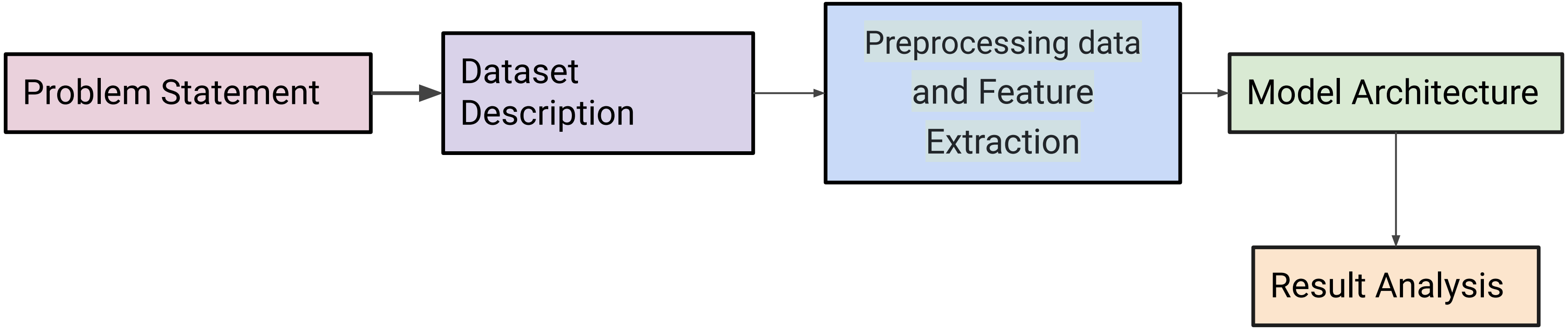
Bengali Spoken Digit Classification: A Hidden Markov Model Approach

By
Dibyendu Das
Under the Guidance Of
Dr. Sujoy Biswas

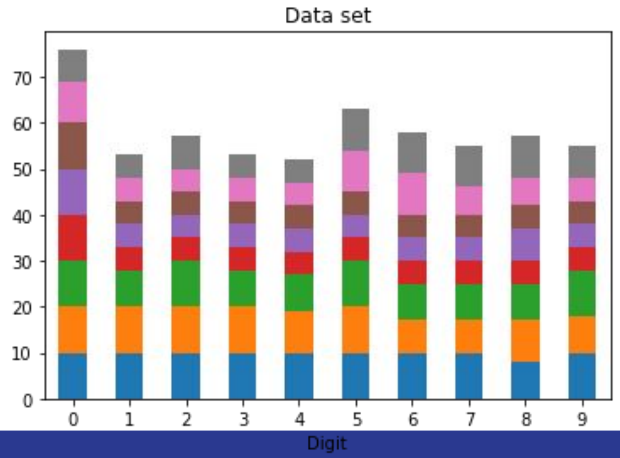
M.Sc in Big Data Analytics
Ramakrishna Mission Vivekananda Educational and Research Institute


Problem Statement
Dataset Description
Preprocessing data and Feature Extraction
Model Architecture
Result Analysis

Dataset Description

Dataset Description

- A dataset containing 600 audio file (.wav format) was created for the experiment.
- Eight people from various parts of the State were asked to give their voice recordings Using “QuickRec” App.



Bengali word	Bengali pronunciation	English word	English numerical
শূন্য	shun-no	zero	0
এক	a-k	one	1
দুই	du-i	two	2

Preprocessing and Feature Extraction

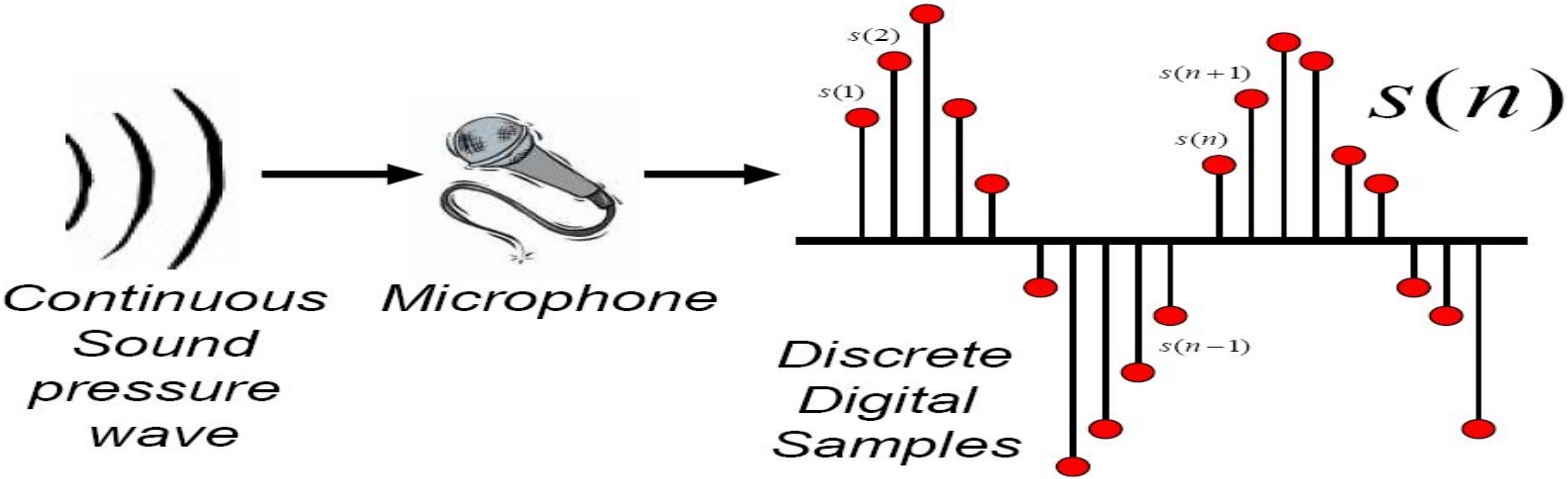
Discrete Representation of Signal :

