AV 331 : Digital Signal Processing Lab

Project

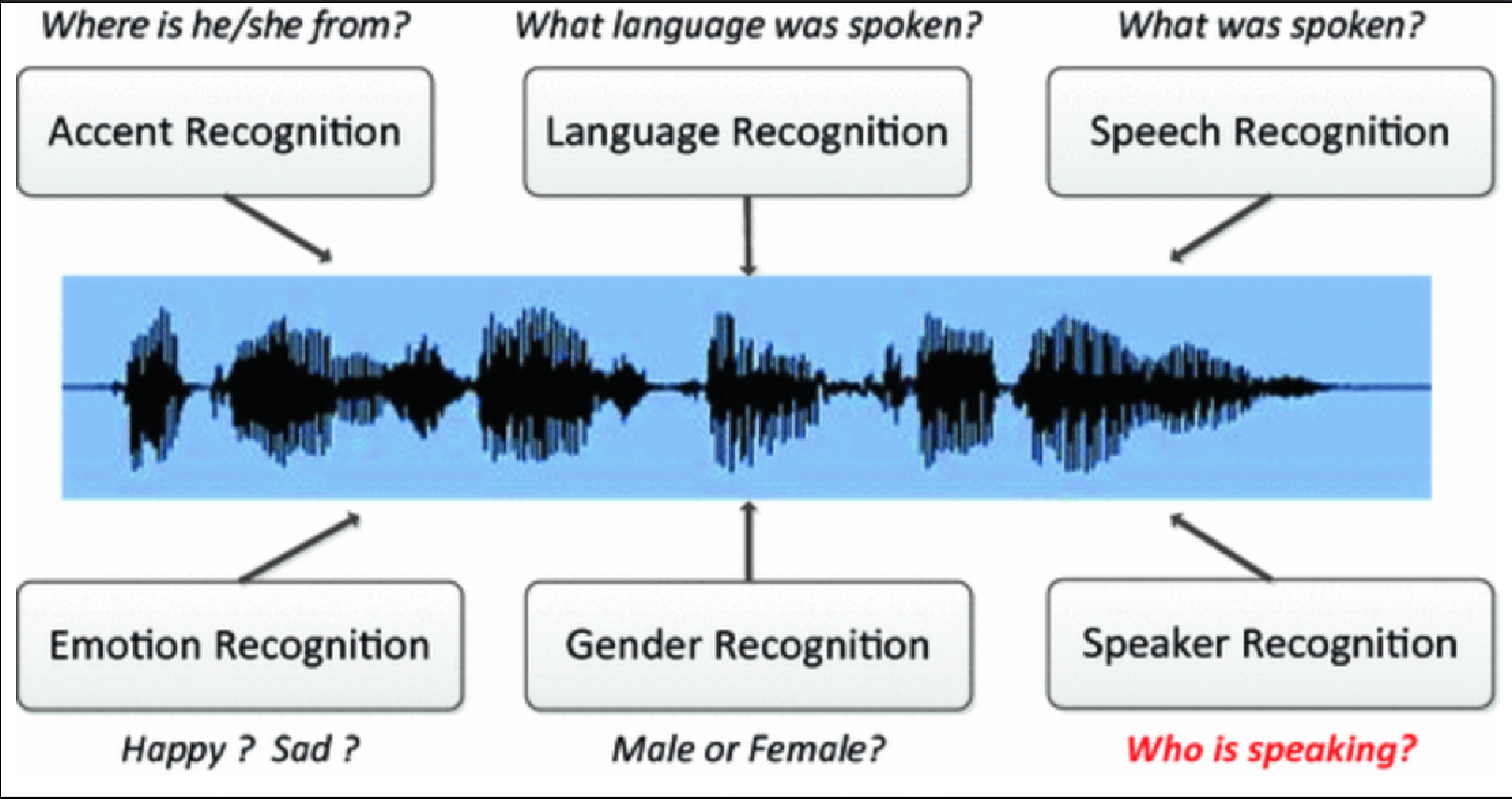
Automatic Speaker Recognition

Sri Aditya Deevi (SC18B080)

ECE (AVIONICS)

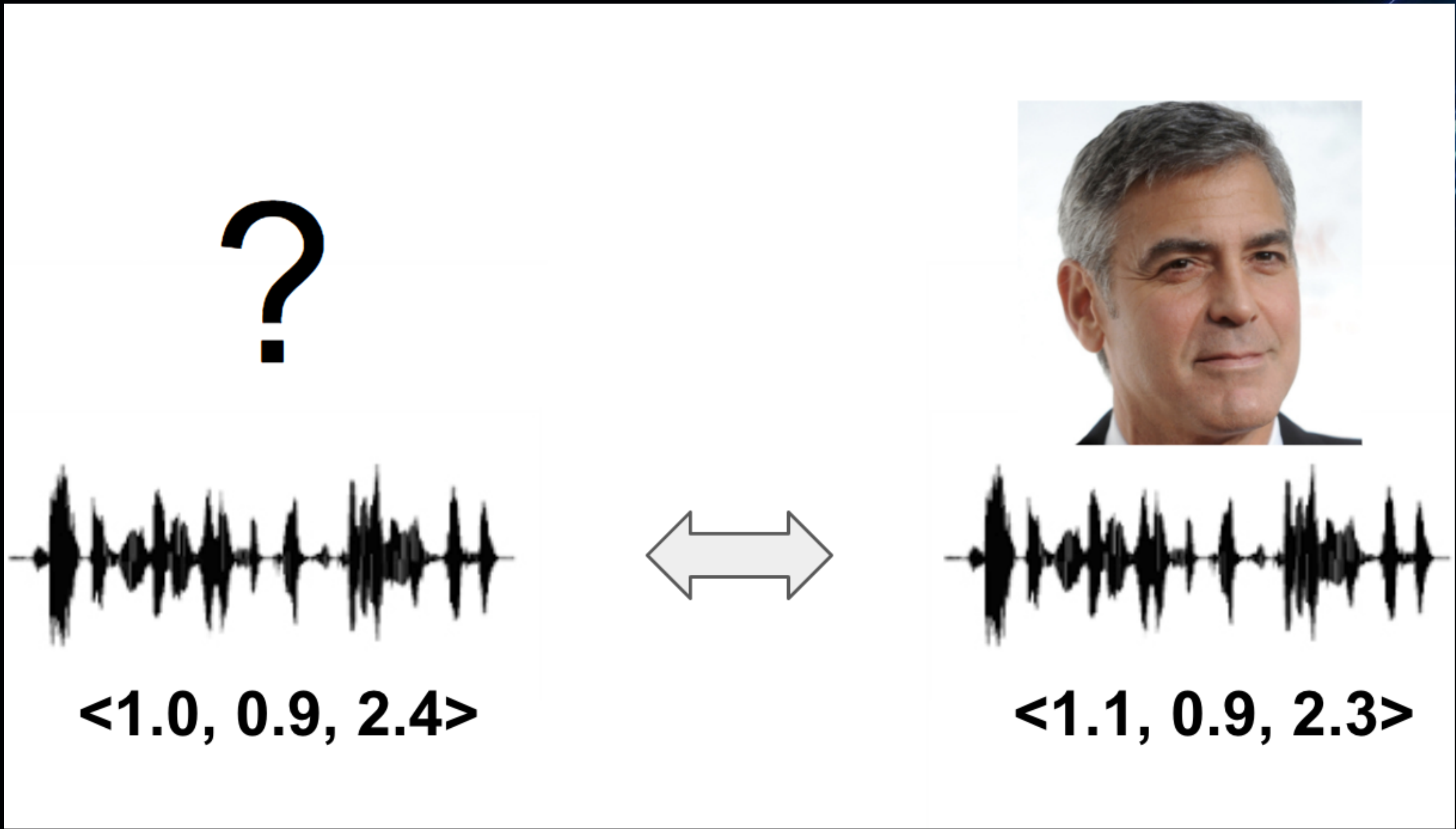
Indian Institute of Space science and Technology

