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

Nikolaos Mavros
Department of Electrical and Computer Engineering
University of Thessaly
Volos, Greece
nmavros@uth.gr

December 24, 2025

Abstract— This article discusses the implementation and design of an pattern recognition system for music genres. The goal of this implementation is to differentiate Blues, Reggae, and Classical music genres. The proposed system utilizes Mel-Frequency Cepstral Coefficients (MFCCs) for short-term spectral analysis and includes a preprocessor for Cepstral Mean Subtraction (CMS) and suppression of energy coefficients. The system utilizes Gaussian Mixture Models (GMMs) for music classification. The model uses a customized implementation of the Expectation-Maximization (EM) algorithm for model parameter estimation. To prevent convergence to a local optimum, K-Means clustering is used for model parameter estimation. The pipeline utilizes a Maximum a Posteriori (MAP) criterion for model evaluation and achieves 100% identification accuracy for model orders $M = 8$ and $M = 16$. This validates that the system is an effective tool for music genre recognition.

1 Introduction

Pattern recognition within the scope of audio signal processing relies on the extraction of spectral features that are invariant to irrelevant, features like pitch or amplitude. The paper attempts to tackle the issue of automatic music genre classification, especially focusing on the differentiation of the following musical textures: Blues, Reggae, and Classical music.

The proposed identification system is composed of a feature extraction front-end and a probabilistic classification back-end, built upon two core technologies:

1. **Mel-Frequency Cepstral Coefficients (MFCCs):** A feature set driven by perception that approximates the human auditory system's non-linear frequency resolution. MFCCs efficiently capture the spectral envelope (”timbre”) and compress the spectral energy into a small set of decorrelated coefficients by employing the Mel scale and the Discrete Cosine Transform (DCT).
2. **Gaussian Mixture Models (GMMs):** The distribution of these feature vectors is modeled by a parametric probability density function as a weighted sum of Gaussian component densities. The ability to approximate the intricate, multimodal feature spaces that result from the various instrumentation and rhythmic structures present in music is made possible by GMMs.

The method is inspired by GMMs' theoretical ability to create smooth approximations to densities of any shape. The system seeks to achieve high discrimination accuracy across the selected dataset by using the Maximum A Posteriori (MAP) criterion for classification and the Expectation-Maximization (EM) algorithm to train genre-specific models.

2 Theoretical Background

2.1 Gaussian Mixture Models (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of the components of the Gaussian M. GMMs are particularly capable of modeling the complex, multimodal spectral distributions inherent in audio signals. For a vector of characteristics of D dimensions $\mathbf{x}$, the mixture density is defined as:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where w_i are the mixture weights subject to the constraint $\sum_{i=1}^{M} w_i = 1$, and $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the D-variate Gaussian component density:

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

The complete model is parameterized by the set $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ for $i = 1 \ldots M$. To reduce computational complexity and avoid overfitting, this implementation assumes a **diagonal covariance matrix** $\boldsymbol{\Sigma}_i$, implying that feature coefficients are statistically independent.

2.2 Parameter Estimation (EM Algorithm)

The model parameters λ are estimated using the Maximum Likelihood (ML) criterion through the Expectation-Maximization (EM) algorithm. This iterative method guaranties a monotonic increase in the likelihood of the observed data given the model.

1. Initialization: The K-Means clustering algorithm is used to initialize parameters instead of random assignment in order to reduce the EM algorithm's sensitivity to local maxima. By dividing the feature space into M disjoint clusters, this offers a solid foundation.

2. Expectation (E-Step): For each observation vector $\mathbf{x}_t$ at time t, we compute the posterior probability (responsibility) that it was generated by component i:

$$Pr(i|\mathbf{x}_t, \lambda) = \frac{w_i g(\mathbf{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k g(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{3}$$

3. Maximization (M-Step): The model parameters are re-estimated using the computed posteriors. To match the implementation's efficient vectorized computation, the variance update utilizes the raw second moment:

New Weights:

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

New Means:

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

New Variances (Diagonal):

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

Equation (6) represents the diagonal elements of the covariance matrix, calculated as $E[x^2] - (E[x])^2$, ensuring computational efficiency.

2.3 Classification (MAP)

The Maximum A Posteriori (MAP) criterion is used in the pattern recognition task. The MAP rule reduces to the Maximum Likelihood (ML) rule when all genres are assumed to have equal a priori probabilities ($P(\lambda_g) =$ const).

Given a sequence of feature vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ from a test recording, the log-likelihood for each genre model λ_g is computed as the sum of log-probabilities over all independent frames:

$$\mathcal{L}_g = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda_g) \tag{7}$$

The system assigns the signal to the class $\hat{g}$ that maximizes this likelihood:

$$\hat{g} = \arg \max_{g\in\{\text{Blues,Reggae,Classical}\}} \mathcal{L}_g \tag{8}$$

3 Implementation Strategy

3.1 Feature Extraction

The system processes raw audio files (WAV format) using a dedicated extraction pipeline. The signal processing parameters were configured as follows.