Analog-to-digital conversion has two steps :

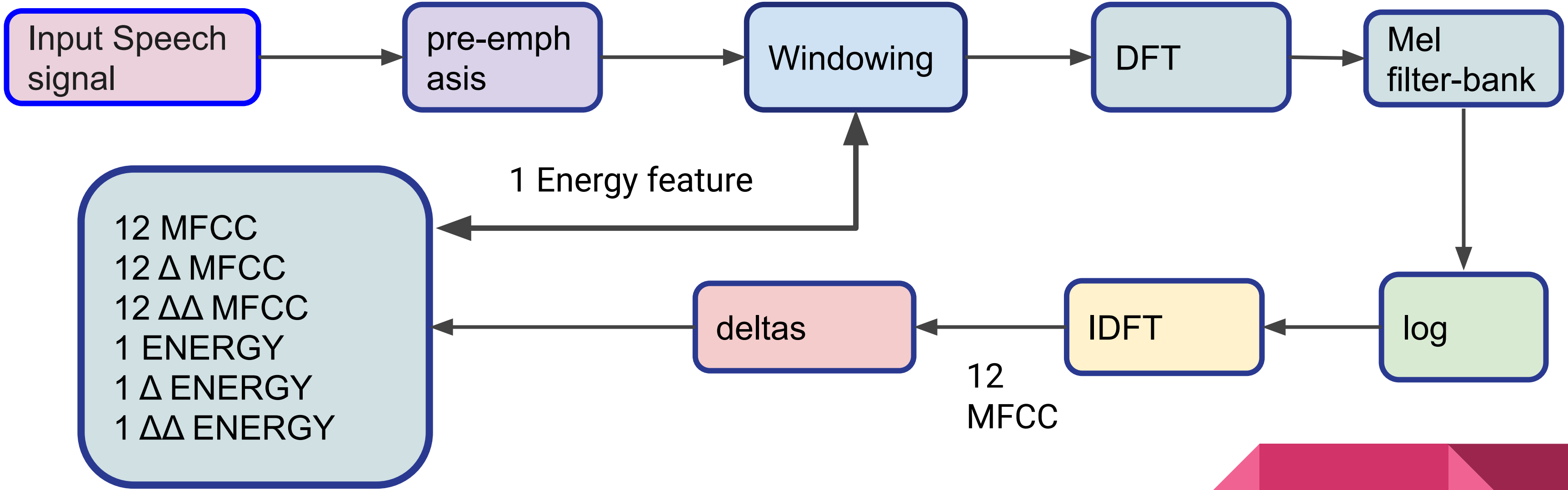
Sampling = 2 * Nyquist frequency

+

Quantization = store the amplitude value as 8 bit or 16 bit



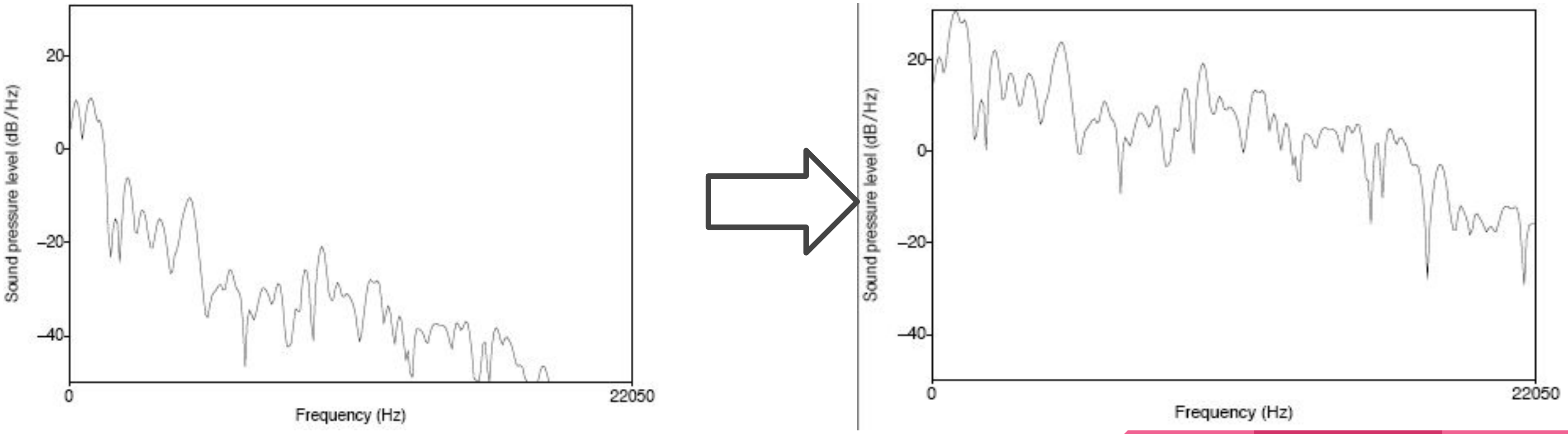
Preprocessing data and MFCC Feature Extraction :



