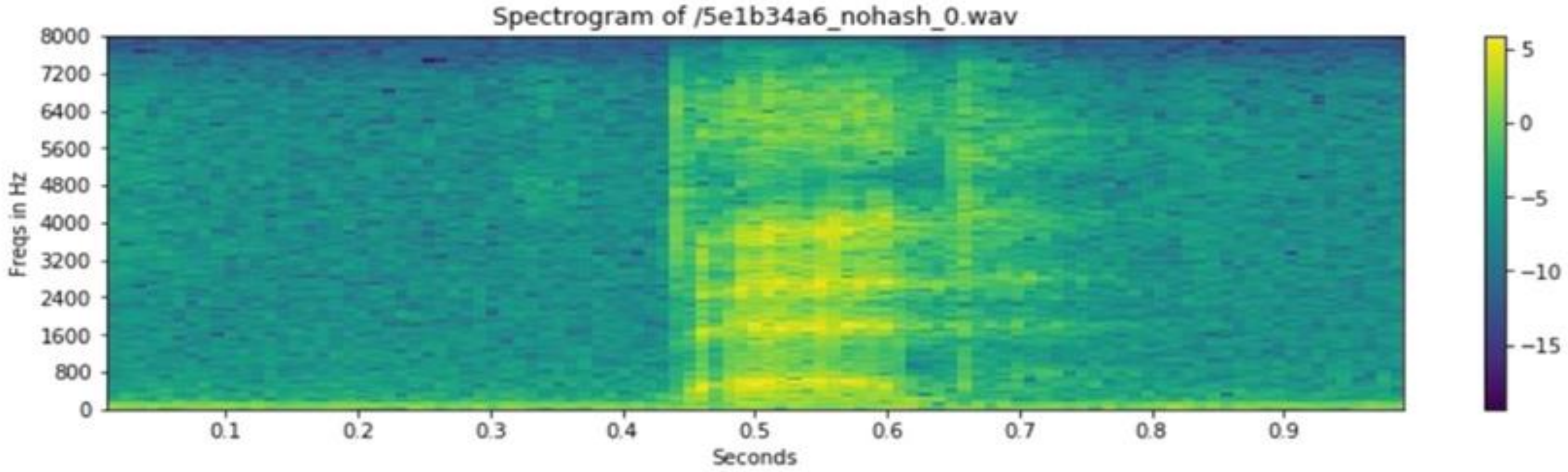
- VGG Features on the Mel Spectrogram

http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Classical Feature (MFCC)



