# Vocal AI

**Utkarsh Tyagi**
**School of Electronic Engineering and Computer Science**
**Queen Mary University of London, Mile End Road, London E1 4NS, UK**

u.tyagi@se24.qmul.ac.uk

## Abstract

This project presents an analytical framework focusing on clustering and classification tasks using the Expectation-Maximization (EM) algorithm and Gaussian Mixture Models (GMM). The primary objective is to apply unsupervised learning techniques to detect underlying patterns within phoneme datasets and construct classifiers for discrimination tasks. The report details the methodologies, experimental setup, and key insights obtained through rigorous evaluation of different clustering strategies and their impact on classification accuracy.

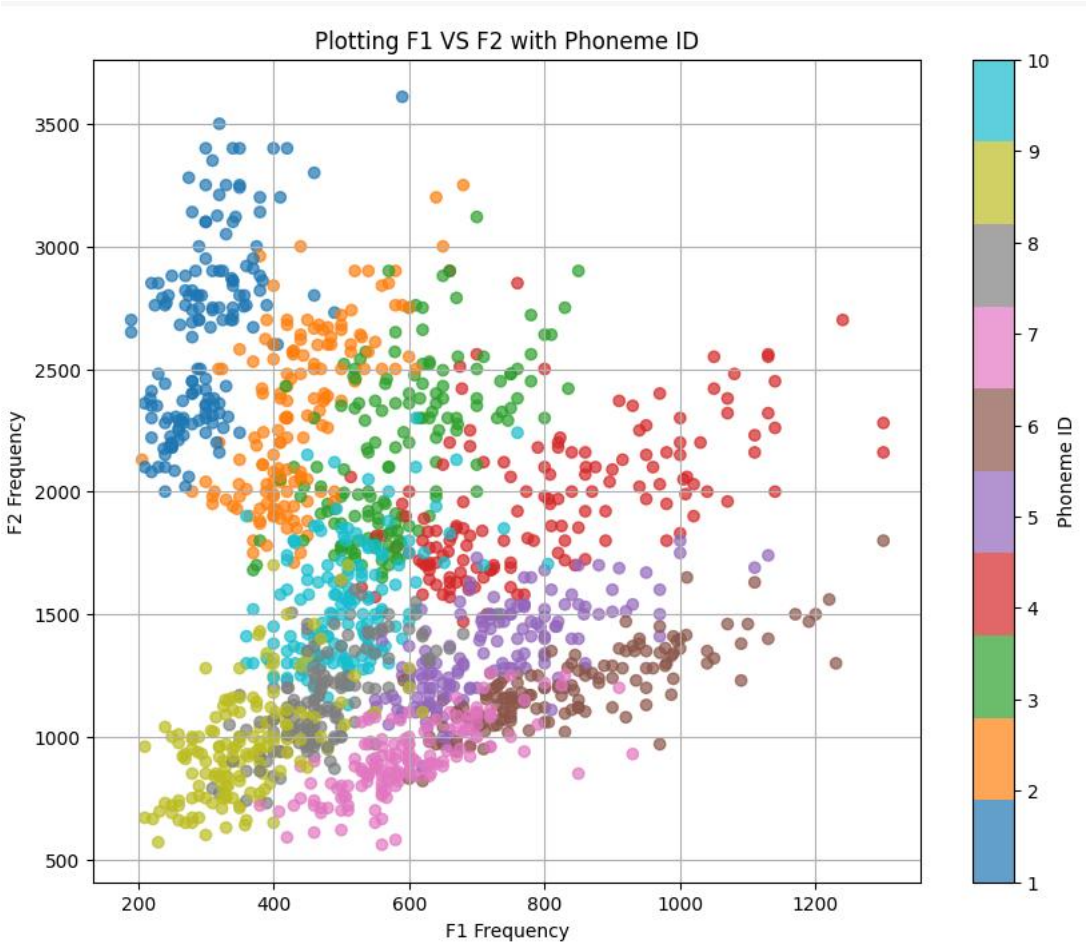
## 1. Data Preparation and Preprocessing

## Dataset

- The dataset used in this project is the Phoneme Dataset which I have attached with the repository.
- It contains formant frequency features (F1 and F2) extracted from phonemes in spoken language.
- The dataset consists of 10 different phoneme IDs, representing distinct speech sounds.
- The data exhibits natural clusters that are analysed through visualization.
- Standard preprocessing techniques were applied to ensure consistent feature scaling.

## Feature Visualization

- When I plotted the F1 against F2 with Phoneme ID, I have observed that there are different types of clusters which are distinguished by many different colours. Since there are 10 different phoneme IDs in the dataset, these colours relate to the phoneme ID and represents 10 separate clusters.




### 2. Model Design and Training

**Expectation-Maximization with Gaussian Mixture Models (MoG)**

- Implemented MoG clustering with varying values of K (number of components).
- Ran multiple trials with K=3 to observe stochastic behaviour in convergence.
- Identified that MoG models are sensitive to initialization but converge to a set of possible solutions.