Human Speech Analysis



Speaker Recognition

- Process of automatically recognizing who is speaking on the basis of individual information included in speech waves



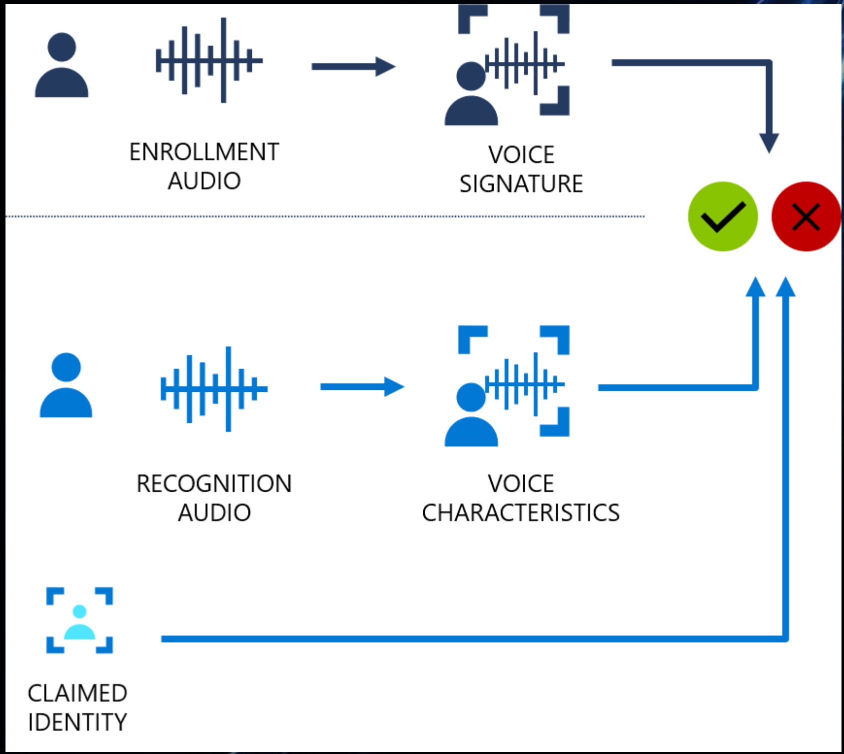
Why Speaker Recognition?

- Some applications are :

Identity verification for telephone banking, database access services, extra security control

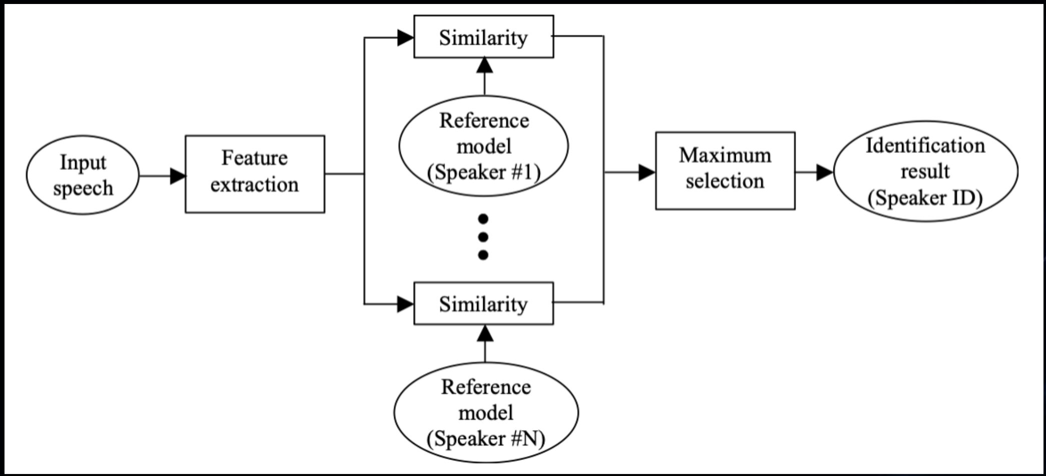
in remote computer access,

laboratory access etc.