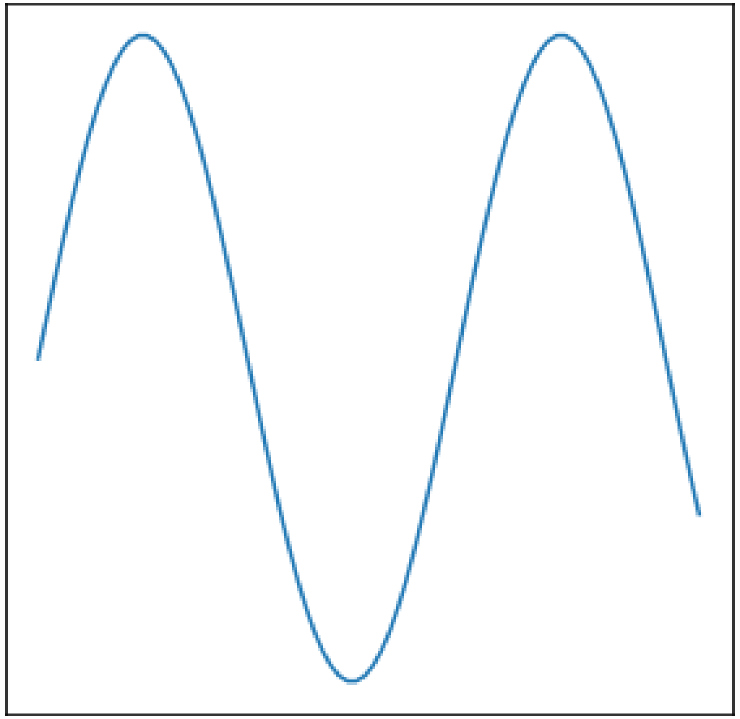
Amplitude
Vs
Time

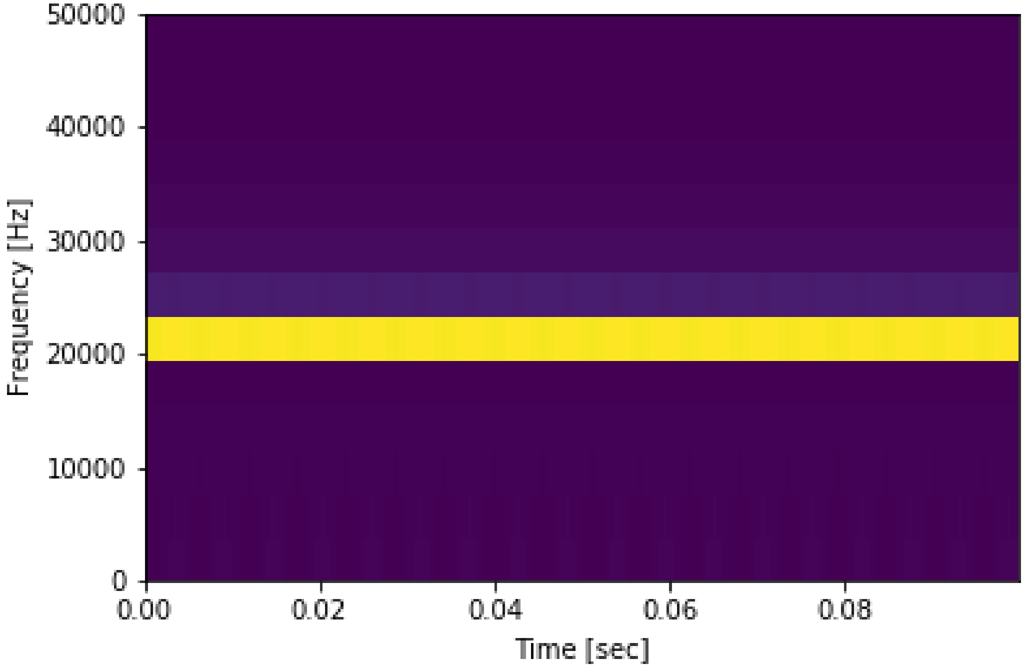


Frequency
Vs
Time

Code Available

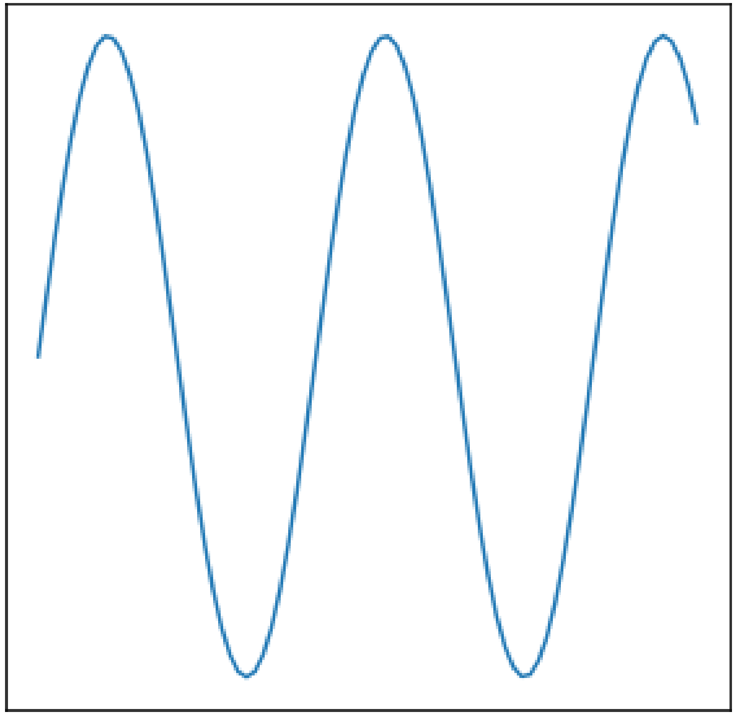
Sine wave (f= 20KHz)

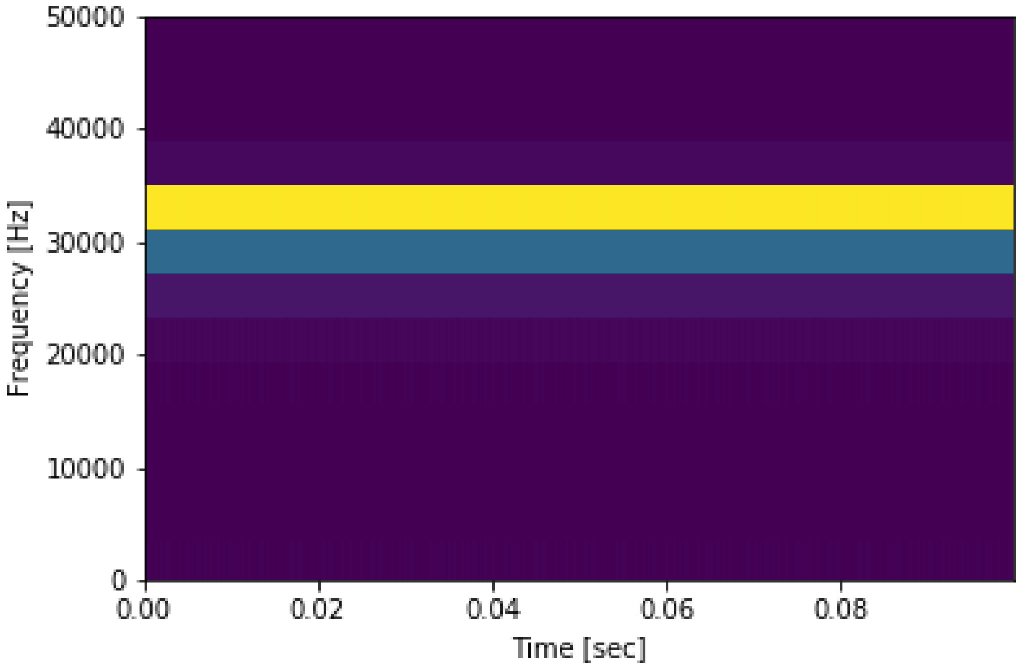




Code Available

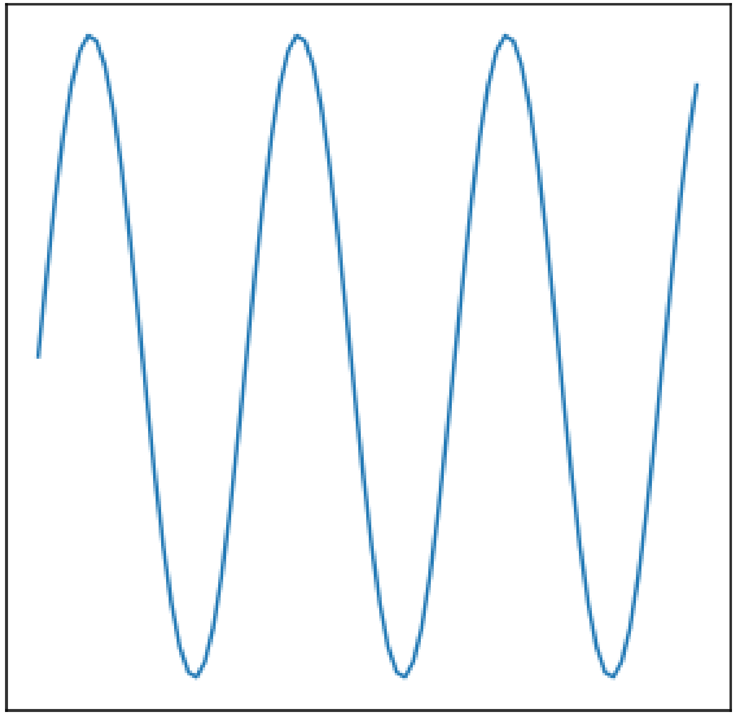
Sine wave (f= 30KHz)

