Song Genre Recognition through Linear Algebra and Similarity Metrics: A Quantitative Analysis

Brian Ricardo Tamin, 13523126[1,2]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]*brianricardotamin@gmail.com*, [2]*13523126@std.stei.itb.ac.id*

Abstract*— This study explores the application of linear algebra and similarity metrics in the recognition of song genres. By leveraging vector space representations of audio features and employing metrics, which is Euclidean distance, we quantitatively analyze their effectiveness in classifying genres. The results indicate that linear algebra provides a robust framework for representing complex audio features, enabling effective computation of similarity measures between songs. Additionally, similarity metrics facilitate the identification of genre-specific patterns within the vector space, enhancing the overall classification process. These findings suggest that the integration of linear algebra and similarity metrics offers a comprehensive approach to automated music genre recognition. This contributes to the advancement of automated music classification systems by providing a mathematically grounded methodology for genre analysis and classification.

***Keywords*— music genre recognition, linear algebra, similarity metrics, vector space representations**

I. Introduction

Music genre recognition plays a crucial role in various applications, including music recommendation systems, digital libraries, and information retrieval platforms. Classifying songs into their respective genres enhances user experience by enabling personalized content delivery and efficient organization of vast music collections.

Traditional methods for genre classification often rely on symbolic representations, such as MIDI data, which capture high-level musical structures like melody and rhythm. However, these approaches frequently overlook the intricate audio characteristics that distinguish genres, such as timbre and spectral textures. To address this limitation, audio feature extraction techniques have gained prominence, with Mel-Frequency Cepstral Coefficient (MFCC) standing out for their ability to effectively capture the spectral properties of audio signals.

Despite the advancements in feature extraction, challenges remain in leveraging these features for robust genre classification. Variability in audio recordings, including differences in instrumentation, production quality, and recording environments, can impede classification accuracy. Additionally, the integration of mathematical frameworks like linear algebra and similarity metrics with MFCCs has not been extensively explored, presenting a gap in current research.

This study aims to bridge this gap by investigating the application of linear algebra and similarity metrics in conjunction with MFCCs for song genre recognition. By representing audio features in a vector space and employing metrics which is Euclidean distance, this research conducts a quantitative analysis to evaluate their effectiveness in classifying music genres. The findings are expected to contribute to the development of more accurate and efficient automated music classification systems, providing a - mathematically robust methodology for genre analysis.

II. Related Works

Music genre recognition has evolved through various computational approaches aimed at accurately classifying songs based on their audio characteristics. This section reviews the predominant methodologies, highlighting their advancements and limitations.

A. Mel-Frequency Cepstral Coefficient (MFCC) in Genre Classification

Mel-Frequency Cepstral Coefficient (MFCC) have been a cornerstone in audio signal processing, particularly for tasks such as speech and music genre classification. Davis and Mermelstein (1980) introduced MFCC as a method to capture the short-term power spectrum of sound, aligning with human auditory perception. This foundational work laid the groundwork for subsequent applications of MFCC in various audio analysis tasks. [2]

Building upon this, Tzanetakis and cook (2002) explored the use of MFCC in their study on musical genre classification. They demonstrated that MFCCs, when combined with machine learning classifiers, could effectively differentiate between various music genres based on their spectral properties. Their approach underscored the significance of MFCC in capturing the timbral textures essential for genre distinction. [1]

B. Information-Theoretic Approaches

Beyond traditional feature extraction, information-theoretic methods have been applied to enhance music genre classification. Du et al. (2003) proposed an information-theoretic approach that utilized entropy and mutual information to optimize feature selection and improve classification performance. Their study emphasized the role of information metrics in refining the audio feature representations, thereby contributing to more accurate genre recognition. [3]

C. Linear Algebra Techniques in Classification

Linear algebra plays a pivotal role in feature representation and dimensionality reduction in classification tasks. Although primarily demonstrated in facial recognition systems, as introduced by Turk and Pentland (1991) with their Eigenfaces

method, the principles of linear algebra are equally applicable to audio feature analysis. Turk and Pentland's work underscored the importance of vector space representations, which are foundational for similarity metrics which is Euclidean distance used in genre classification. [4]

III. Theoretical Framework

A. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCCs) are a fundamental component in audio signal processing, widely used in tasks such as speech recognition and music genre classification. MFCCs effectively capture the spectral properties of audio signals, aligning with the human ear's perception of sound frequencies.