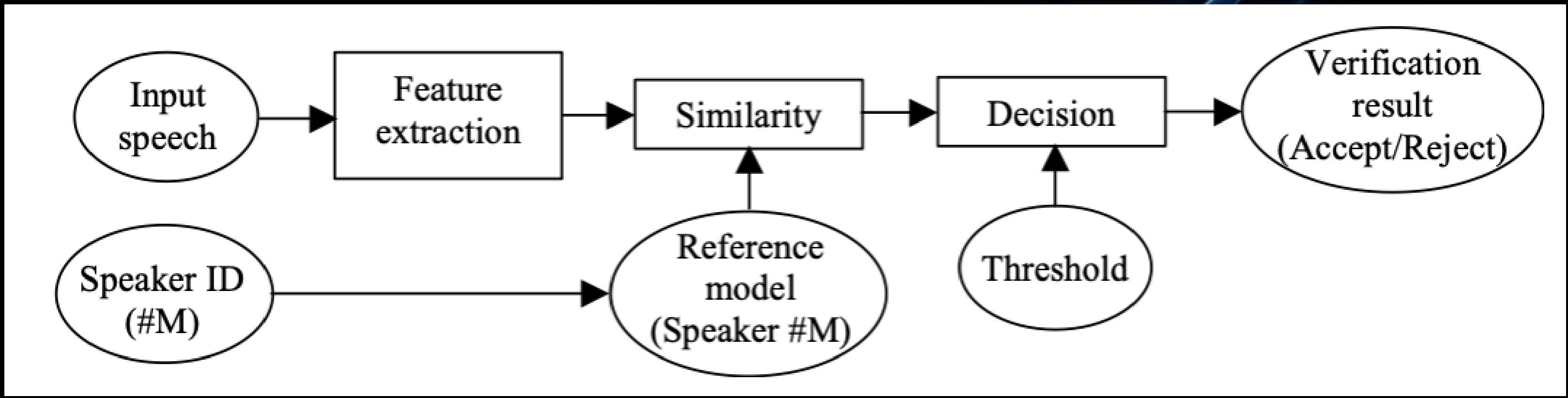


Types of Speaker Recognition

(a) Speaker Identification



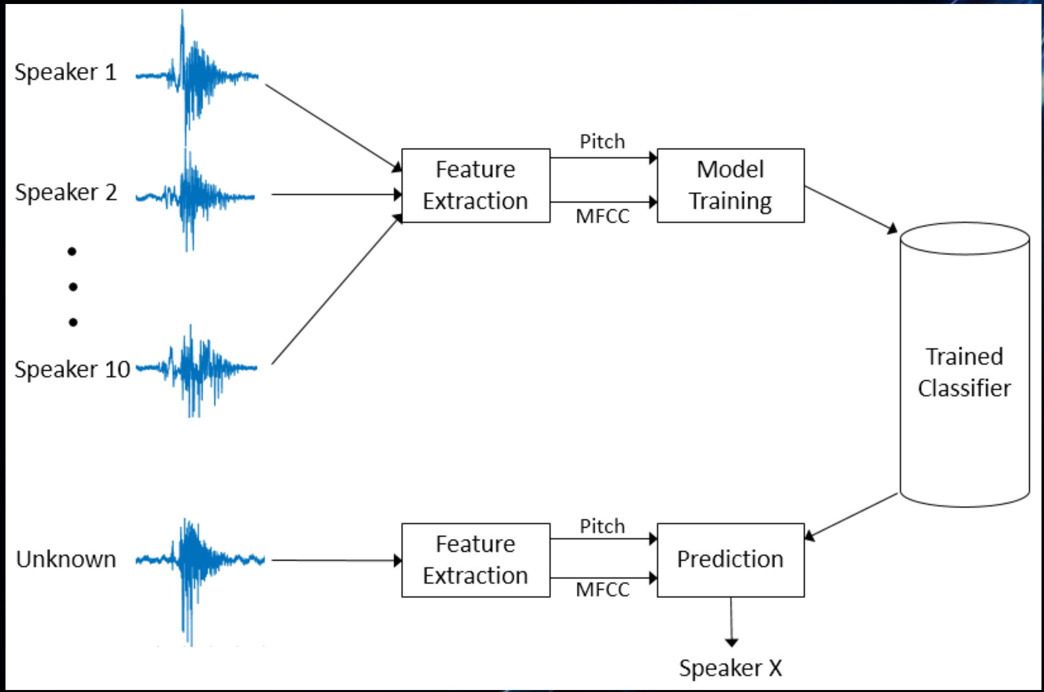
(b) Speaker Verification



Phases of ASR System

- **Training/Enrolment Phase :**

  Samples of speech from registered speakers to build specific reference models



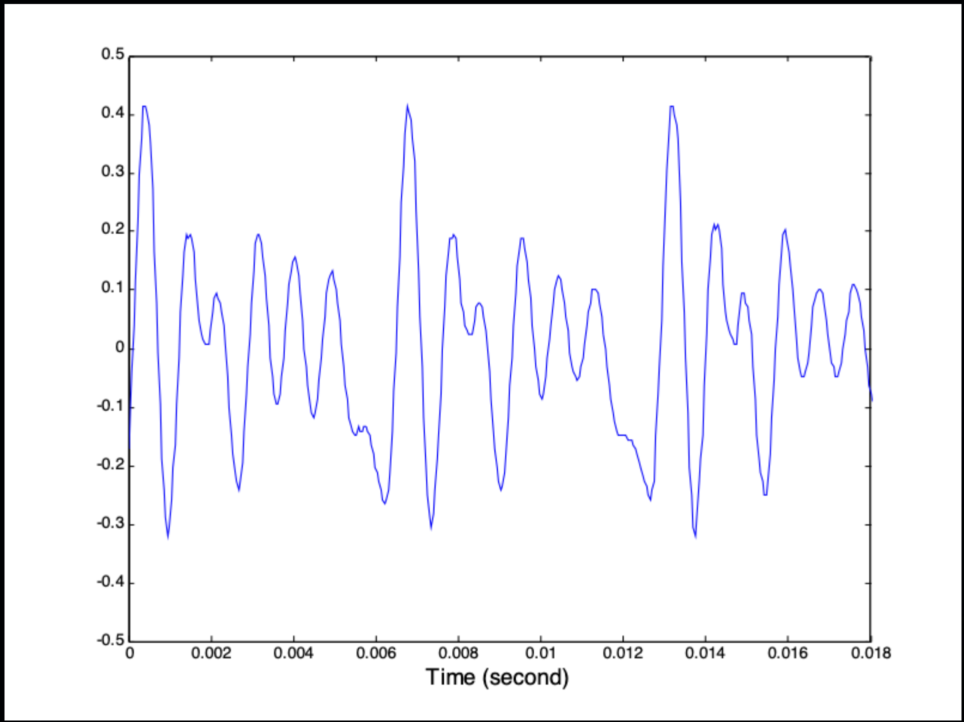
- **Testing Phase :**

  Unknown samples of speech matched with the bank of speaker reference models

*In this work, we are considering Speaker Identification systems

Speech Feature Extraction

- First stage in both Enrolment and Testing phases.



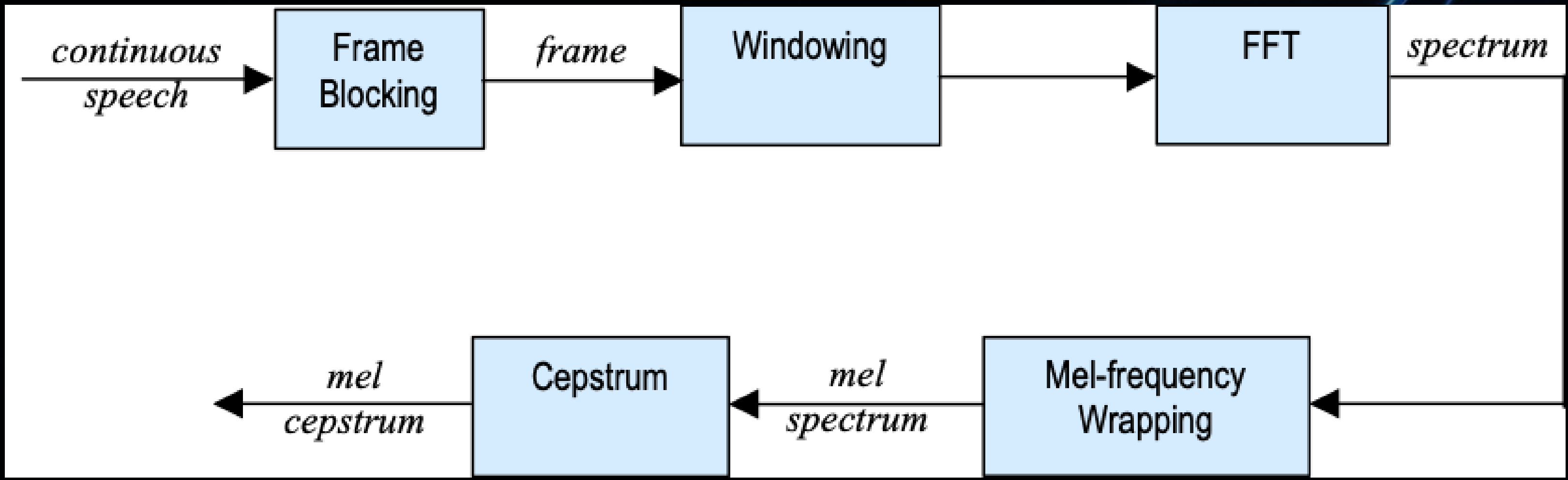
Speech signal is "Quasi-stationary"