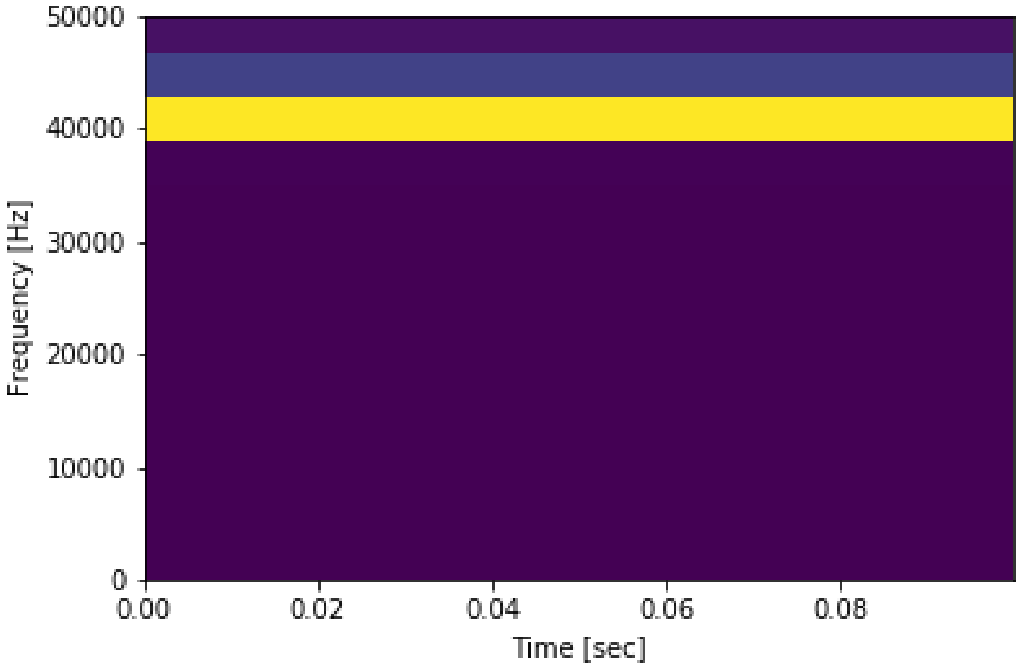




Code Available

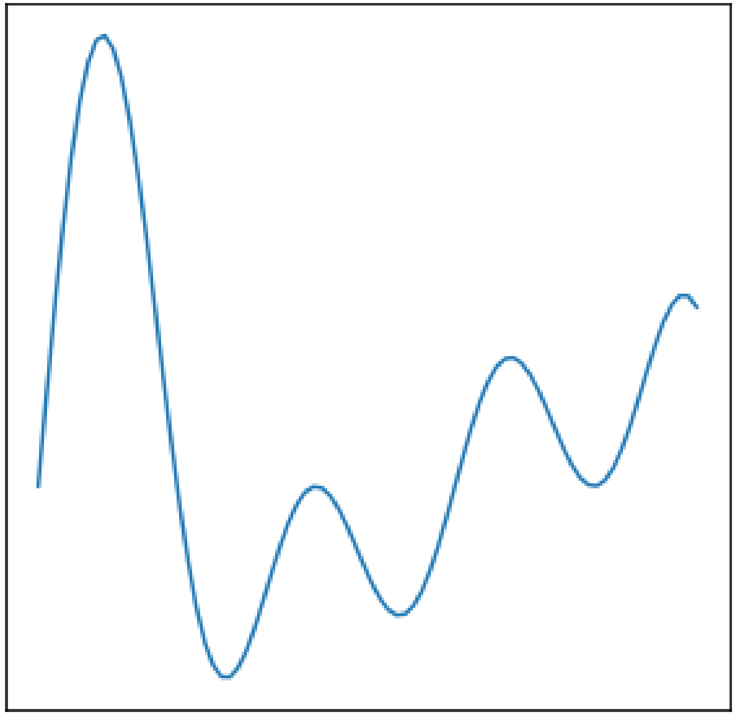
Sine wave (f= 40KHz)

