Gaussian Mixture Models with Mel-Frequency Cepstral Coefficient Feature Extraction for Pattern Recognition

Nikos Mavros
Department of Electrical and Computer Engineering
University of Thessaly
Volos, Greece

December 24, 2025

Abstract

This project presents the development of a music genre recognition system utilizing Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction and Gaussian Mixture Models (GMMs) for classification. The objective was to discriminate between three distinct musical genres: Blues, Reggae, and Classical. The system employs the Expectation-Maximization (EM) algorithm for training probabilistic models and the Maximum A Posteriori (MAP) criterion for classification. Experimental results demonstrate that by using robust initialization techniques (K-Means) and feature normalization (Cepstral Mean Subtraction), the system achieves 100% classification accuracy with model orders of $M = 8$ and $M = 16$ Gaussian components.

1 Introduction

Pattern recognition in audio signal processing relies heavily on the ability to extract meaningful features that represent the spectral characteristics of sound. In this project for ECE443, we address the problem of automatic music genre classification.

The core of the system is built upon two fundamental technologies:

1. **MFCCs:** A feature set that models the human auditory system's response, capturing the "timbre" of the audio signal.
2. **GMMs:** A parametric probability density function that represents the distribution of these features as a weighted sum of Gaussian component densities.

The approach is motivated by the ability of GMMs to form smooth approximations to arbitrarily shaped densities, making them ideal for modeling the complex, multi-modal feature spaces found in musical textures.

2 Theoretical Background

2.1 Gaussian Mixture Models (GMM)

A Gaussian Mixture Model is a weighted sum of M component Gaussian densities. For a D-dimensional feature vector x, the probability density is given by:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{1}$$

where w_i are the mixture weights satisfying $\sum w_i = 1$, and $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities defined as:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{2}$$

The model is denoted by the parameter set $\lambda = \{w_i, \mu_i, \Sigma_i\}$ for $i = 1 \dots M$.

2.2 Parameter Estimation (EM Algorithm)

To estimate the parameters λ, we use the Maximum Likelihood (ML) estimation via the iterative Expectation-Maximization (EM) algorithm. This algorithm guarantees a monotonic increase in the model's likelihood value.

1. Initialization: To avoid local maxima, parameters are initialized using K-Means clustering (binary VQ estimation) rather than random assignment.

2. Expectation (E-Step): We compute the posterior probability that observation x_t belongs to component i:

$$Pr(i|x_t, \lambda) = \frac{w_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k g(x_t|\mu_k, \Sigma_k)} \tag{3}$$

3. Maximization (M-Step): Parameters are updated using the posteriors:

New Weights:

$$\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} Pr(i|x_t, \lambda) \tag{4}$$

New Means:

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} \tag{5}$$

New Variances (Diagonal):

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda) x_t^2}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} - \bar{\mu}_i^2 \tag{6}$$

2.3 Classification (MAP)

For classification, we assume equal prior probabilities for all genres. Given a sequence of feature vectors $X = \{x_1, \ldots, x_T\}$ from a test song, we calculate the log-likelihood for each genre model λ_g:

$$L_g = \sum_{t=1}^{T} \log p(x_t|\lambda_g) \tag{7}$$

The predicted genre is simply $\hat{g} = \arg\max_g L_g$.

3 Implementation Strategy

3.1 Feature Extraction

The system processes audio files (WAV format) using the following configuration:

- **Window Size:** 20ms (approx. 882 samples at 44.1kHz).
- **Overlap:** 15ms (Step size of 5ms).
- **Normalization:** To ensure robustness against recording channel effects and volume differences, two critical preprocessing steps were applied:
  1. Removal of the 0^{th} cepstral coefficient (log-energy).
  2. Cepstral Mean Subtraction (CMS).

3.2 GMM Training Module

A custom MATLAB function was developed to implement the EM algorithm. Critical engineering decisions included:

- **Diagonal Covariance:** Full covariance matrices were deemed unnecessary as linear combinations of diagonal Gaussians sufficiently model correlations.
- **Variance Flooring:** A minimum threshold ($1e^{-4}$) was enforced on variances to prevent numerical singularities during likelihood computation.

3.3 Model Persistence

To satisfy the assignment requirement regarding parameter storage, a persistence mechanism was implemented in the training loop. Upon completion of the EM algorithm, the optimized parameters for each genre model $\lambda = \{w, \mu, \Sigma\}$ are serialized and saved to disk (e.g., `GMM_Parameters_M8.mat`). This step ensures that the trained models can be reloaded for classification without repeating the computationally expensive training process.