Can be characterised by Short-Time Spectral Analysis

Mel Frequency Cepstrum Coefficients

"Perceptually-Relevant Time-Frequency Representation"

- Can be used for characterization of speech (frame).
- MFCC's can be calculated as follows :

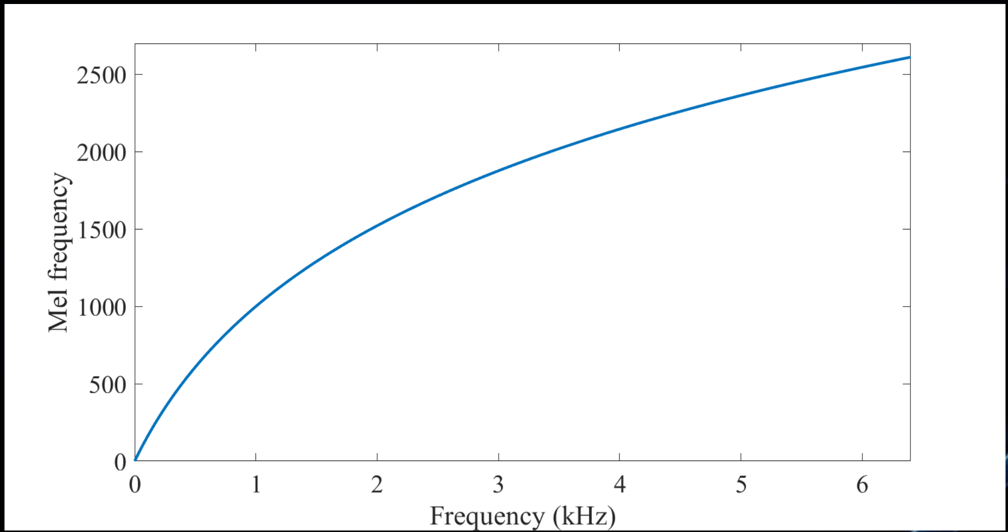


Mel Frequency

Cepstrum Coefficients

Mel Frequency Scale

- From Psychoacoustic experiments, scientists concluded that human beings perceive frequencies logarithmically.



$$mel = 2595 * log_{10}(1 + \frac{f}{700})$$

Mel Filter Banks

- Convert the frequency scale of STFT (Short Time Fourier transform) to Mel-Scale to represent the perceptual difference.

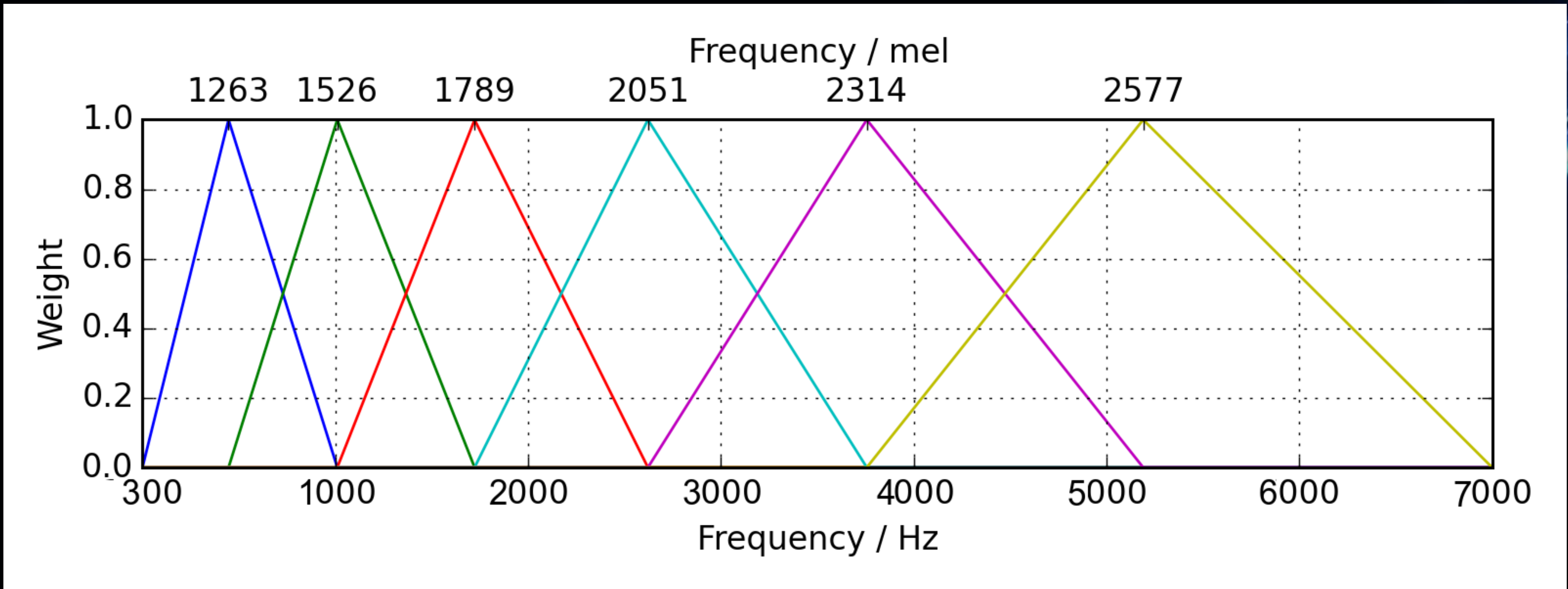
(Mel Frequency Wrapping)

- A Mel Filter Bank consists of a number Mel bands.

(which is a hyper-parameter that needs to be optimized)

Here, we consider No. of Mel Bands = **26**

Mel Filter Banks



Human Voice Frequency : (300 - 7000) Hz

$$f_s = 12500Hz$$

(To avoid aliasing)

Mel Frequency Cepstrum Coefficients

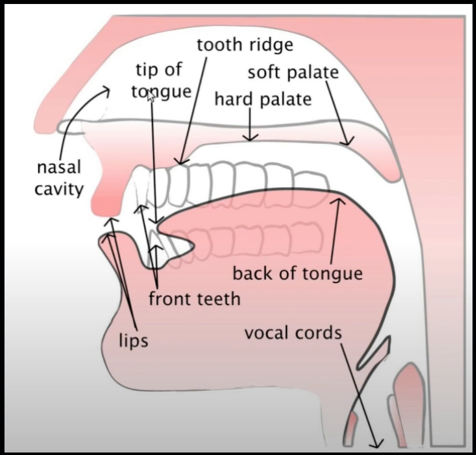
Cepstrum

Cepstrum	Quefrency	Liftering	Rhamonic
↓	↓	↓	↓
Spectrum	Frequency	Filtering	Harmonic

$$C(x(t)) = F^{-1}(log(F(x(t))))$$

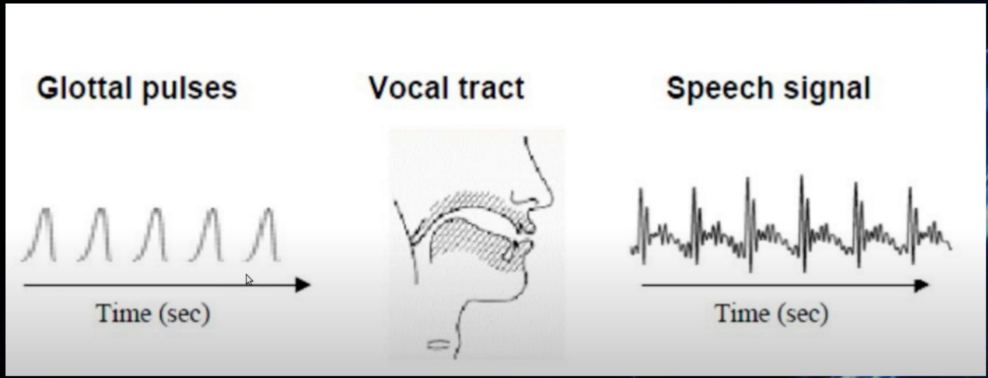
Understanding Human Speech Generation

Speech is generated by Vocal Tract



(Speech Carrier)