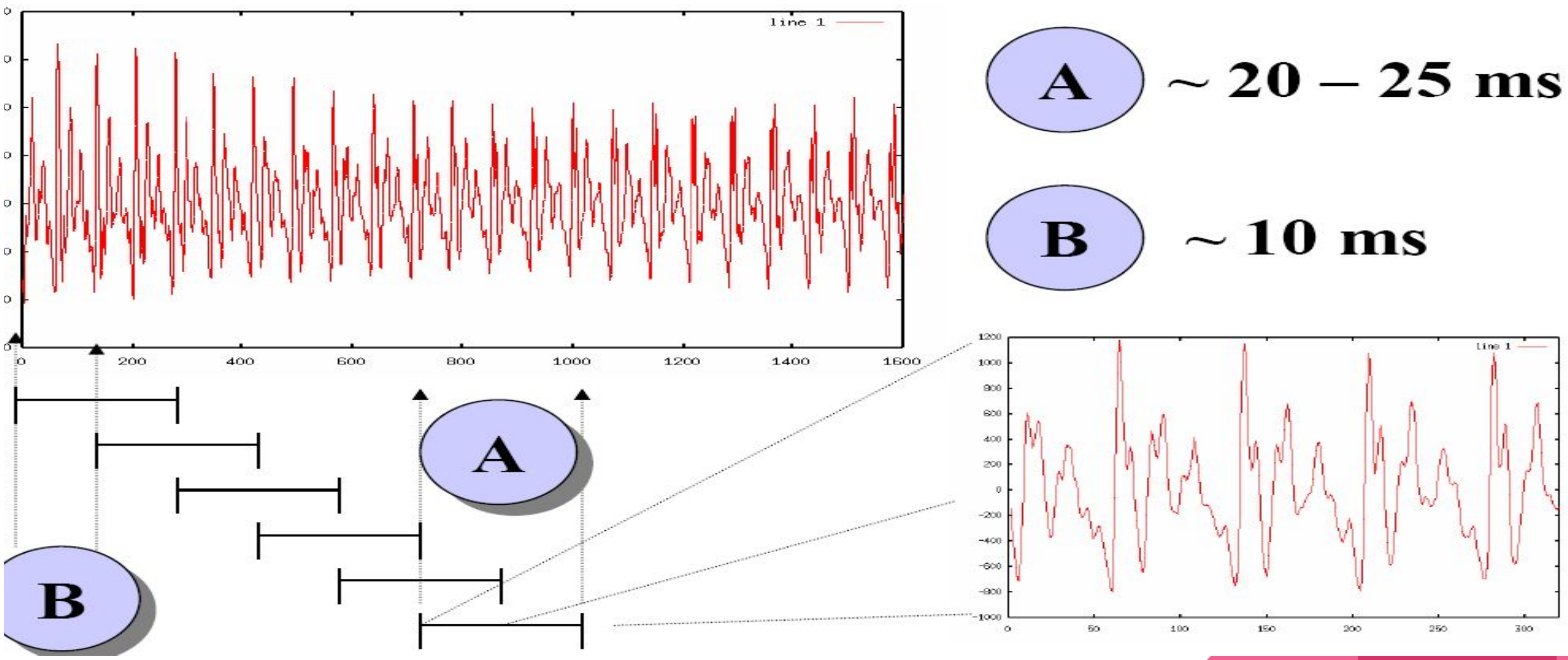
Preemphasis :

- The spectrum for voiced segments has more energy at lower frequencies than higher frequencies. This is called **spectral tilt**
- Spectral tilt is caused by the nature of the glottal pulse

Spectral slice from the vowel [aa]



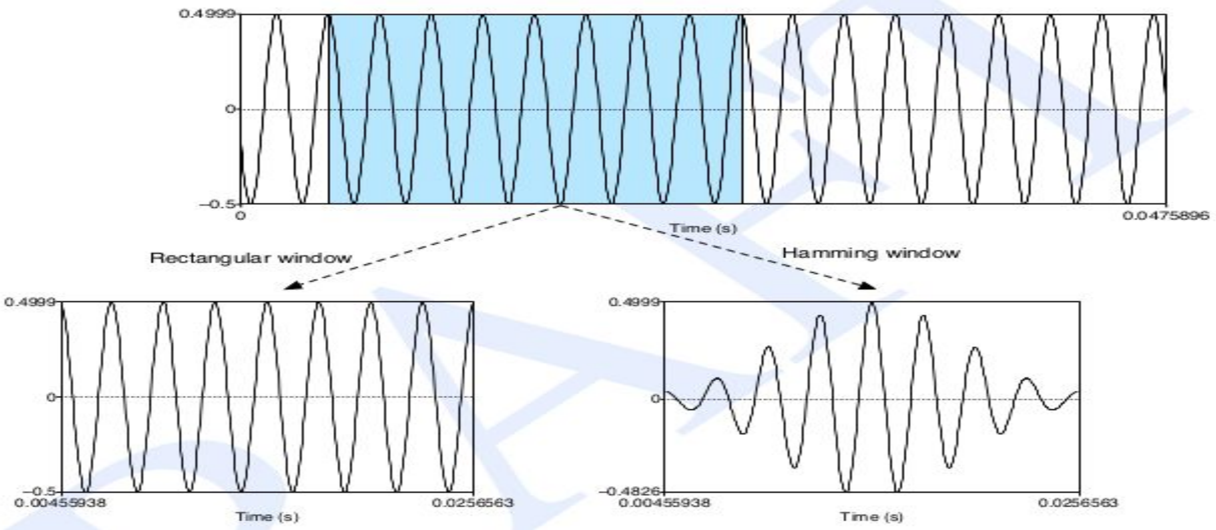
Windowing :