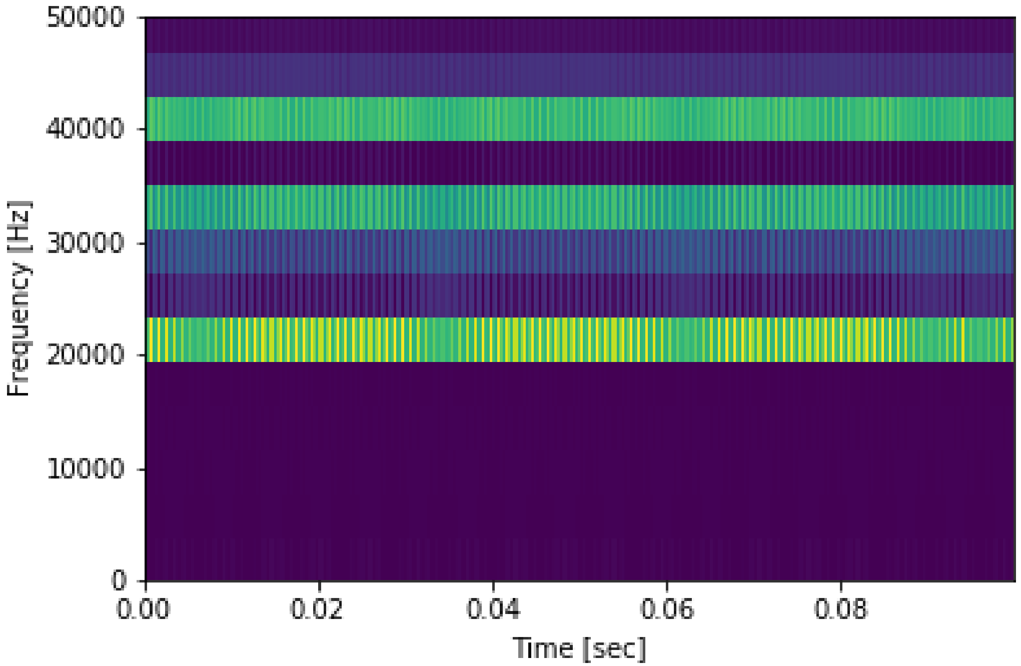




Code Available

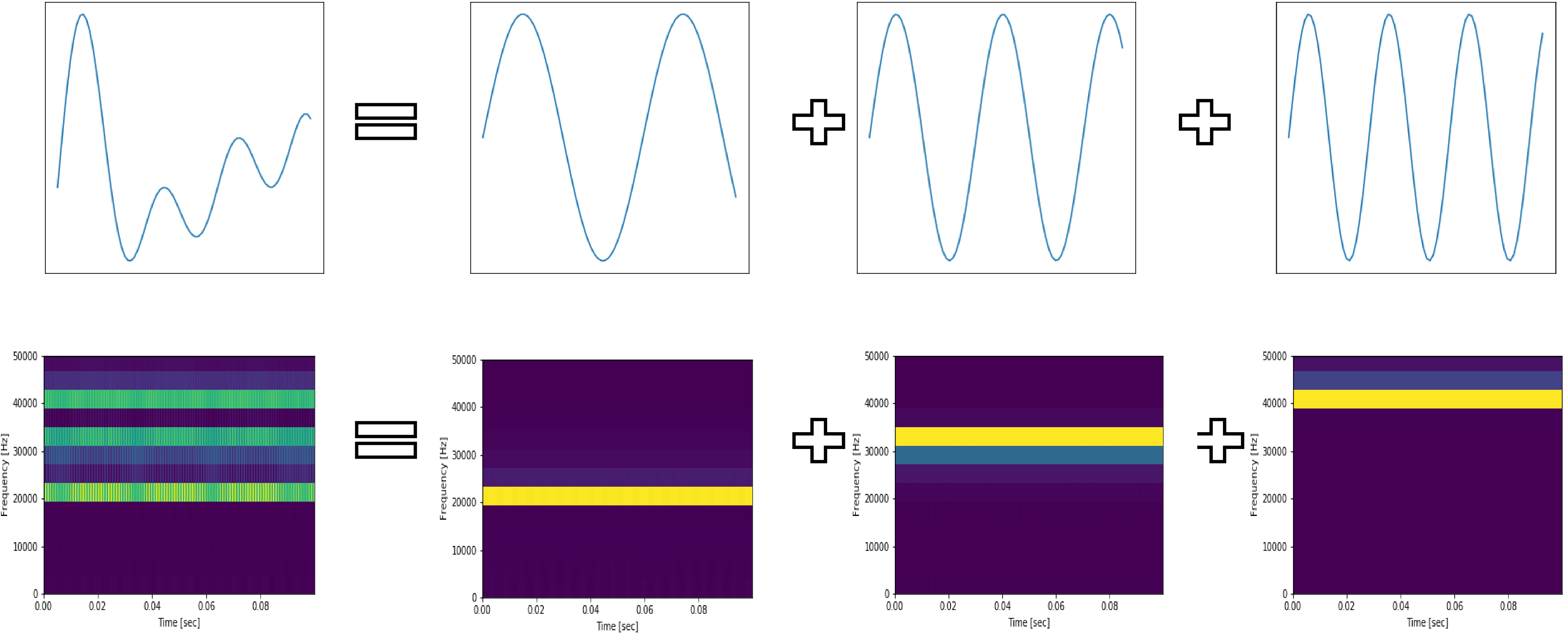
Any wave is a combination of many sine waves





Code Available

Any wave is a combination of many sine waves

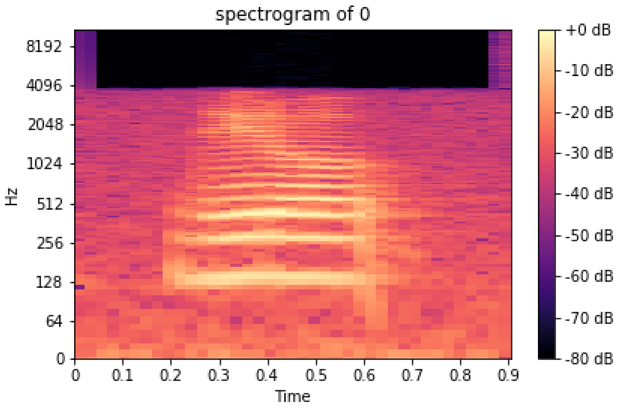


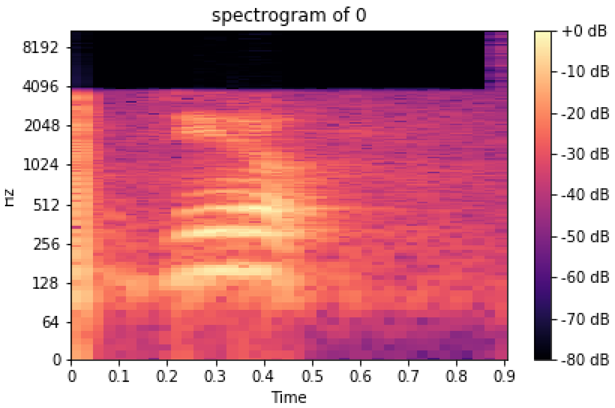
Example Problem

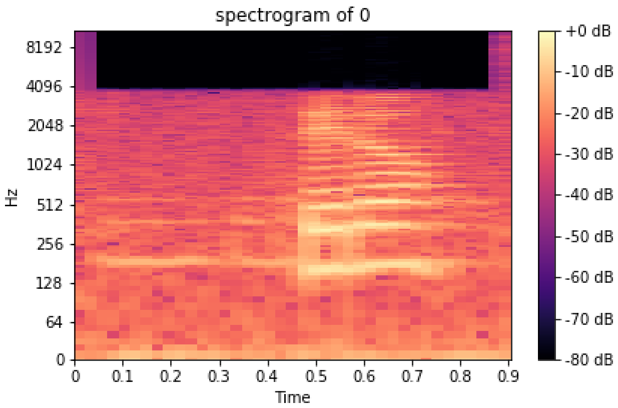
- Sound waves (.wav files)
- 10 short commands (“zero”, “one”, “two”)
- 1 sec duration
- 5000 samples (many people)