Glottal Pulses generated by Vocal Cords



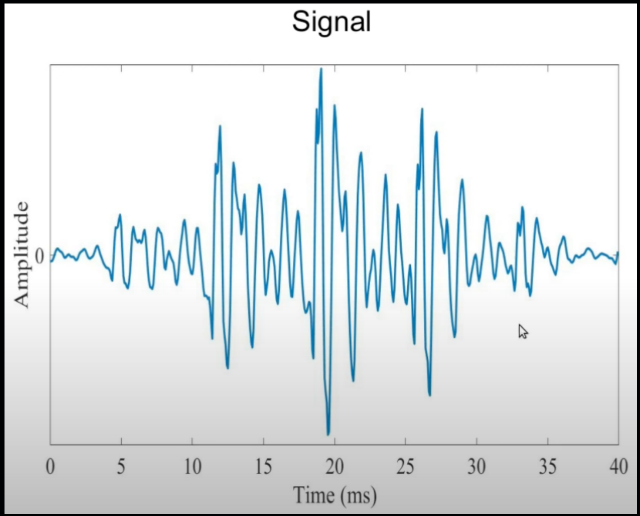
- Vocal Tract can be modelled as a filter.

$$Speech = h(t) * g(t)$$

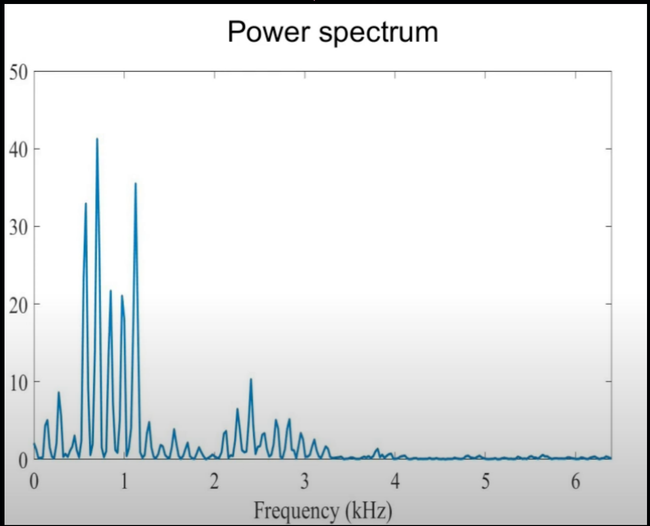
g(t) -> Glottal Pulses

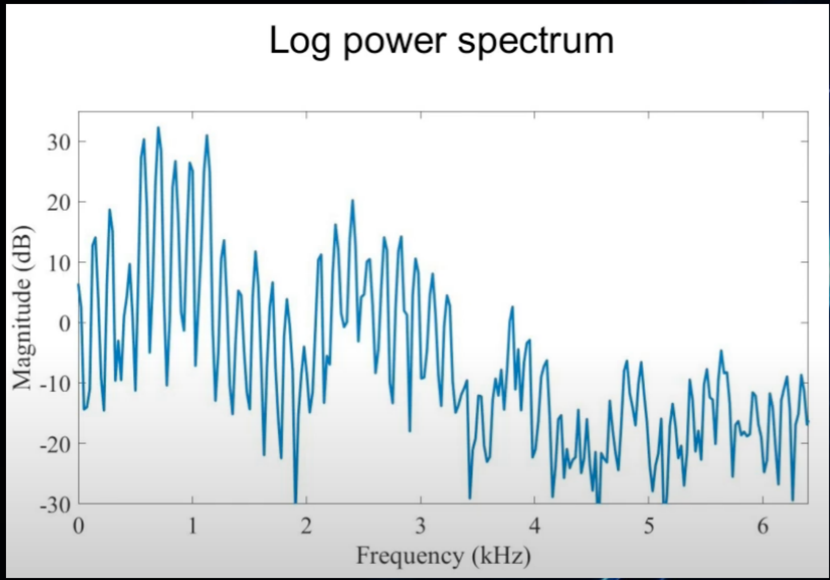
h(t) -> Impulse Response of Vocal Tract

Understanding Cepstrum



DFT





Understanding Cepstrum

$$X(f) = H(f).G(f)$$

$$X(f)_{dB} = H(f)_{dB} + G(f)_{dB}$$

Speech

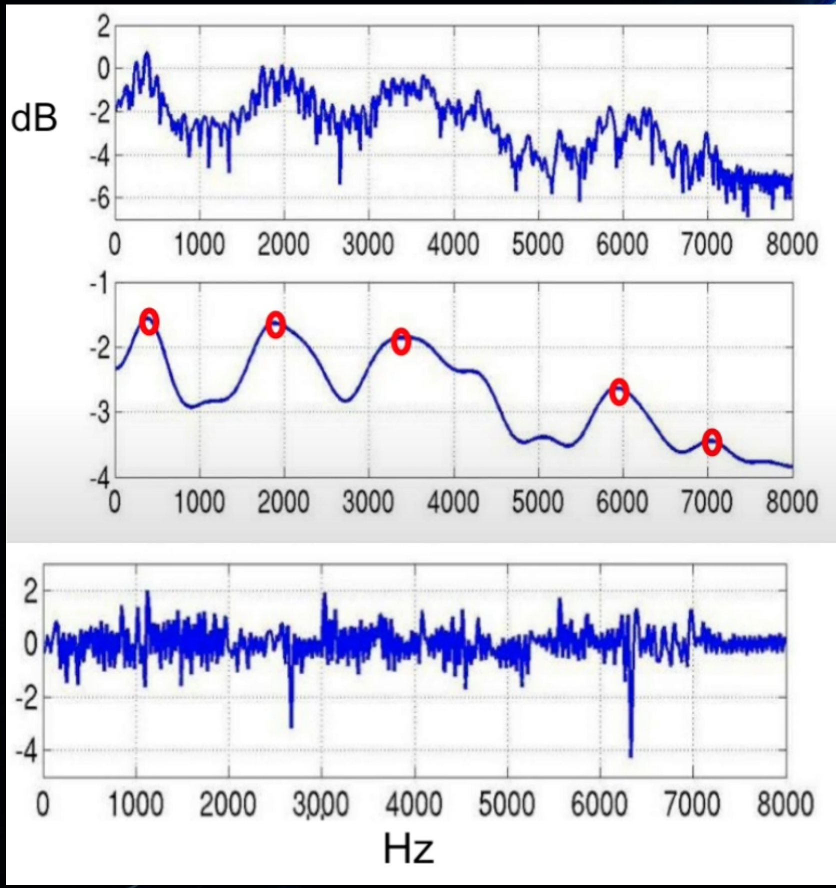
(Log Spectrum)

