SPEAKER RECOGNITION SYSTEM

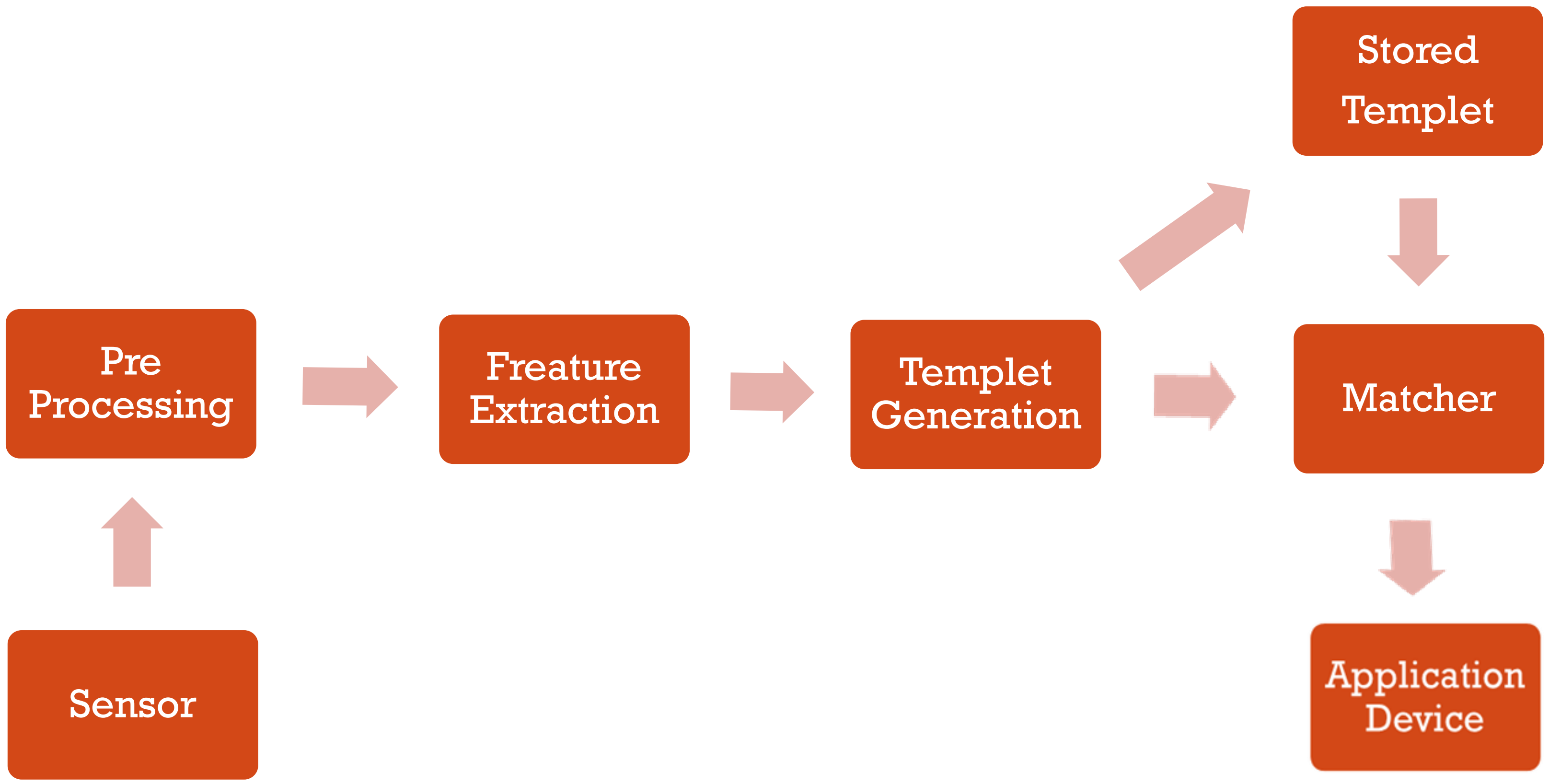
- Akshay Gupta
- Avinash Bansal

Department of Computer Science and Technology, Z.H.C.E.T,
Aligarh Muslim University