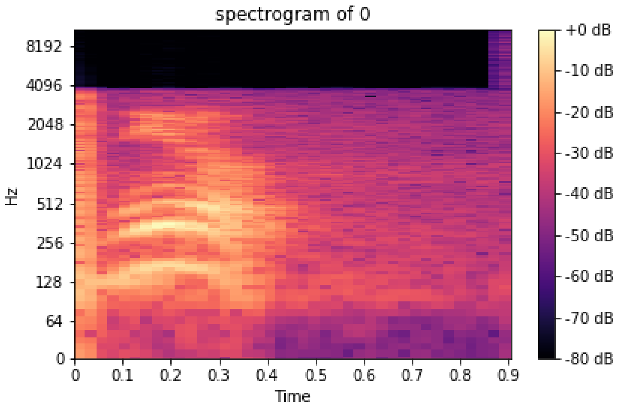
Code Available

Utterance of the Word Zero



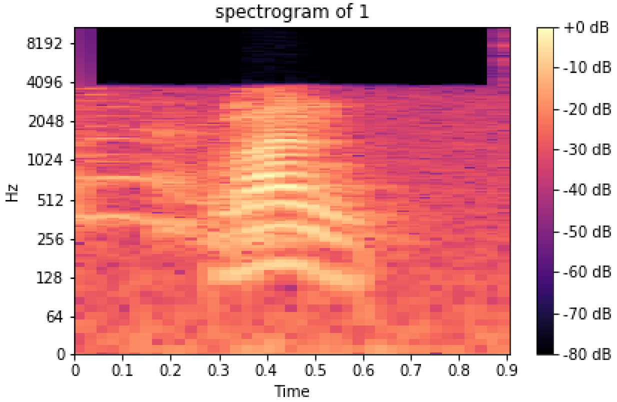


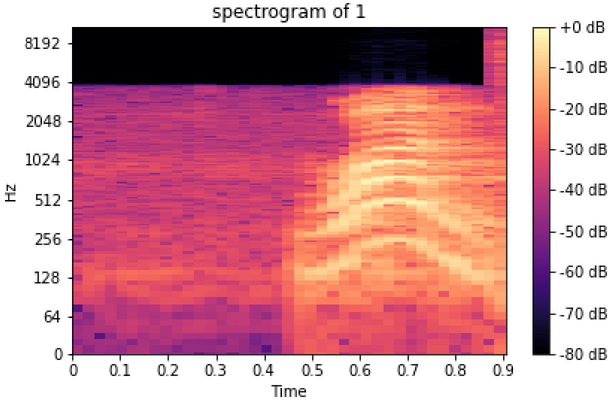


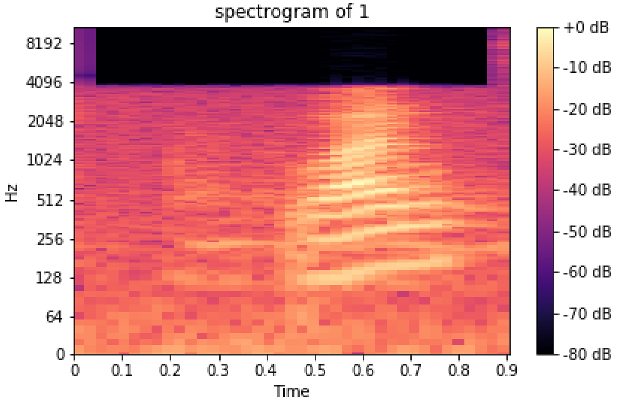


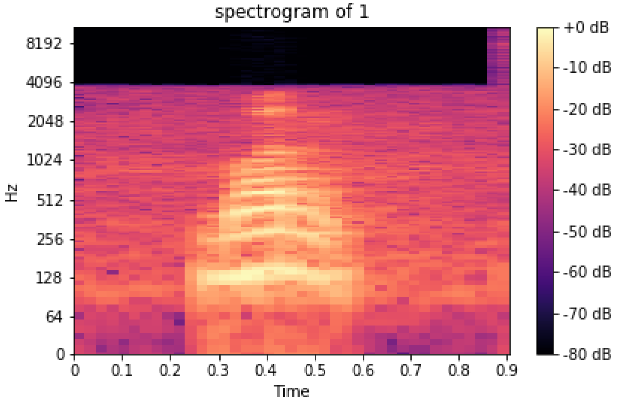
Code Available

Utterance of the Word One



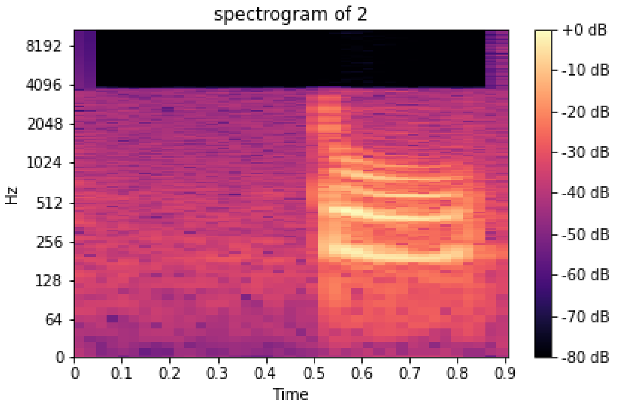