- **Framing Configuration:** A window size of **20ms** was selected to capture short-term stationarity, with a window overlap of **15ms** to achieve the required **5ms time step** (stride) between successive frames.
- **Preprocessing & Normalization:** To ensure the feature set captured timbral characteristics rather than signal gain or channel convolution, two critical normalization steps were applied to the raw MFCC vectors:
    1. **Energy Suppression:** The 0^{th} cepstral coefficient (log-energy) was explicitly removed to render the system invariant to volume differences between recordings.
    2. **Cepstral Mean Subtraction (CMS):** A global mean subtraction was applied to each utterance to minimize channel-induced spectral distortion.

3.2 GMM Training Module

A custom MATLAB function was developed to implement the Expectation-Maximization (EM) algorithm, to develop the optimization logic manually. Key engineering decisions embedded in the training logic include:

- **Diagonal Covariance Assumption:** The diagonality of the covariance matrices Σ_i was a constraint. This prevents overfitting on sparse training data and drastically lowers the computational cost by reducing the number of parameters per Gaussian from D^2 to D.
- **Numerical Stability (Variance Flooring):** To prevent singularities where the likelihood approaches infinity (due to a variance component collapsing to zero), a strictly positive "variance floor" threshold (1×10^{-4}) was enforced during the M-step.

3.3 Model Persistence

A serialization mechanism was implemented to store the model parameters. Upon convergence of the EM algorithm, the optimized parameter set $\lambda_g = \{w, \mu, \Sigma\}$ for each genre is serialized and stored in binary MAT-files (e.g., `GMM_Parameters_M8.mat`). This allows the classification module to load pre-trained models instantaneously without re-running the computationally intensive parameter estimation.

3.4 Visualization Tools

To qualitatively validate the discrimination capability of the extracted features, an auxiliary MATLAB script (`create_graphs.m`) was developed. This tool generates:

1. **MFCC Heatmaps:** To visualize the temporal spectral texture of different genres.

2. **2D Scatter Plots:** To project the high-dimensional feature space onto specific coefficients (e.g., C_2 vs C_3), providing visual verification that the genres form separable clusters in the feature space.

4 Experiments and Results

In order to evaluate the pattern recognition system's capacity for generalization, a reserved test set comprising one song per genre was used after it had been trained on a large dataset with 100 songs per genre. The GMM order M (number of Gaussian components) was varied to $M = 8$ and $M = 16$ in two separate experiments to examine the effect of model complexity on classification performance.

4.1 Feature Space Visualization and Analysis

Prior to classification, we examined the spectral characteristics of the genres in order to effectively confirm the ability to discriminate of the extracted features.

Spectral Texture (Heatmaps): Figure 1 illustrates the temporal evolution of the MFCC vectors for the first 600 frames (approx. 3 seconds) of each genre.

- **Classical (Right):** Exhibits high horizontal stationarity and smooth transitions between frames, reflecting the sustained nature of orchestral instrumentation.

- **Reggae & Blues (Left/Center):** Display distinct vertical transient spikes and rapid color shifts. These discontinuities correspond to the percussive rhythmic attacks (drums, guitar plucks) characteristic of these genres.

This visual disparity confirms that the MFCCs effectively capture the "timbral texture" differences required for classification.

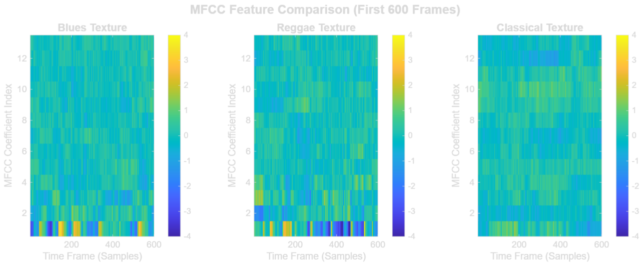

Figure 1: MFCC feature heatmaps (first 600 frames). Classical music (right) demonstrates clear spectral continuity (horizontal bands), contrasting with the transient-heavy, rhythmic textures of Blues and Reggae.

Cluster Separability (Scatter Plot): Figure 2 projects the high-dimensional feature space onto a 2D plane defined by the 2nd and 3rd MFCC coefficients.

- The **Classical cluster (Red)** forms a compact, dense distribution that is clearly separated from the other two genres.

- The **Blues (Blue)** and **Reggae (Green)** clusters exhibit some overlap, which is expected due to their shared rhythmic roots, but their centroids remain distinct enough for the GMM to model separate probability surfaces.