H(f) (in dB)

(Spectral Envelope)

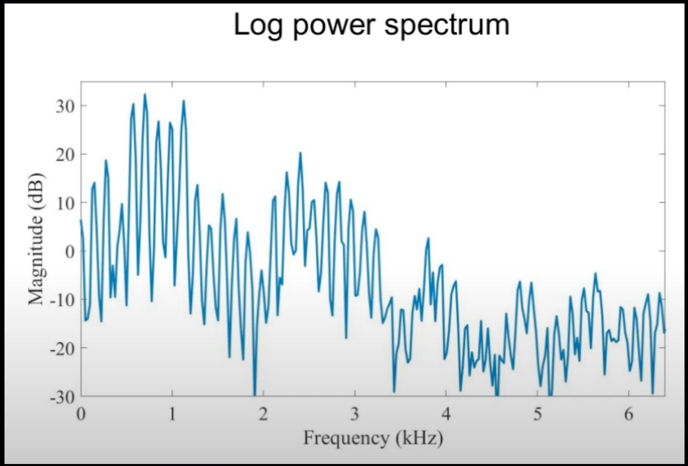
G(f) (in dB)

(Spectral Detail)



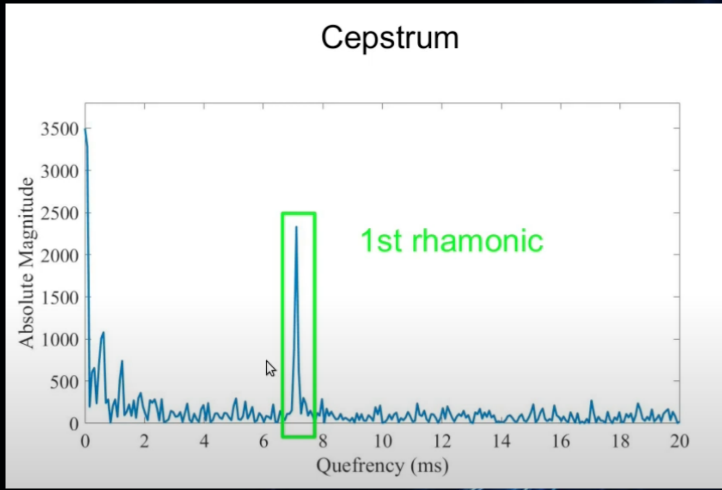
Formants carry the identity of sound unique to speaker

Understanding Cepstrum

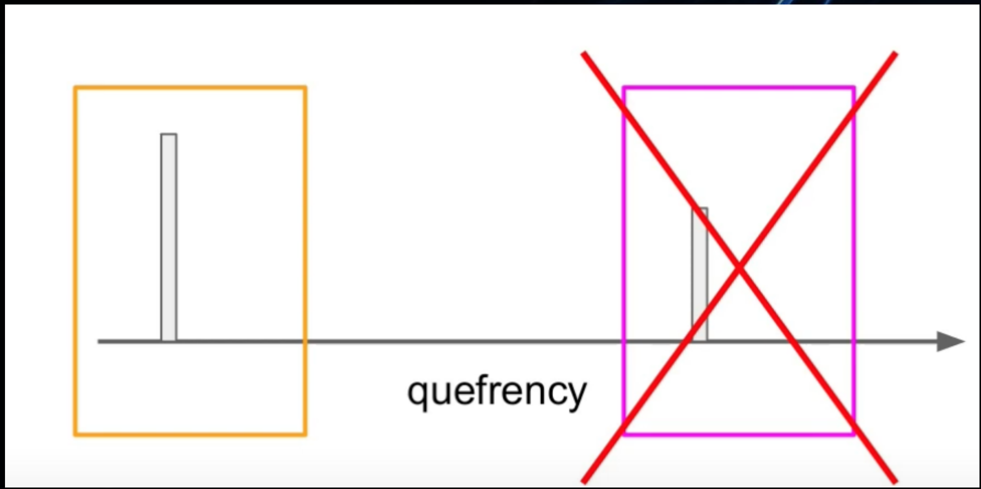


Inverse

Fourier Transform



Choose intermediate set of coefficients (Ex: 2-20)



Reject 0 quefrency and very high quefrencies

Implementation Flow

(Feature Extraction)

Speech Signal

fs = 12500 Hz

Short-Time Fourier Transform

Frame Size = 256 samples

Overlap = 156 samples

Window = Hamming

Convert to Power Spectrum

Mel-Frequency Wrapping

Number of Mel Banks = 26

Applying Logarithm

Take Discrete Cosine Transform

MFCC

Choose Intermediate Quefrencies

Why DCT ?

$$\tilde{c}_n = \sum_{k=1}^{K} (log\tilde{S}_k) cos[n(k - \frac{1}{2})\frac{\pi}{K}]$$

$$n = 0, 1, 2, ..., K - 1$$

where $\tilde{S}_k$ = Mel power spectrum coefficients

K = 26

ADVANTAGES

- Real Coefficients
- Decorrelation of Energy in Mel Bands
- Basis Functions are cosines
- Dimensionality Reduction

Final Remark on MFCC's

- The Frame Size considered is 256 samples (~21 msec)

- An ***acoustic vector*** (MFCC Coefficients) is computed for each frame and stored as an ***acoustic matrix***.

CHARACTERISTIC of the speaker

- The dimensions of the acoustic matrix are :

$$(\#MFCC_{coefficients}, \#Frames)$$

Speech Feature Matching

- Problem at hand is that of "Supervised Pattern Recognition"

Working Principle : Based on the assumption, that acoustic vectors are a unique feature representation of a speaker's voice.

- In this project, **Vector Quantization (VQ)** is considered for Feature Matching.

Vector Quantization

*Acoustic vectors here

- Process of mapping vectors from a large vector space to a finite number of regions in that space.