Common window shapes :

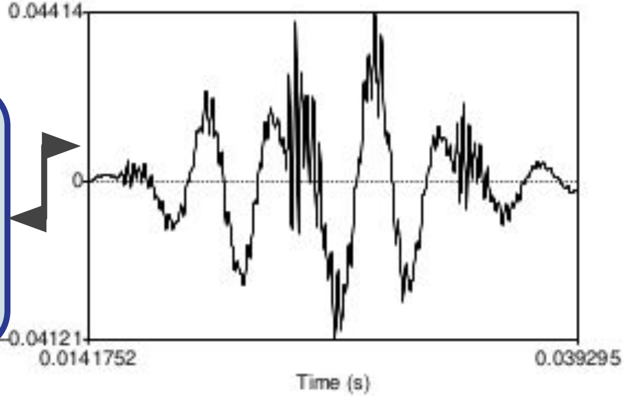
$$\text{rectangular} \qquad w[n] = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$

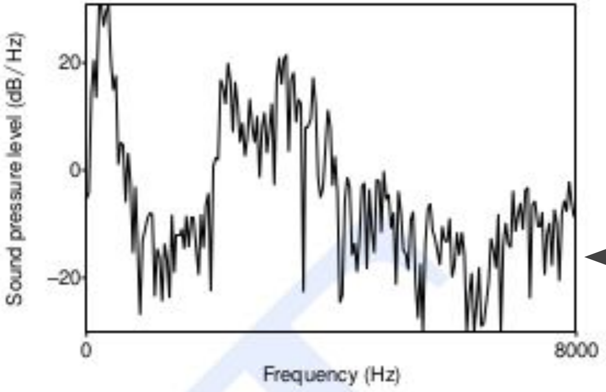
$$\text{hamming} \qquad w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right) & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$



DFT :

25 ms windowed portion





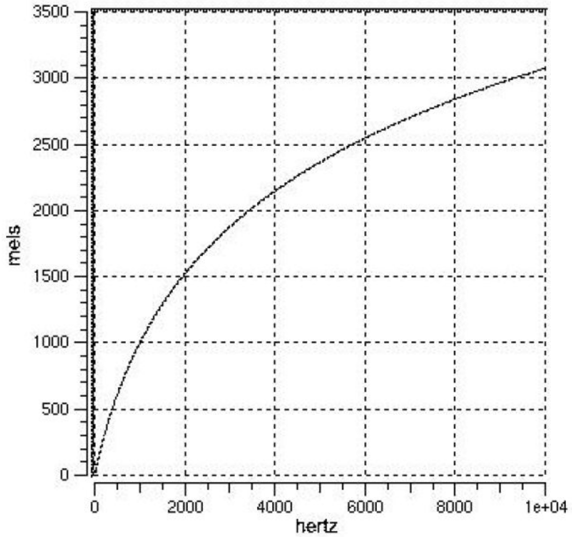
Spectrum compute d by DFT

Mel Scale :

- **Human hearing is not equally sensitive to all frequency bands**
- **Less sensitive at higher frequencies, roughly > 1000 Hz**
- **I.e. human perception of frequency is non-linear:**

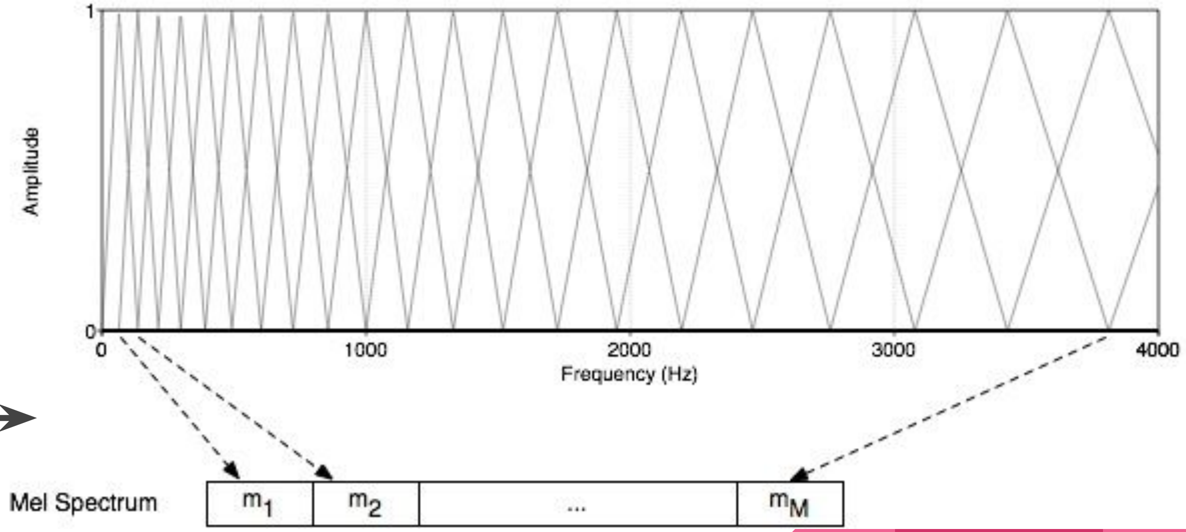
A mel is a unit of pitch
Definition: Pairs of sounds perceptually equidistant in pitch
Are separated by an equal number of mels