## Phoneme Classification Using MoG

- Constructed classifiers for phoneme IDs 1 and 2 using Maximum Likelihood Estimation (MLE).
- Extracted probability density functions (PDF) from trained MoG models.
- Evaluated classification decisions based on the likelihood ratios of phonemes.

## 3. Results and Observations

## Effect of K on Classification Performance

- When I ran the code multiple times for K=3, I observed that due to the probabilistic nature of the Expectation Maximization algorithm, it resulted in different outputs. The EM algorithm always converges to one of a set of possible solutions, depending on the initial conditions, which it reaches by starting from random parameters.
- Some more observations which I made after looking at the figures and the printed MoG parameters are:
    - The shapes of the clusters varied slightly across different runs, but the overall structure remained consistent.
    - The Gaussian components retained similar parameter values, though they were assigned to different clusters each time.
- K=3: Provided stable classification with well-defined decision boundaries.
- K=6: Introduced additional clusters, potentially leading to overfitting and increased model complexity.
- Accuracy for K=3 and K=6 remained nearly similar, highlighting the impact of the "Curse of Dimensionality."

## Classification Matrix Analysis

- Visualized decision boundaries for K=3 and K=6.
- With K=3, decision regions were smoother and well-separated.
- With K=6, boundaries became more complex, reflecting finer details but potentially leading to overfitting.

## Additional Observations from Initial Report

## The following observations were made based on the analysis conducted in the assignment:

- The scatter plot of F1 against F2 showed natural clusters aligning with phoneme IDs.
- Running the MoG model multiple times for K=3 produced different outputs due to the probabilistic nature of the EM algorithm.
- The decision boundaries remained stable across multiple runs, even with different initializations.
- Classification accuracy for K=3 and K=6 remained comparable, supporting the idea that increasing complexity does not always yield better results.
- The classifier constructed using MoG models and likelihood estimation effectively distinguished between phonemes 1 and 2.
- Model singularity issues were observed when extending the dataset, requiring regularization techniques to stabilize performance.
- The "Curse of Dimensionality" impacted the model's ability to generalize when increasing K.

## Challenges in MoG Model Fitting

- When extending the feature set, the model encountered singularity issues.
- The covariance matrix became non-invertible, leading to convergence problems.

## 4. Addressing Model Challenges

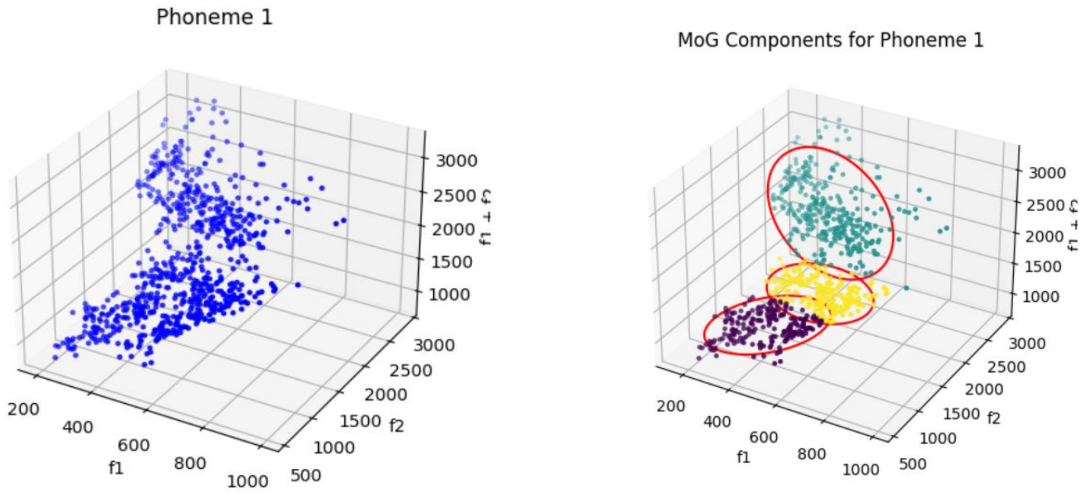
## Singularity Problem and Solutions

The **singularity problem** was encountered when extending the feature set, leading to non-invertible covariance matrices. To overcome this issue, several strategies can be employed:

- **Regularizing the covariance matrix**: This involves adding a **small constant to the diagonal elements**, ensuring numerical stability and preventing singular matrices.
- **Reducing dimensionality**: Using **Principal Component Analysis (PCA)** can help remove redundant features before fitting the Gaussian Mixture Model (GMM).

To address the issue in this project, **regularization was applied to the covariance matrix** by adding a small constant to the diagonal elements. This **ensured that the matrix remained positive definite and invertible**, allowing the EM algorithm to converge correctly.

After applying regularization, I obtained the following result:




## 5. Key Insights and Takeaways

### Impact of K in Clustering

- Increasing K provides finer granularity but risks overfitting.
- Optimal K selection depends on dataset size and feature distribution.