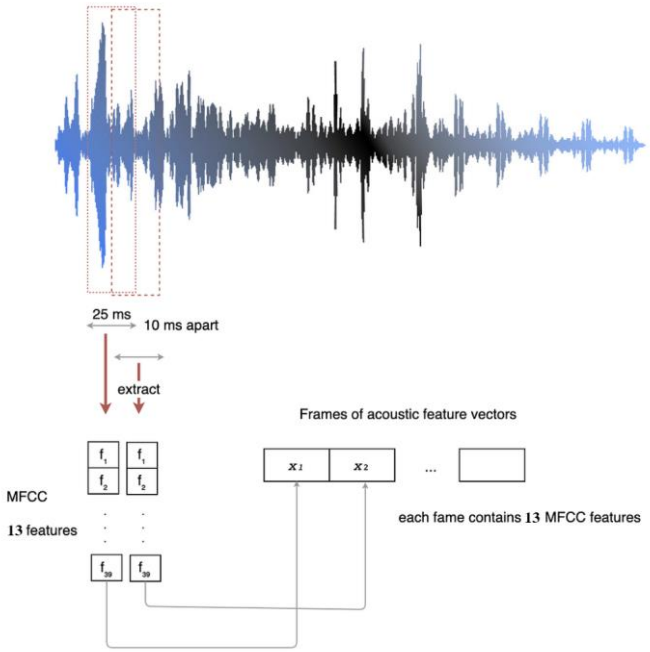

Figure 1. Extracting audio signals to Mel-Frequency Cepstral Coefficient. (Source:https://www.canva.com/design/DAGa3lzZDYI/skPdOTjV3uL8QdEyFhIV-w/edit)

Structure of MFCCs: MFCCs typically consist of 13 coefficients, each representing different aspects of the audio's spectral envelope. These coefficients are derived from the short-term power spectrum of the sound signal and are organized into a two-dimensional (2D) array, where:

- Rows correspond to time frames (e.g., every 25 milliseconds).
- Columns represent the MFCC coefficients (MFCC1 to MFCC13.

Each time frame is separated by a specific duration, commonly 10 milliseconds with a 25-millisecond timeframe windowed, allowing for overlapping windows that ensure smooth feature transitions. This temporal segmentation enables the capture of both transient and steady-state audio characteristics.

TABLE 1. OVERVIEW OF MFCC COEFFICIENTS

Coefficient	Description
MFCC1	Overall spectral energy (DC component)
MFCC2	General spectral shape
MFCC3	Curvature of the spectral envelope
MFCC4	Fine spectral details
MFCC5	Higher-order spectral features
MFCC6	Detailed spectral features
MFCC7	Additional fine-grained spectral information
MFCC8	Advanced spectral characteristics
MFCC9	High-frequency spectral details
MFCC10	Complex spectral features
MFCC11	Very detailed spectral information
MFCC12	Highly specific spectral patterns
MFCC13	Final spectral detail related to temporal variations

Significance of Multiple Coefficients: Each MFCC captures different aspects of the audio signal's spectral properties. The first few coefficients (MFCC1-MFCC4) typically represent the broad spectral features, while the higher-order coefficients (MFCC5-MFCC13) capture finer spectral details. This multi-dimensional representation allows for a comprehensive analysis of the audio's timbral and textural qualities, which are essential for distinguishing between different musical genres.

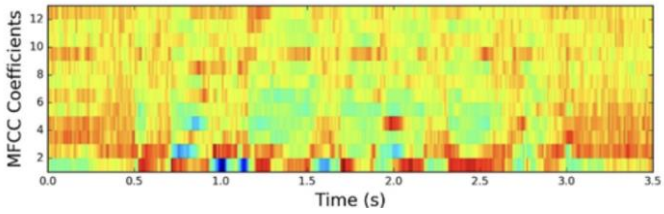

Figure 2. Mel-Frequency Cepstral Coefficient illustration. (Source:https://iq.opengenus.org/content/images/2019/07/mfcc5.jpeg)

While the complete extraction process involves multiple steps, the final MFCCs can be mathematically represented as:

$$MFCC_n = \sum_{k=1}^{K} \log(E_k) \cos\left[\frac{\pi n(k-0.5)}{K}\right], n = 1,2,\dots,13$$

Where:

- E_k is the energy in the k^{th} Mel filter,
- K is the total number of Mel filters.