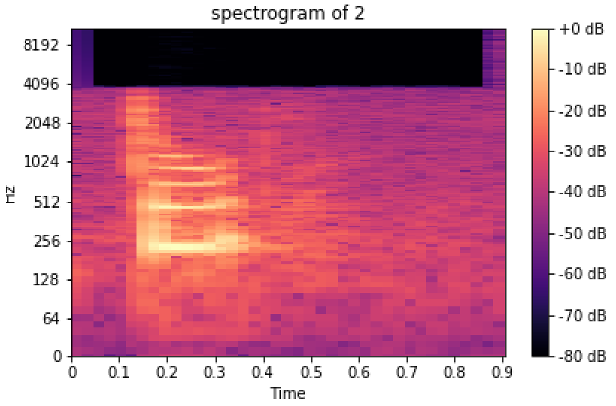


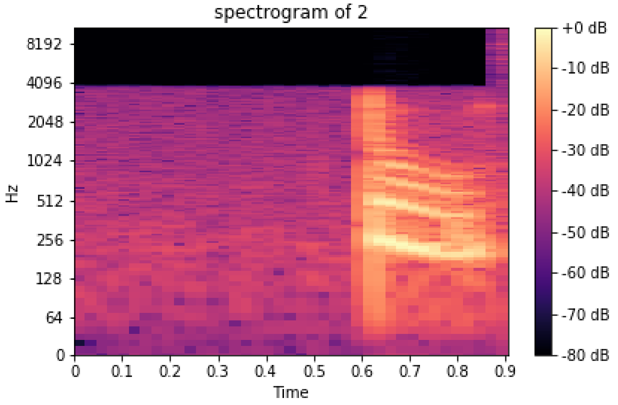


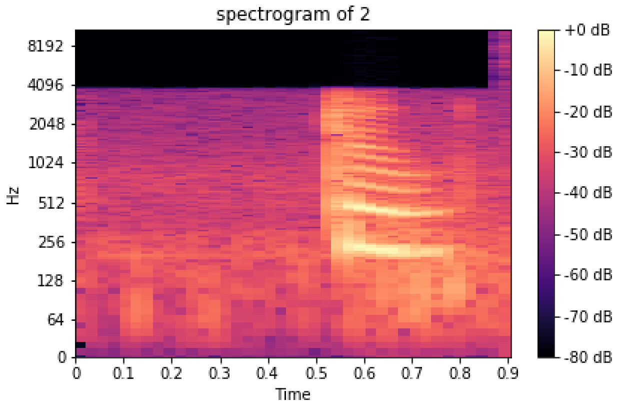
Code Available

Utterance of the Word Two

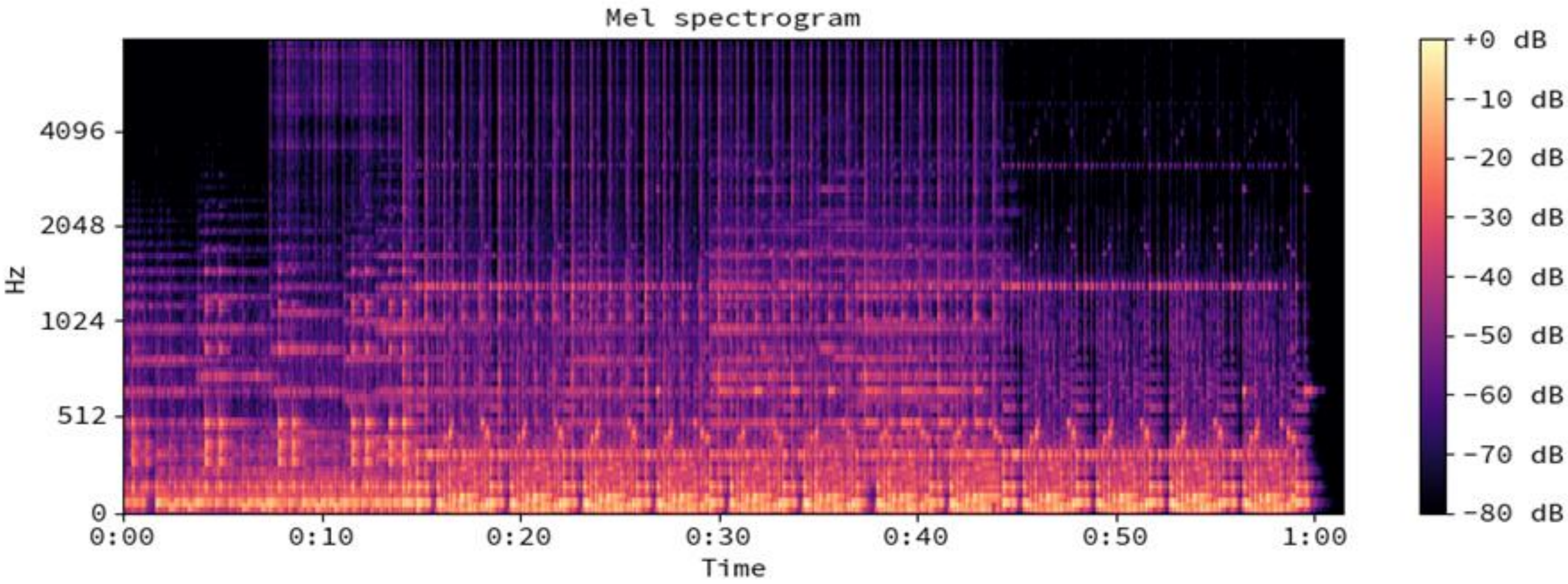








Features from Mel Spectrogram



MFCC
(Hand coded Classic Features)

VGG19-Features
(Trained on Mel spectrograms)