SPEAKER RECOGNITION SYSTEM

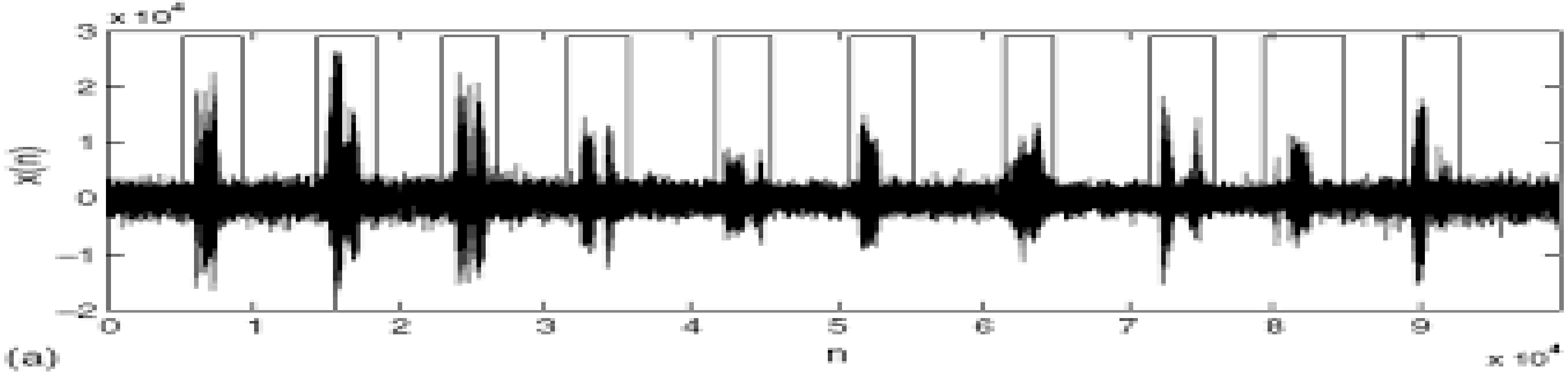
- Identification of person who is speaking by characteristics of their voice biometrics.
- Recognition of speaker is divided into two parts identification and verification.

BASIC BLOCK DIAGRAM



VOICE ACTIVITY DETECTION

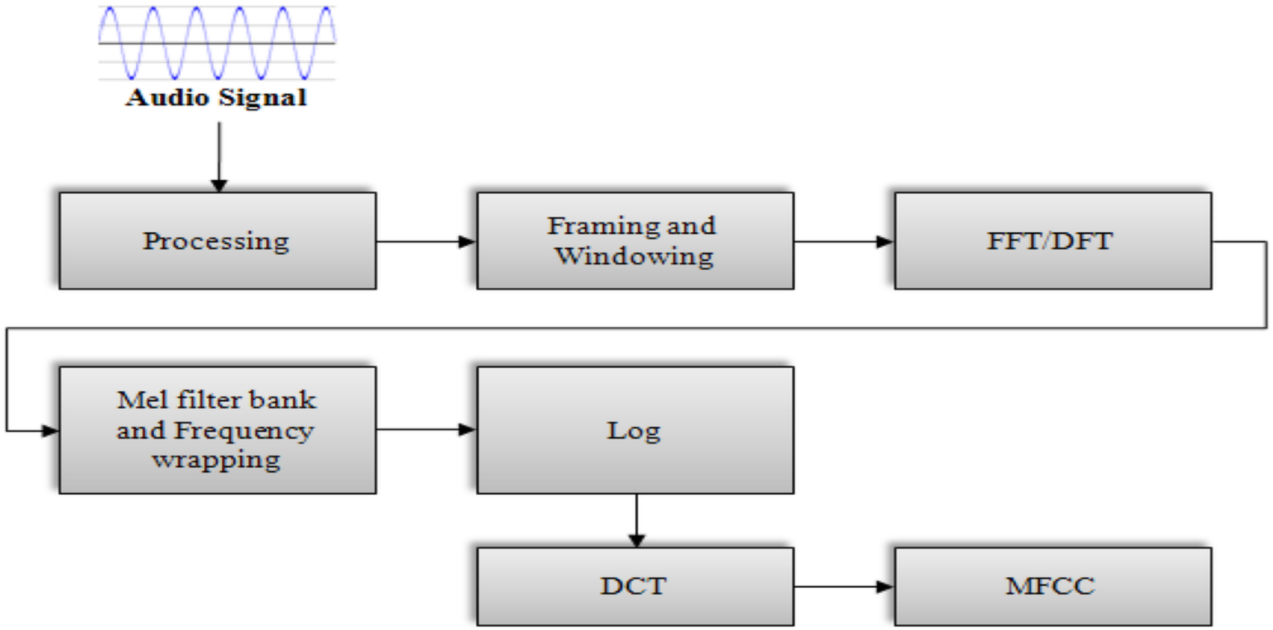
- Energy Based
  - Filter out the interval with relatively low energy
  - Work perfectly for high quality recording
  - Sensitive to noise



