# Mini Project 1

Team 08 : EE17B047(Kommineni Aditya) - EE17B035(V Sai Krishna)

April 2020

## 1 Problem Statement

The objective is to build a **Vanilla Gaussian Mixture Model** system to do the task of **text independent speaker identification.** The dataset contains 200 speakers and is taken from the TIMIT/NTIMIT databases.

## 2 Feature Extraction from data

Before training the GMMs, MFCC features were extracted from both the TIMID and NTIMID data. Features were extracted at 100 frames per second. Both **MFC and Delta Coefficients** were extracted from the data. To reduce the influence of noise Voice Activity Detection (VAD) was used. The threshold for VAD was set to 0.06. The total number of coefficients generated were 38 (19 MFC and 19 Delta Coefficients).

## 3 GMM Training

Now, for each speaker a GMM Model will be built. We used **K means** as our **Vector Quantization (VQ)** approach for initializing the means, variance and weights of each Gaussian. Since, we are building a separate GMM for each speaker, we have to do K means initialization for each speaker data individually.

The initial mean values for our K means algorithm is chosen by sampling the dataset uniformly between the minimum and the maximum value along every direction. We have tried a range of values for the number of clusters (which is in turn equal to the number of Gaussians per GMM). The equations used are as follows:

**Cluster Assignment Step**

$$S_i^{(t)} = \{x_p : ||x_p - \mu_i^{(t)}||^2 \leq ||x_p - \mu_j^{(t)}||^2 \ \forall j, 1 \leq j \leq k\}$$

where S = {$S_1$,$S_2$,...$S_k$} are the Cluster identities and $\mu_i$ is the mean of points in $S_i$.

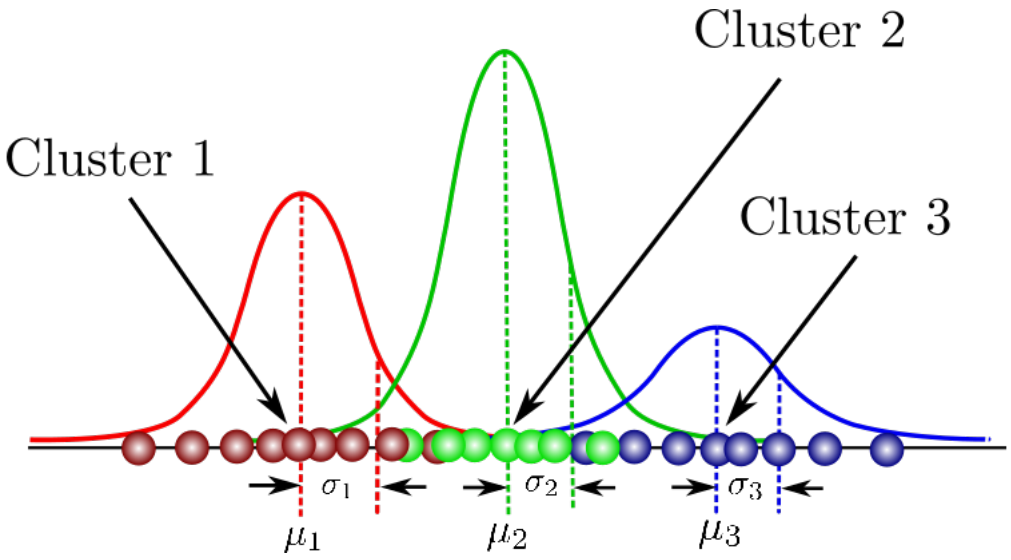


Figure 1: One dimensional model of Gaussian Mixture Model clustering

**Centroids Update Step**

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \epsilon S_i^{(t)}} x_j$$

After the convergence of our K-means algorithm, the mean, variance and fraction of points assigned to each cluster is calculated. These are assigned as the initial mean, variance and weights for the Gaussians for our GMM. Then our GMM's are trained using the **Expectation-Maximization algorithm**. The E and M steps are stated below:

**Expectation Step**

$$\gamma_j^{(i)} = P(z^{(i)} = S_j |\ x_i; \phi, \mu, \Sigma)$$

**Maximization Step**

$$\phi_j = \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x_i}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

# 4 Testing and Observations

While testing, we will first find the conditional probability of the MFCC data given the GMM parameters of a speaker. This is done for every GMM's parameters and whichever speaker's model gives the **highest conditional probability** is chosen as our prediction. However, we have taken **Top 3 and Top 5** predictions for calculating accuracies too.

As mentioned before, we varied the number of Gaussians used per GMM. The following table depicts the results obtained:

| **No of Gaussians per GMM** | **Top 1 Accuracy** | **Top 3 Accuracy** | **Top 5 Accuracy** |
|---|---|---|---|
| 3 | 9% | 16% | 26.5% |
| 5 | 21% | 28.5% | 36.5% |
| 10 | 40% | 59.5% | 69% |
| 15 | 53.5% | 68% | 76.5% |
| 20 | 64.5% | 75% | 81.5% |

Table 1: Accuracy obtained vs No of Gaussians in our GMM

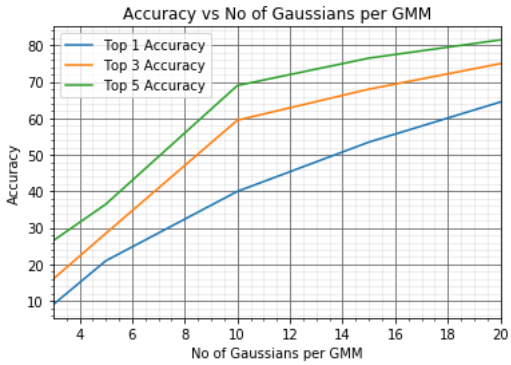


Figure 2: Accuracy vs No of Gaussians per GMM

From the table and graph, we can observe that our model clearly does a better job than random guessing.

# 5 Conclusions

- We could observe that as we increase the number of gaussians in our model, we get better results. This is exactly what we expected as, using more gaussians would enable us to represent the data better.
- Our approach requires us to build a separate GMM from the scratch for each speaker. This is a huge waste of time and computing resources. This is why a **UBM-GMM approach is preferred**, as a brand new GMM for every speaker will not be built from scratch.
- Increasing the number of Gaussians is necessary for better results. However, it increases both training and testing duration substantially. This is why in practice **Speaker verification** is used over **Speaker Identification**.

## Note

The accuracy obtained is significantly higher when the number of speakers were less. For example, we obtained an accuracy of 100% when we trained and tested on only 25 speakers.

This tells us that the accuracy can be improved if the amount of training data per speaker is increased.

# 6 References


- https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f
- https://en.wikipedia.org/wiki/K-means_clustering
- The link to the code: https://drive.google.com/file/d/1jfPoQHcLSjv1tm6FwaCVEBJtAlVJQ8aA/view?usp=sharing