http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

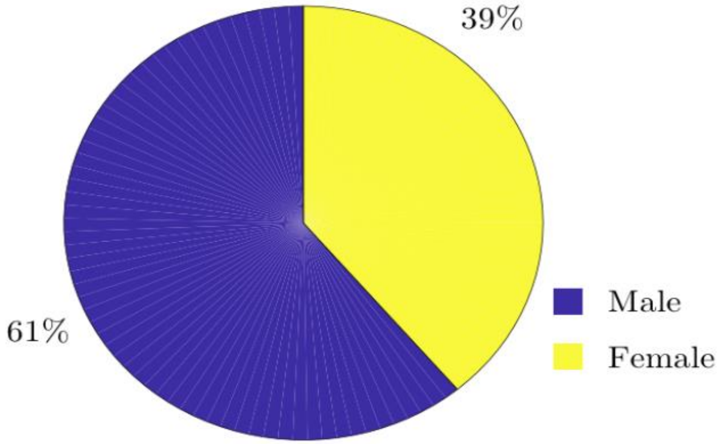
Example problem

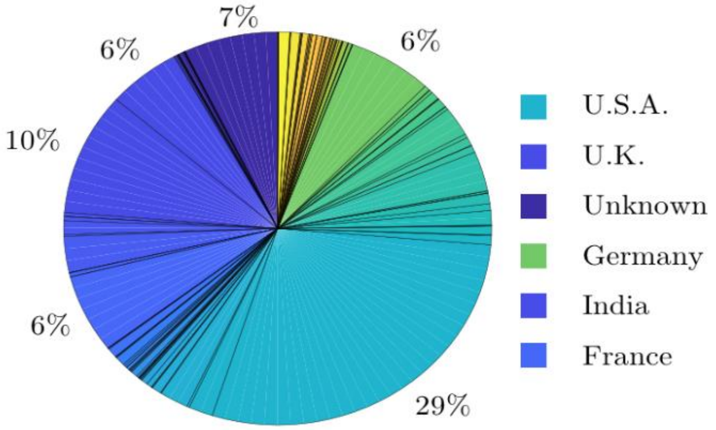
VoxCeleb

A large scale audio-visual dataset of human speech

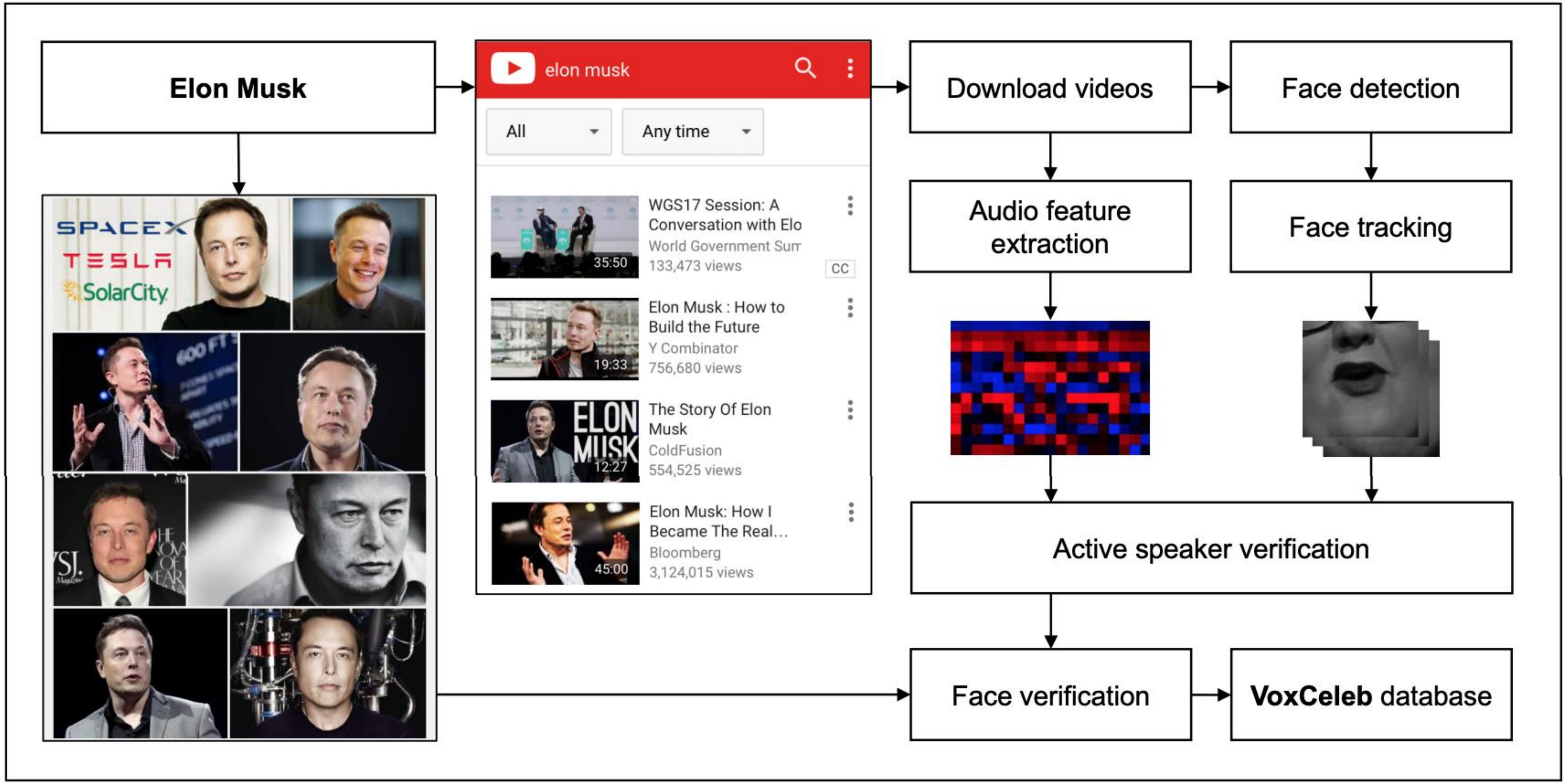
VoxCeleb2

VoxCeleb2 contains over a million utterances for 6,112 identities.





Example problem



elon musk

All

Any time

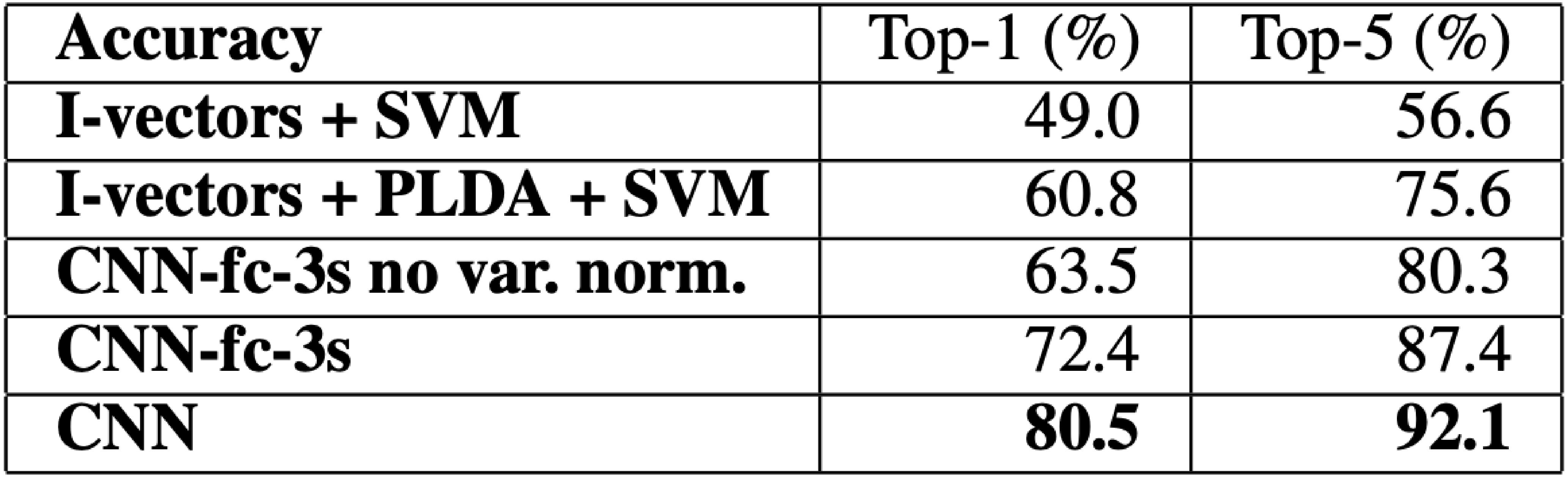
WGS17 Session: A Conversation with Elo
World Government Sum
133,473 views
35:50
CC