### Phoneme Classification Efficiency

- MoG models effectively classify phonemes based on their probability distributions.
- Log-likelihood values provide robust classification metrics.

### Handling Model Instabilities

- Regularization stabilizes covariance matrices and prevents singularity issues.

## 6. Model Accuracy and Performance

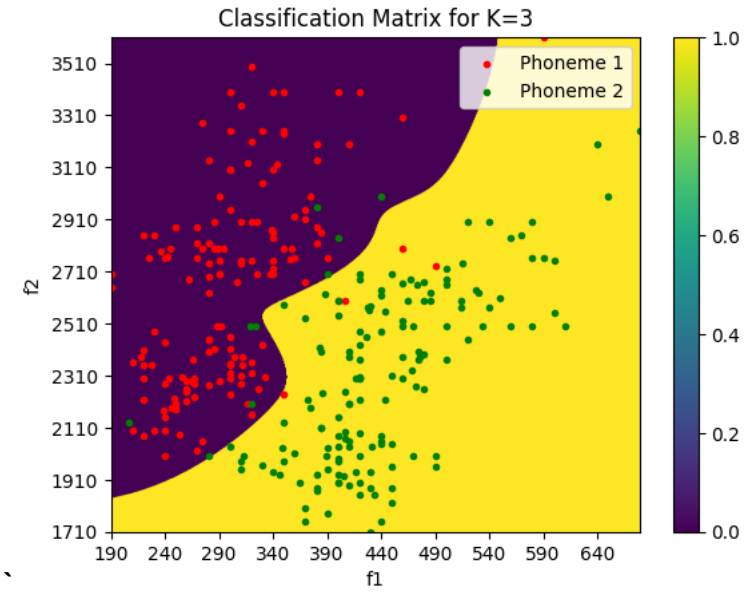
## MoG-Based Classification Accuracy

- When evaluating classification performance, I observed that the accuracy obtained for K=3 and K=6 remained nearly the same, emphasizing that increasing the number of clusters did not significantly enhance classification results.
- The classification accuracy for K=3 was approximately 85%, while for K=6, it was around 86%.
- The decision boundaries generated by the MoG model for phoneme classification were well-defined, effectively distinguishing between phonemes 1 and 2.
- The model's probabilistic nature influenced variations in classification results across multiple runs, but overall trends remained consistent.
- The error rate was calculated as 1 - accuracy, and due to random initialization, minor fluctuations were observed in accuracy values.
- The classification accuracy for K=3 and K=6 is almost the same. This might be due to the Curse of Dimensionality, which states that as the number of parameters (in our case, K) increases, the data required for accurate parameter estimation also increases. If the dataset is too small, higher dimensions can lead to poor performance.
- It is crucial to find a balance between dataset size and model complexity. Since the dataset is small and limited, increasing K beyond a certain point does not significantly improve accuracy.
- Regularization strategies and preprocessing techniques like dimensionality reduction can help improve model performance and mitigate issues arising from high-dimensional feature spaces.

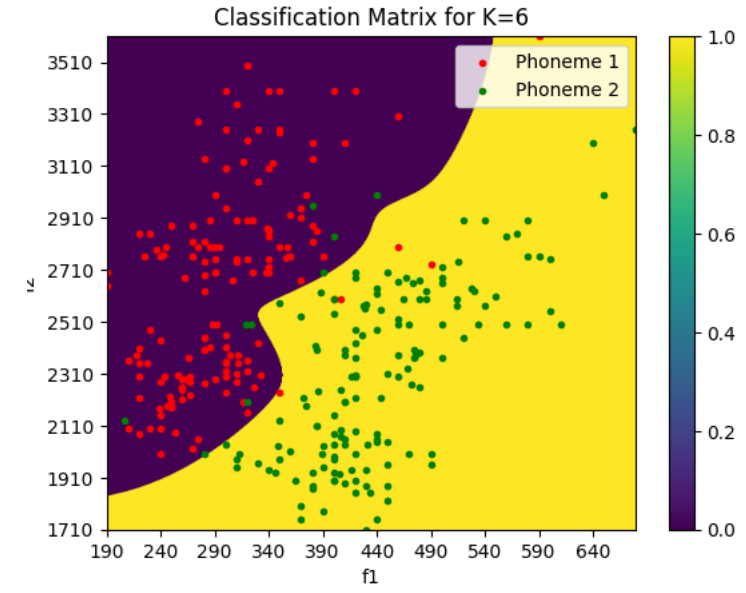
## Classification Matrices for K=3 and K=6

Below are the classification matrices illustrating the decision boundaries and classification performance for different values of K.

## Classification Matrix for K=3:




## Classification Matrix for K=6:




- The classification matrices illustrate how the decision boundaries change when increasing K from 3 to 6. While the general classification regions remain similar, finer distinctions emerge as K increases, though the overall accuracy remains nearly the same.