$$\mathrm{Mel}(f) = 1127 \ln\left(1+(f/100)\right)$$



Mel Filter bank
Uniformly spaced before 1 kHz
logarithmic scale after 1 kHz

Each triangular filter collects energy from given frequency range



The Cepstrum : IDFT

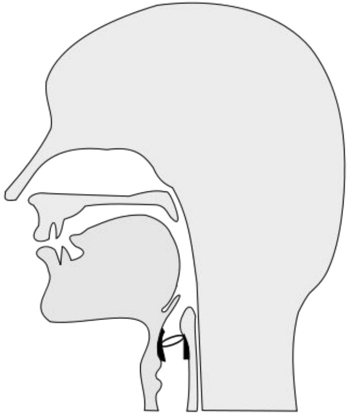
Speech = Convolution of vocal tract frequency response with glottal pulse

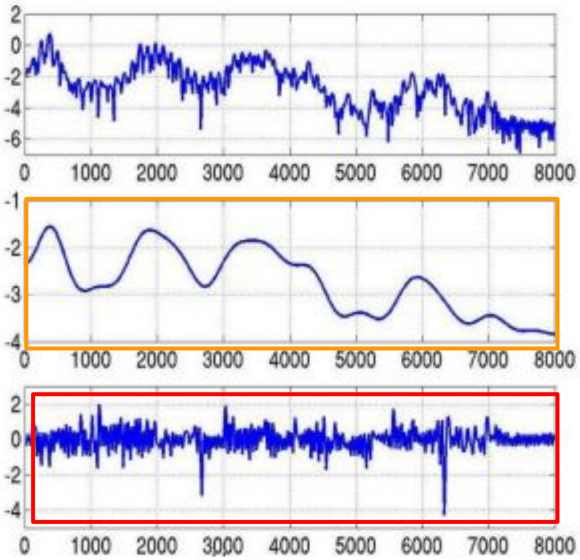
$x(t) = e(t) * h(t) \Rightarrow X(t) = E(t) * H(t) \Rightarrow \log(X(t)) = \log(E(t)) + \log(H(t))$

$\boxed{\text{Log (X(t))}} \equiv \boxed{\text{log(E(t))}} + \boxed{\text{log(H(t))}}$

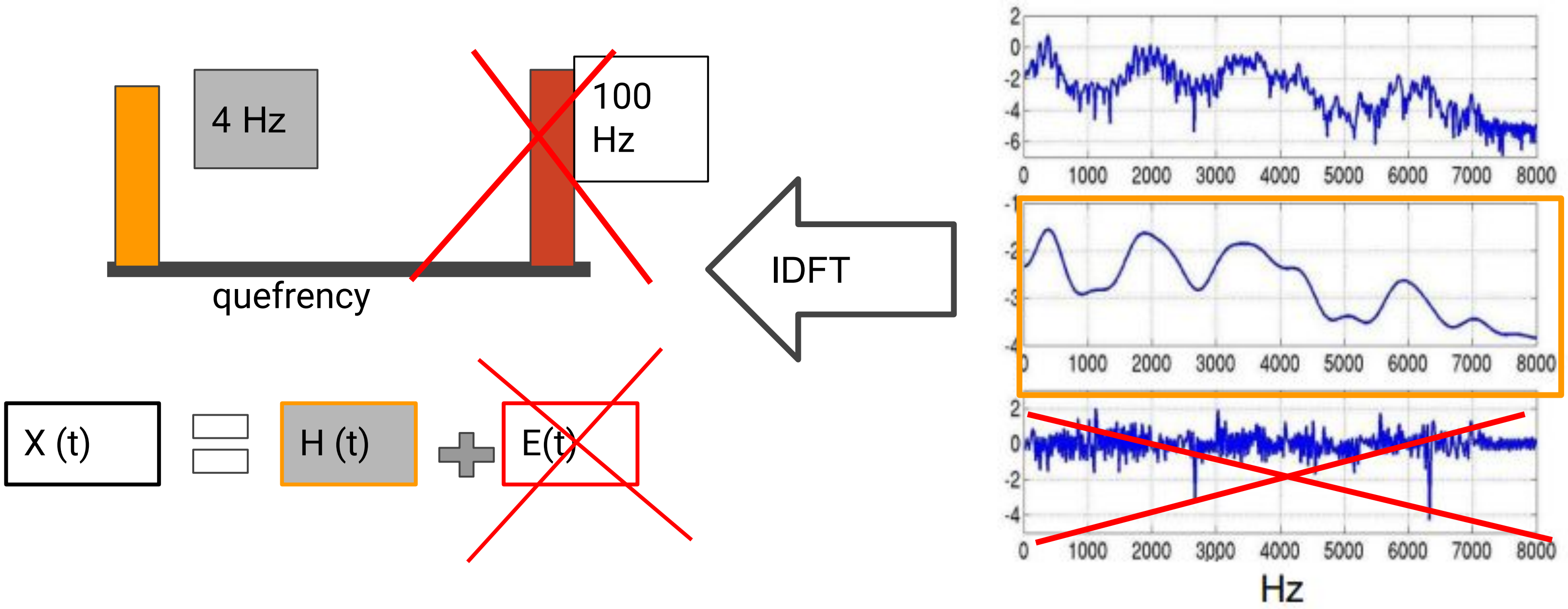
log(H(t)) → Vocal tract frequency response

log(E(t)) → Glottal pulse





The Cepstrum :