(a)

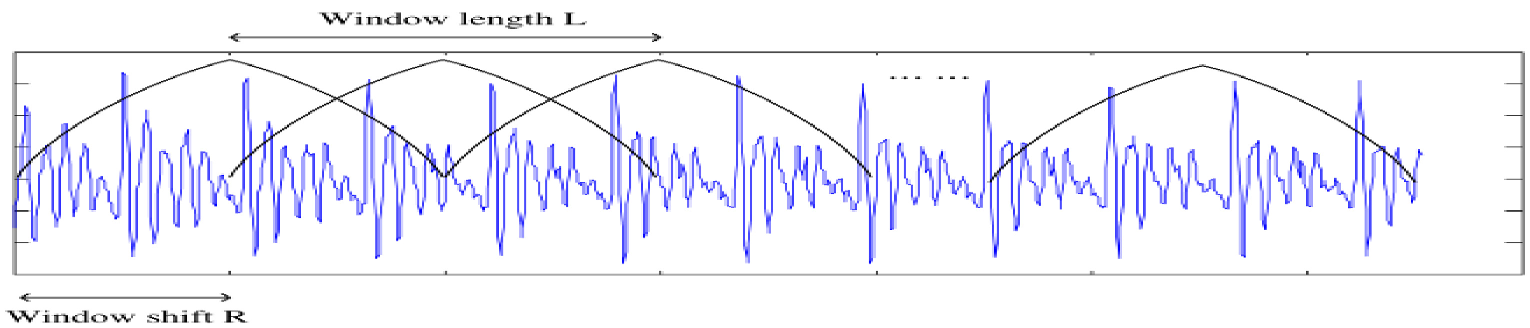
ME-FREQUENCY CEPSTRAL COEFFICIENTS

- Cepstral feature closely approximate human auditory system. Commonly used feature for speech/speaker recognition.
- A pure mathematical model.



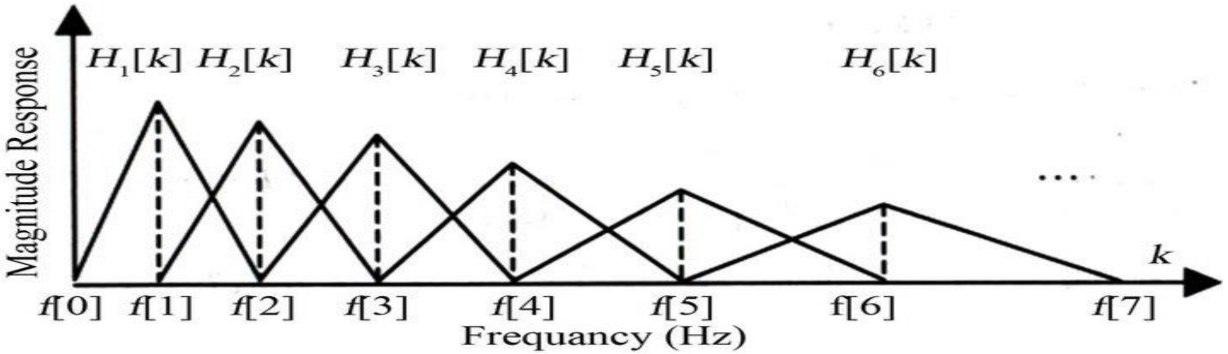
- **Windowing:**

Windowing is the process of taking a small subset of a larger dataset, for processing and analysis