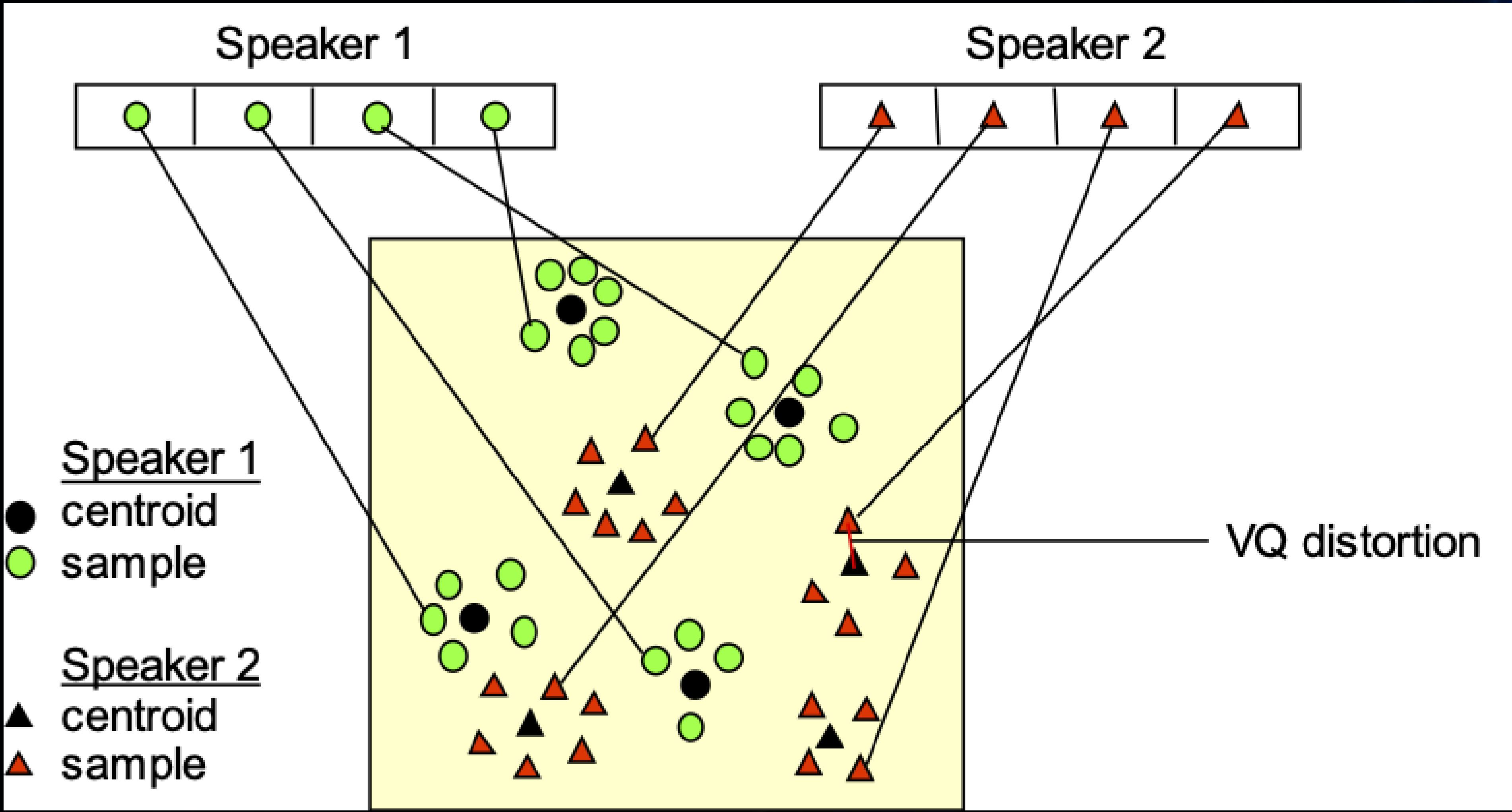
  - Each such region is called a "**Cluster**".

  - Centroid of cluster is called a "**Codeword**".

  - Collection of all codewords is called "**Codebook**".

- In this project, ***speaker-specific*** codebooks are considered.

Vector Quantization



Building Speaker-specific Codebooks

Speech (Particular Speaker)