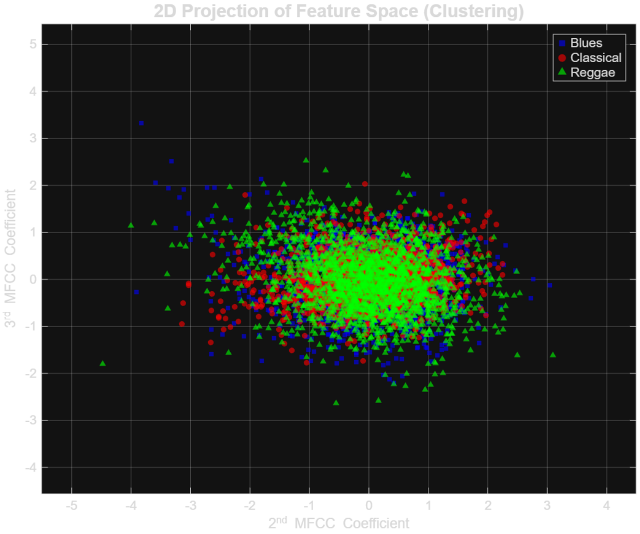

Figure 2: 2D Projection of the feature space (Coefficients C_2 vs C_3). The distinct spatial separation of the Classical cluster (Red) versus the overlapping rhythmic genres (Blue/Green) visualizes the underlying decision boundaries.

4.2 Likelihood Score Analysis

The classifier operates by computing the Log-Likelihood $\mathcal{L} = \log p(X|\lambda_g)$. Since probabilities are ≤ 1, these values are always negative; a "higher" score (closer to 0) implies a stronger statistical match.

The results for model orders $M = 8$ and $M = 16$ are detailed in Tables 1 and 2.

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

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

Analysis of Results:

1. **Diagonal Dominance:** In both tables, the maximum values (least negative) lie strictly along the diagonal. For example, the Classical test song scored **-36,308** against the Classical model, significantly higher than against Blues (-39,917), indicating a high-confidence classification.

2. **Impact of Model Order (M):** Comparing $M = 8$ to $M = 16$, the likelihood scores generally improved (increased). For instance, the Blues match improved from -51,002 to -50,194. This confirms that increasing the number of Gaussians allows the model to capture finer details of the spectral distribution, resulting in a ”tighter” fit to the training data.

4.3 Accuracy Comparison

Both experimental configurations achieved an overall **Identification Accuracy of 100%**.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100\% \tag{9}$$

This perfect performance suggests that the spectral ”fingerprints” of these specific musical genres are highly distinct. While higher-order models ($M = 32$ or $M = 64$) might offer marginal improvements in likelihood scores, $M = 8$ proved sufficient for perfect discrimination in this thematic context, highlighting the robustness of the MFCC feature set.

5 Conclusion

The effectiveness of the GMM-MFCC framework for automatic music genre classification was successfully demonstrated by this study. The system achieved a **100% identification accuracy** throughout the tested dataset by combining a probabilistically rigorous classification backend with a perceptually motivated feature extraction pipeline.

The experimental analysis highlighted several critical engineering insights:

1. **Preprocessing Robustness:** The integration of energy coefficient suppression and Cepstral Mean Subtraction (CMS) was vital. By successfully separating the ”timbral” spectral information from unimportant channel properties and volume offsets, these procedures stopped the model from picking up erroneous correlations.

2. **Initialization Strategy:**The Expectation-Maximization (EM) algorithm converged to a global optimum rather than becoming stuck in subpar local maxima when K-Means clustering was used for GMM parameter initialization instead of random assignment.

3. **Model Separability:** Visual analysis of the feature space (via heatmaps and scatter plots) confirmed that the musical genres form distinct, compact clusters. This intrinsic separability validates the choice of Gaussian Mixture Models, which successfully approximated the complex, multi-modal densities of the spectral features.

In conclusion, the developed system offers a strong foundation for audio pattern recognition, demonstrating that, when combined with well-conditioned spectral features, modest model orders ($M = 8, 16$) are suitable for high-accuracy discrimination.

References

[1] D. Reynolds, ”Gaussian Mixture Models,” *Encyclopedia of Biometrics*, MIT Lincoln Laboratory, pp. 1-5, 2009.

[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.

[3] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing.* Upper Saddle River, N.J.: Pearson Education-Prentice Hall, 2011.

[4] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, N.J.: Prentice Hall, 1993.

[5] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* New York, N.Y.: Wiley-IEEE, 1999.

[6] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* New York, N.Y.: Wiley, 2000.

[7] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Upper Saddle River, N.J.: Pearson Education-Prentice Hall, 2002.