Feature :

- The cepstral coefficients do not capture energy
- So we add an energy feature
- Also, we know that speech signal is not constant
- So we want to add the changes in features (the slopes).
- We call these delta features

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

c(t) = cepstral value at time t

- We also add double-delta acceleration feature

12 MFCC
12 Δ MFCC
12 ΔΔ MFCC
1 ENERGY
1 Δ ENERGY
1 ΔΔ ENERGY

HMM Architecture

Our main goal :

“What is the most likely word out of all words in the language L given some acoustic input O?”

Input O = $o_1, o_2, \ldots, o_T$

L = { শুন্য ,এক,...,,নয় }

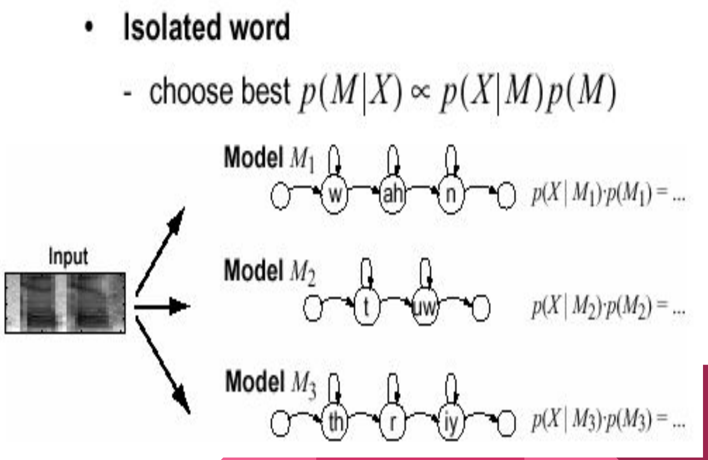
$\hat{W} = \operatorname{argmax} P(W|O) \quad W \in L$

$\hat{W} = \operatorname{argmax} P(O|W)\,P(W) \quad W \in L$

HMM

Acoustic Model

Language Model



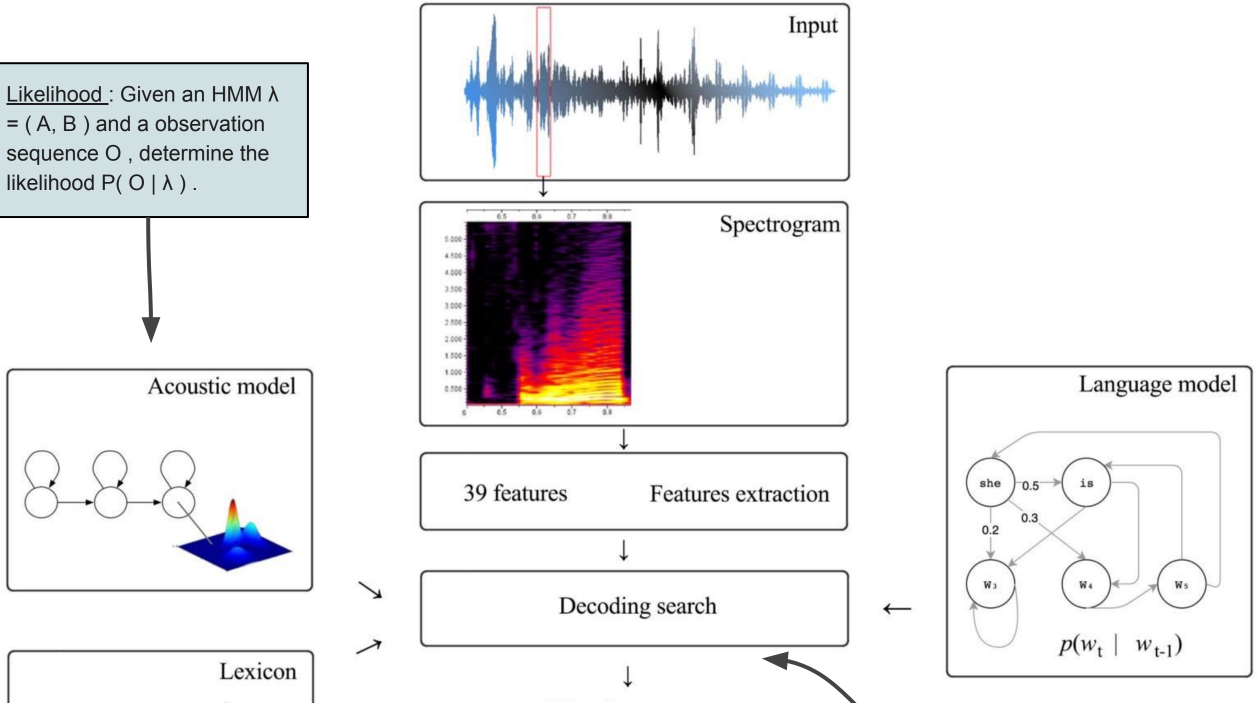
Overall Architecture :

Forward Algorithm

Likelihood : Given an HMM λ = (A, B) and a observation sequence O , determine the likelihood P(O | λ) .