▼

MFCC Processing

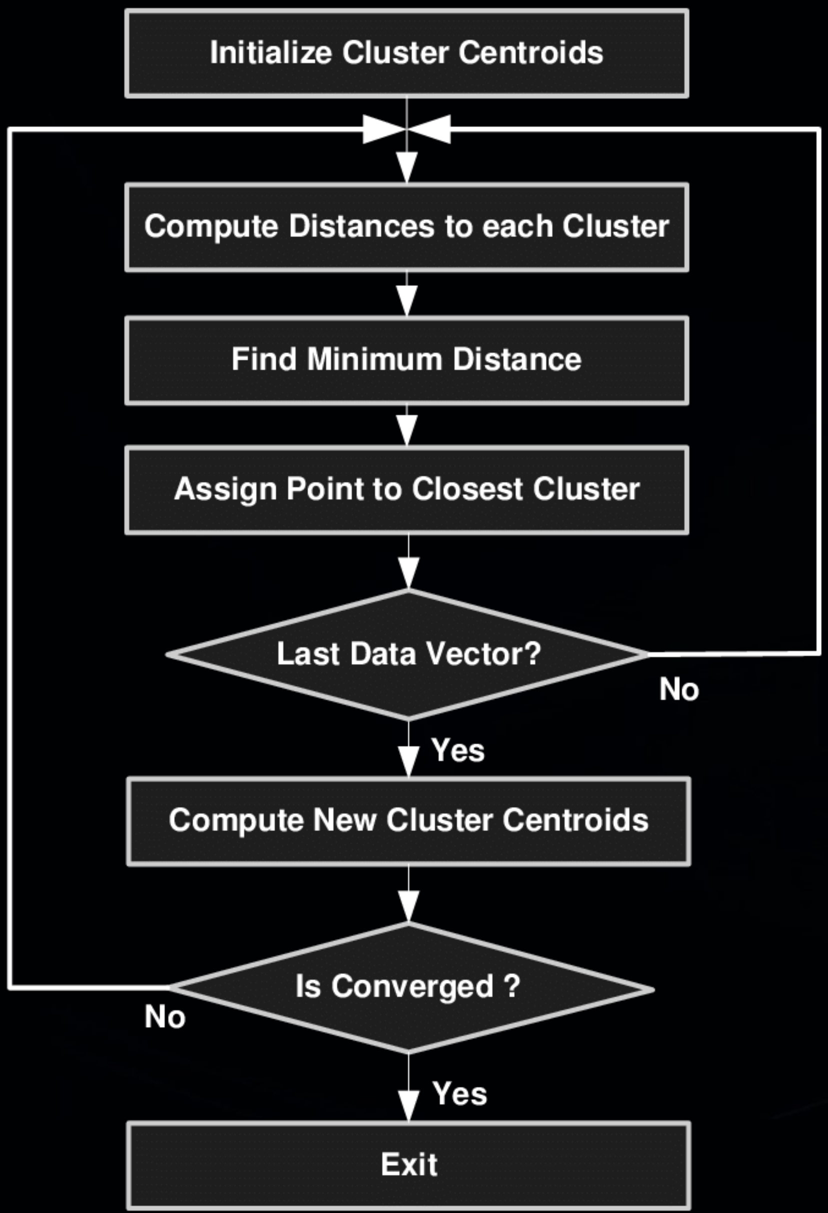
▼

Acoustic Vectors

▼

Codebook Formation by clustering using **K-Means**

K-Means



- Unsupervised ML Algorithm

Number of clusters

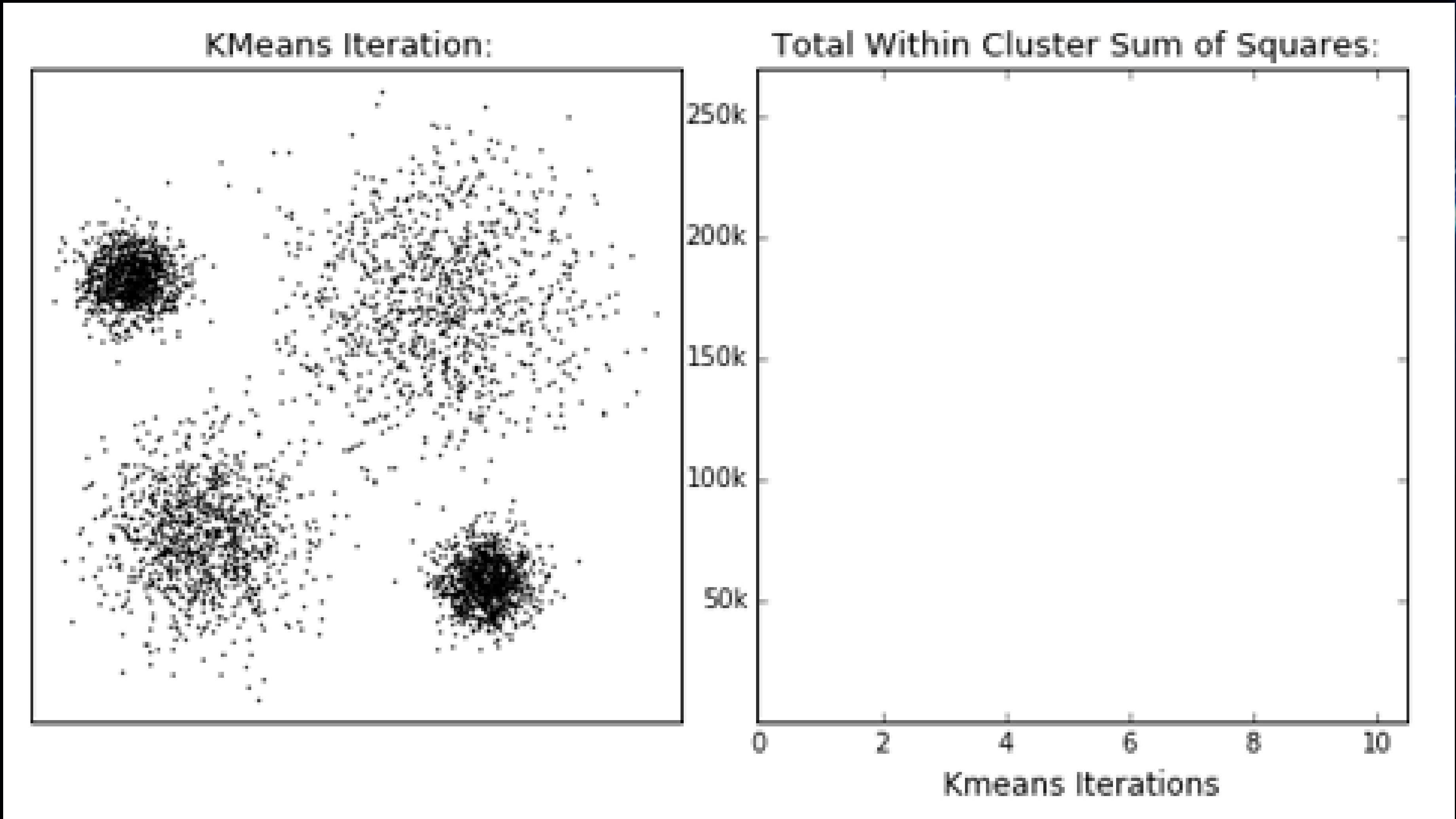
To be optimized

*Here, Clusters = 16

"Initialization Sensitive"

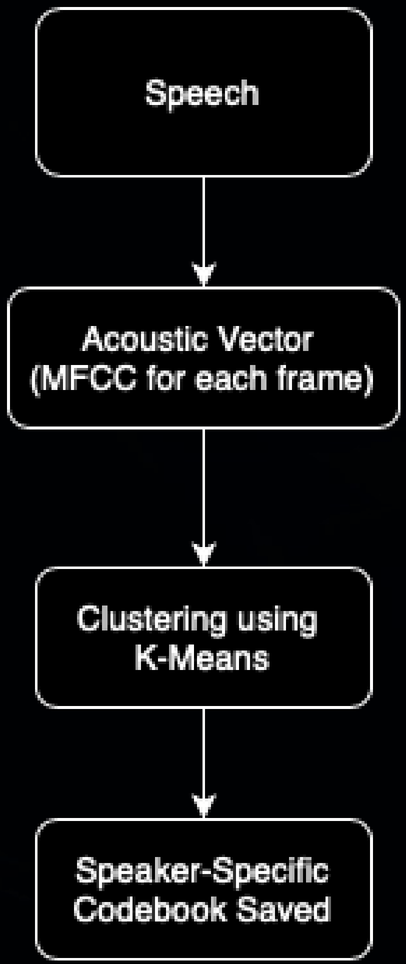
Multiple Random Initializations

*Here, N = 10

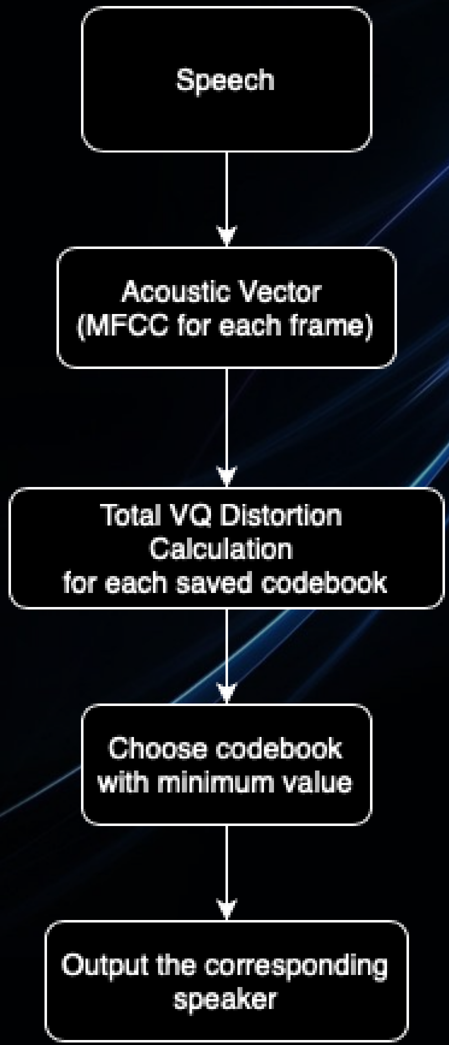

KMeans Iteration:
Total Within Cluster Sum of Squares:
250k
200k
150k
100k
50k
0
2
4
6
8
10
Kmeans Iterations

Flow of Various Phases

Enrolment Phase



Testing Phase



A Note on the Dataset

- The training data consists of 13 audio files (13 different speakers) of 2 second duration sampled at 12500 Hz.

- All the speakers are asked to say "zero" for the experiment.

- The test data also contains similar format audio files (13 in number).

- The test audio files are collected from the speakers after some time (typically some days) to simulate any minor variations in voice.

Live Demonstration and Summary of Results

THANK YOU!

References

- https://minhdo.ece.illinois.edu/teaching/speaker_recognition/
- http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/speaker_recognition.html
- http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
- https://www.youtube.com/playlist?list=PL-wATfeyAMNqIee7cH3q1bh4QJFAaeNv0