- **Mel scale and Filter bank:**

$$Mel(f) = 2595 \log 10(1 + f/700)$$

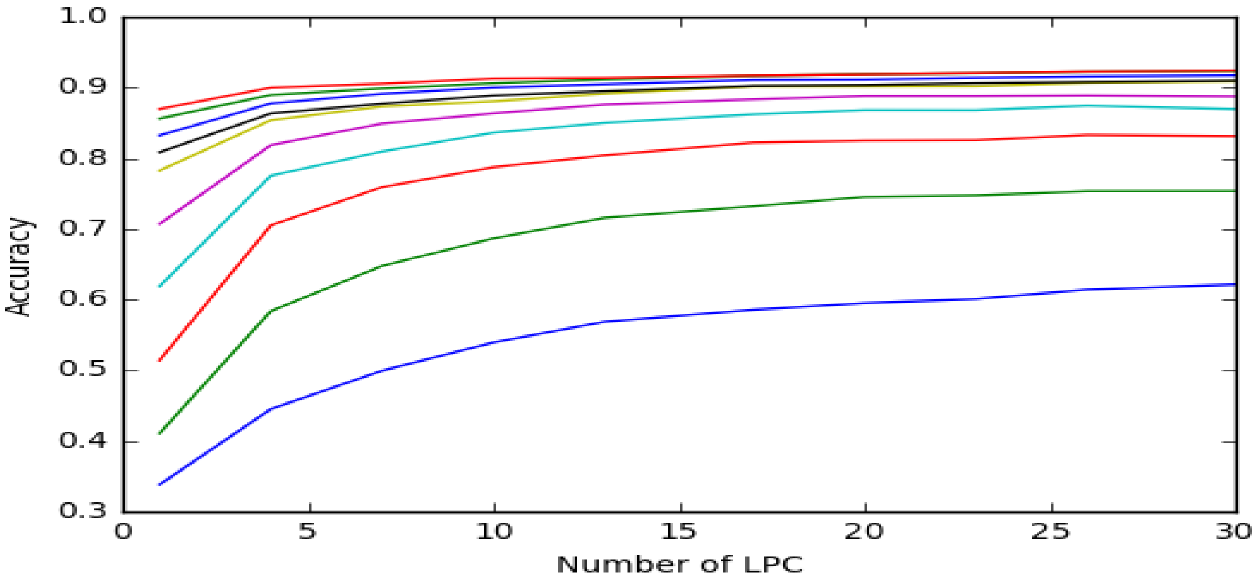


Linear Predictive Coding/Coefficients:

- Assumption

  In a short period, the nth signal is a linear combination of previous p signals: $x'(n) = \sum_{i=0}^{p} a(i)x(n-i)$
- Minimize squared error E [$x'(n)$-$x(n)$] using Levinson-Durbin algorithm.
- Use $a1, a2\ldots.ap$ as features

Experiment with number of features:

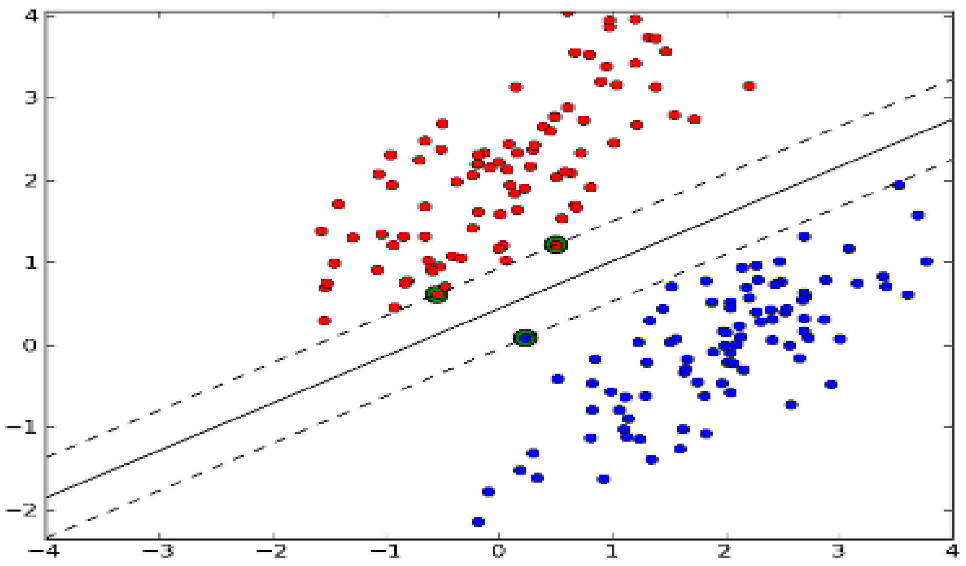


SUPPORT VECTOR MACHINE

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.



Mathematical formulation

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

If $y = 1$, we want $\theta^T x \geq 1$ (not just ≥ 0)

If $y = 0$, we want $\theta^T x \leq -1$ (not just < 0)

CONCLUSION:

- We managed to get voice sample of 36 speakers and from that we get around 80% training and test set accuracy
- Our model has good accuracy but upto now it is not considering noise.
- Our learning model is not over fitted.(i.e. training accuracy=test accuracy)

REFERENCES

- **Book:** Fundamentals of Speaker Recognition by Homayoon Beigi
- **Wikipedia:** MFCC
- **Scikit learn Machine Learning in Python:**

  www.scikit-learn.org/stable/index.html

Thanks !