Baum-Welch Algorithm

Learning : Given an observation sequence O and the set of states in the HMM , Learn HMM parameter A and B



Word sequence

$$W^* = \arg\max_{W} P(W \mid X)$$

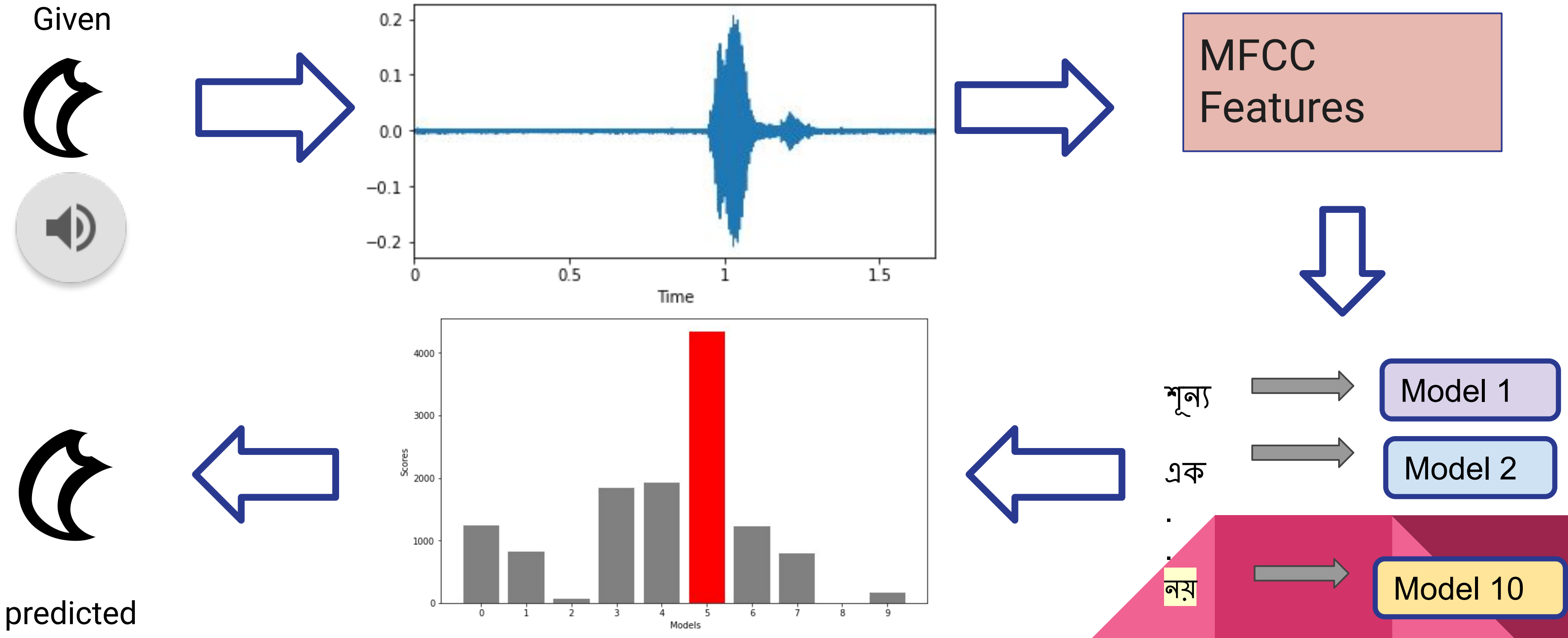
$$W^* = \arg\max_{W} p(X|W) \quad P(W)$$

word sequence acoustic model language model

Viterbi Algorithm:
Given an observation sequence O and an HMM λ = (A, B) , discover the best hidden sequence Q

Tutorial on HMM
Link :
https://docs.google.com/presentation/d/1JUBvTYGM1cepqNkQ39oM3zVpAzEWJni_dF51pz5_scg/edit?usp=sharing

Example :

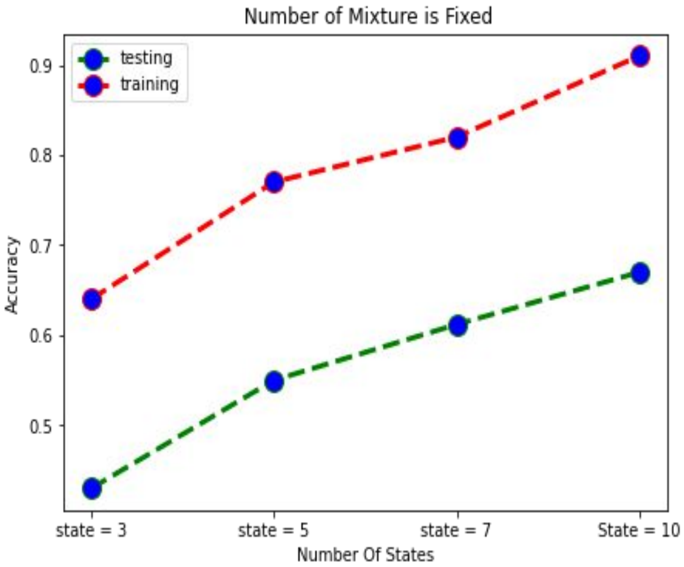


Given
৫
0.2
0.1
0.0
−0.1
−0.2
0
0.5
1
1.5
Time
MFCC
Features
শূন্য
এক
নয়
Model 1
Model 2
Model 10
4000
3000
2000
1000
0
Scores
0 1 2 3 4 5 6 7 8 9
Models
৫
predicted