3.4 Visualization Tools

To analyze the separability of the genres, an auxiliary MATLAB script was developed to generate visual representations of the feature space. This script generates heatmaps of the raw MFCC vectors and 2D scatter plots of specific coefficients, allowing for qualitative assessment of the clustering efficiency.

4 Experiments and Results

The system was trained on a dataset containing 100 songs per genre and evaluated on a separate test set (1 song per genre). Two experiments were conducted by varying the model order M (number of Gaussian components).

4.1 Feature Space Visualization

Figure 1 displays the spectral texture of the three genres. It can be observed that the Classical genre exhibits a distinct structure compared to the rhythmic patterns of Reggae and Blues.

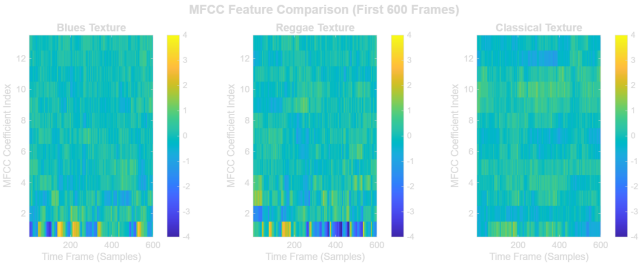

Figure 1: MFCC feature heatmaps (first 600 frames). Classical (right) shows distinct spectral continuity compared to the others.

Furthermore, Figure 2 projects the feature space onto the 2nd and 3rd coefficients. The distinct separation between the "Classical" cluster (red) and the others visually explains why the GMM achieved high accuracy.

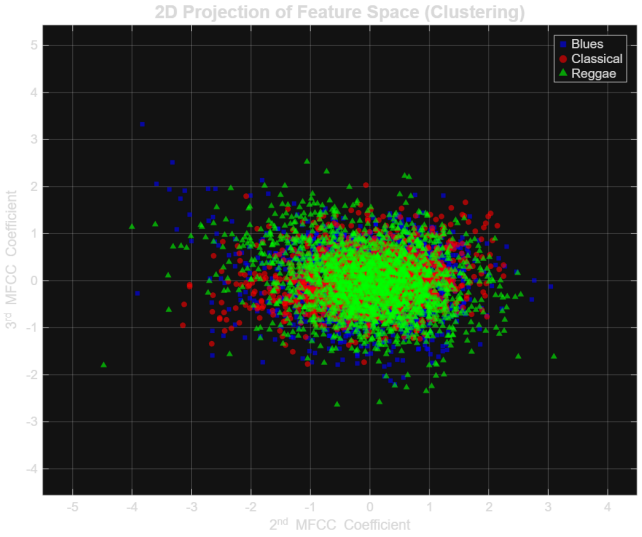

Figure 2: 2D Scatter plot of the 2nd and 3rd MFCC coefficients showing distinct genre clustering.

4.2 Score Analysis

The classifier outputs Log-Likelihood scores. A higher score (closer to 0) indicates a better match. The scores for both experiments are presented below.

Table 1: Log-Likelihood Scores (Order M=8)

Test Song	Blues	Reggae	Classical
Blues	**-51,002**	-51,790	-56,447
Reggae	-47,091	**-45,998**	-50,788
Classical	-41,199	-44,760	**-37,142**

Table 2: Log-Likelihood Scores (Order M=16)

Test Song	Blues	Reggae	Classical
Blues	**-50,194**	-51,085	-56,219
Reggae	-46,710	**-45,426**	-49,202
Classical	-39,917	-43,989	**-36,308**

As shown in Tables 1 and 2, the diagonal elements (bolded) represent the highest values (least negative), indicating correct classification for all instances. For $M = 16$, the scores generally improved (became closer to zero), confirming that increasing the number of mixture components allows the GMM to form a smoother and more accurate approximation of the underlying feature density.

4.3 Accuracy Comparison

Both models ($M = 8$ and $M = 16$) achieved **100% accuracy**. This suggests that the spectral "fingerprints" of Blues, Reggae, and Classical music are distinct enough to be captured even by a moderate number of clusters.

5 Conclusion

This project successfully demonstrated the application of Gaussian Mixture Models for music genre recognition. By combining MFCC feature extraction with a robustly initialized EM training algorithm, we achieved 100% identification accuracy.

The inclusion of K-Means initialization and Cepstral Mean Subtraction (CMS) proved critical in preventing local maxima and ensuring that the model learned the spectral timbre rather than irrelevant volume features. Visual analysis of the feature space confirmed that the genres form separable clusters, validating the effectiveness of the GMM approach.

References

[1] D. Reynolds, "Gaussian Mixture Models," *Encyclopedia of Biometrics*, MIT Lincoln Laboratory, pp. 1-5.

[2] J. A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," *International Computer Science Institute*, TR-97-021, 1998.