Elon Musk : How to Build the Future
Y Combinator
756,680 views
19:33

The Story Of Elon Musk
ColdFusion
554,525 views
12:27

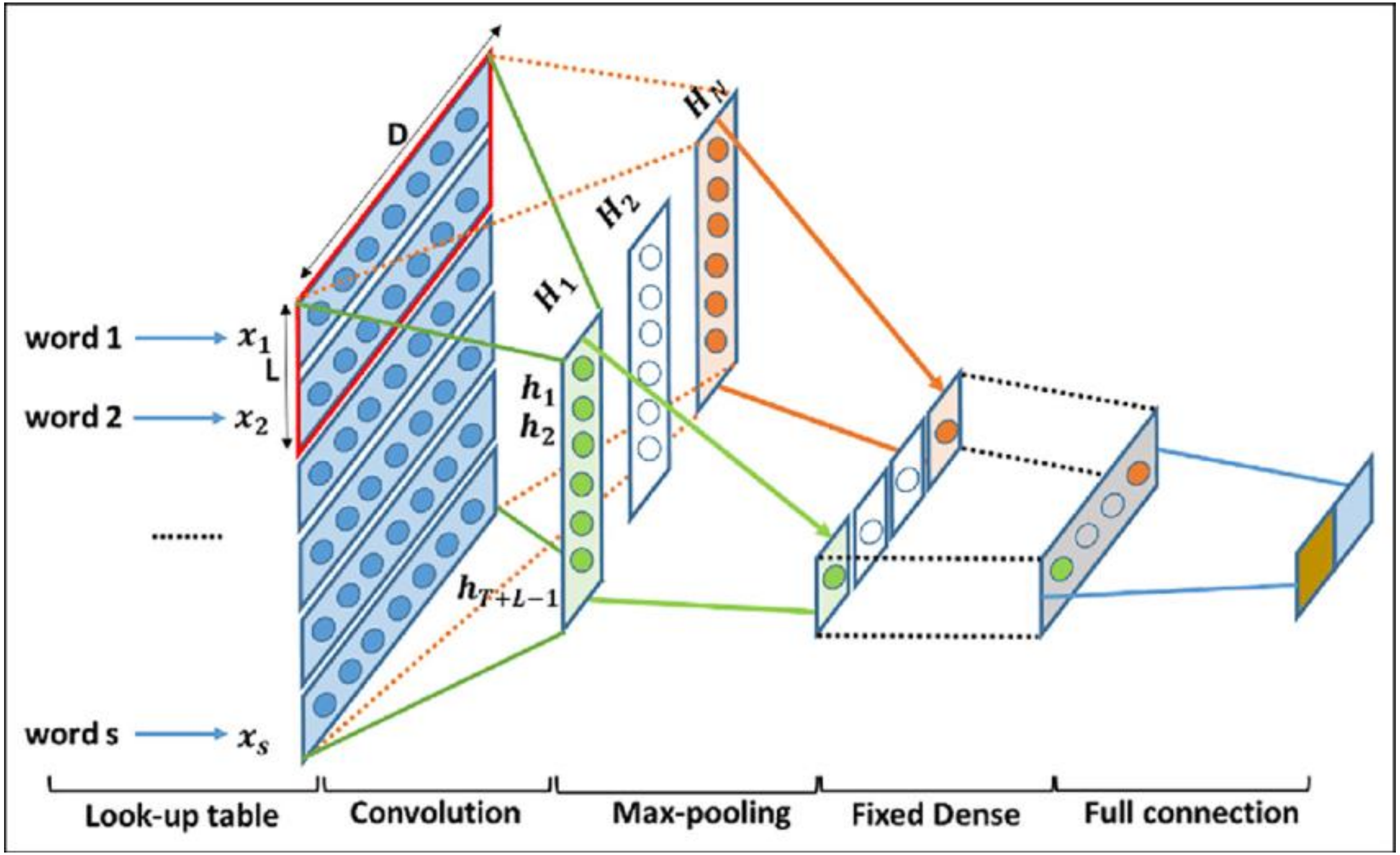
Elon Musk: How I Became The Real…
Bloomberg
3,124,015 views
45:00

Performance on VoxCeleb



Accuracy	Top-1 (%)	Top-5 (%)
I-vectors + SVM	49.0	56.6
I-vectors + PLDA + SVM	60.8	75.6
CNN-fc-3s no var. norm.	63.5	80.3
CNN-fc-3s	72.4	87.4
CNN	**80.5**	**92.1**

Neural Networks + word2vec for text



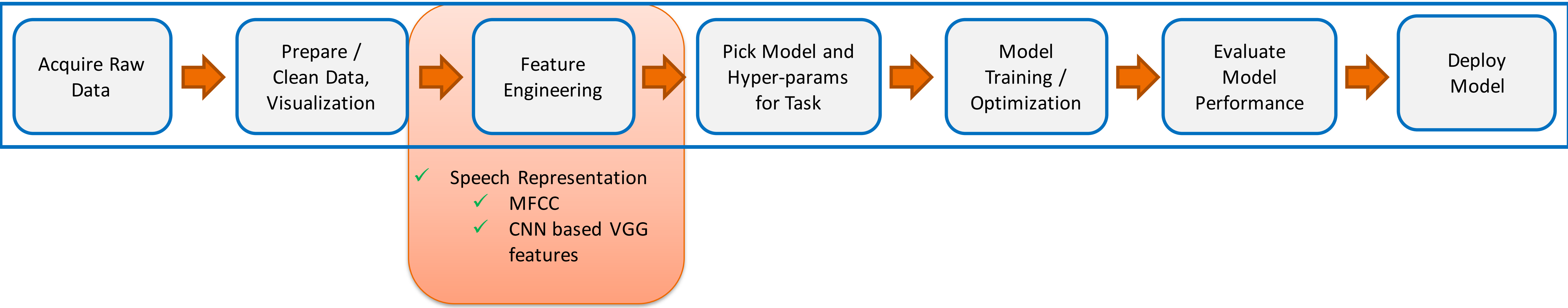
Sentiment Classification
(Positive / Negative)

https://www.researchgate.net/publication/333752473/figure/fig1/AS:769346934673412@1560438011375/Standard-CNN-on-text-classification.png

Summary

- Data driven features are now effective for many data.
  - “Learn from some one else’s data”.
  - “Refine to your problem” (more later)
- Many recognition/classification tasks in the image and speech space are reachable.

Summary



Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- ✓ Speech Representation
  - ✓ MFCC
  - ✓ CNN based VGG features

Thanks!!

Questions?