Result Analysis

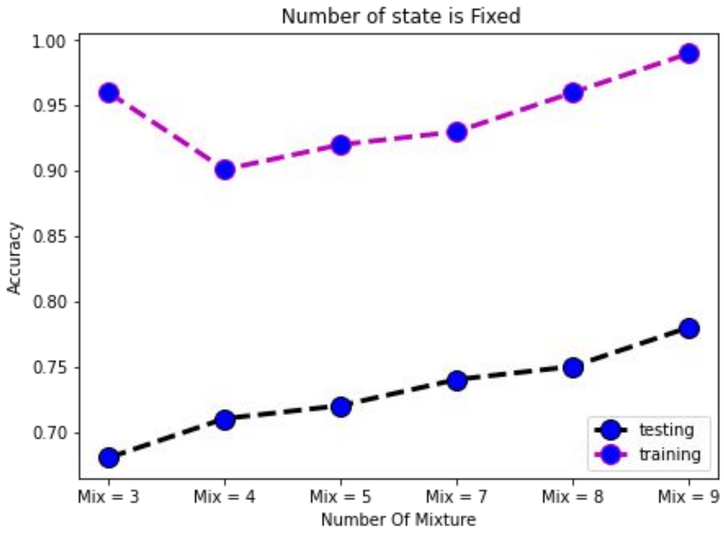
Total Data = 600
Train Data = 478
Test Data = 122

Number Of States	Number Of MIxture	Number Of Iteration	Training Accuracy	Testing Accuracy
3	2	100	64%	43%
5	2	100	77%	55%
7	2	100	82%	61%
10	2	100	91%	67%



Changing Parameters :

Number Of States	Number Of MIxture	Number Of Iteration	Training Accuracy	Testing Accuracy
10	3	100	96%	68%
10	4	100	90%	71%
10	8	100	96%	74%
10	9	400	96%	76%

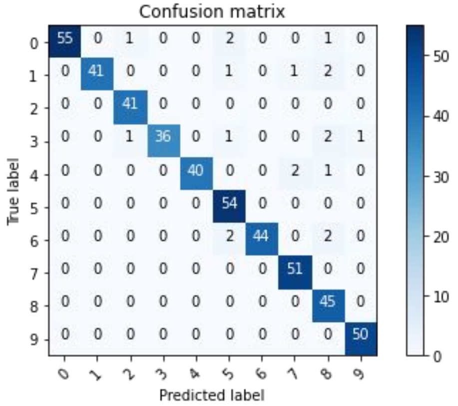


Data = 478

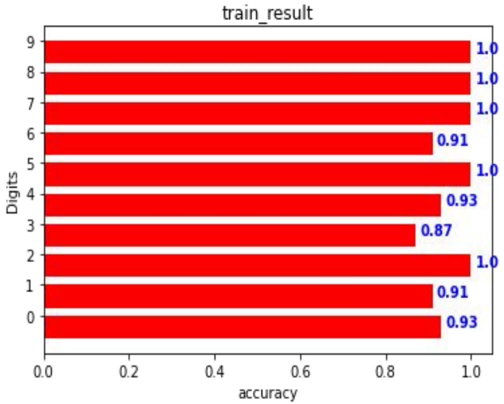
Training Analysis

Accuracy = 96%

Confusion Matrix for training set



Digit wise accuracy for training set

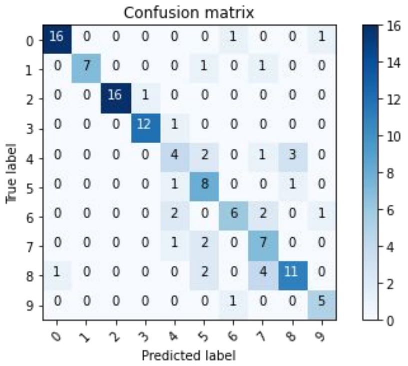


Data = 122

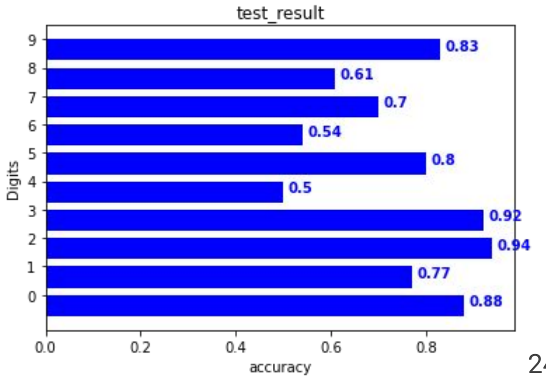
Testing Analysis

Accuracy = 76%

Confusion Matrix for test set



Digit wise accuracy for test set



Any Questions

?


Thank
You