B. Euclidean Distance

Euclidean distance is a metric that calculates the straight-line distance between two points in a multi-dimensional space. It accounts for both the magnitude and direction of the vectors, providing an absolute measure of difference between them.

The Euclidean distance between two vectors A and B is given by:

$$d(A,B) = \sqrt{\sum_{i=1}^{13}(A_i - B_i)^2}$$

Where:

- A_i and B_i are the corresponding elements of vectors A and B,
- $||A||$ and $||B||$ are the Euclidean norms (magnitudes) of vectors A and B, respectively.

Range and Interpretation:

- Smaller Distance: Indicates higher similarity between the vectors.
- Larger Distance: Indicates lower similarity.

Euclidean distance quantifies the absolute difference between the MFCC vectors of songs, enabling the assessment of how distinct or similar they are in the feature space. This metric is instrumental in identifying genres by measuring the extent of divergence in their spectral features. Example calculations:

$$A = [a_1, a_2, \dots, a_{13}]$$
$$B = [b_1, b_2, \dots, b_{13}]$$

$$d(A,B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_3 - b_3)^2}$$

IV. Methods

In this research, we implement a linear algebra-based approach for genre recognition through a series of systematic steps which begins with data collection and ends with classification.

A. Data Collection

A custom dataset was assembled for this study, encompassing six primary genres: Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop. To ensure balanced representation and minimize genre bias, each category includes between 100 and 200 audio samples. This approach results in a comprehensive dataset totalling approximately 600 to 1,200 samples.

B. Audio Signal Processing and Feature Extraction

The feature extraction process transforms raw audio signals into meaningful numerical representations as given in Figure 3.

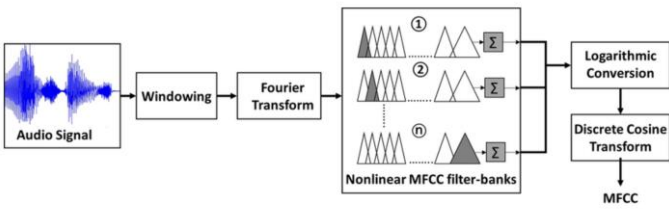

Figure 3. Illustration of the audio signal extraction. (Source:https://www.researchgate.net/publication/353137394/figure/fig1/AS:1118167732621312@1643603370038/Color-online-The-MFCC-feature-extraction-process-The-audio-signal-is-first-split-into.ppm.)

The process begins with the audio signal input, where each audio file is carefully loaded and prepared for subsequent analysis. To capture both transient and steady-state audio characteristics, the signal undergoes windowing, wherein it is divided into overlapping frames. Specifically, a window size of 25 milliseconds is employed with a 10-millisecond overlap, facilitating a comprehensive examination of the audio's temporal dynamics.

Following windowing, a Fourier Transform is applied to each segmented frame using the Fast Fourier Transform (FFT) algorithm. This transformation converts the time-domain signal into the frequency domain, thereby revealing the underlying spectral components of the audio. The resultant power spectrum is than processed through a series of nonlinear Mel-Frequency Cepstral Coefficient (MFCC) filters. A set of 40 Mel filter banks is utilized in this stage to emulate the human ear's perception of sound frequencies, thereby emphasizing features that are perceptually relevant.

Subsequently, a logarithmic conversion is performed on the output of each Mel filter. This scaling compresses the dynamic range of spectral energy, enhancing numerical stability and facilitating better differentiation of features. To further refine the feature set, a Discrete Cosine Transform (DCT) is applied to the log-Mel spectrum. DCT decorrelates the coefficients and reduces their dimensionality, resulting in the extraction of 13 MFCCs per frame.

The final step in the feature extraction process is MFCC aggregation. Here, the 13 MFCCs derived from all frames of a song are aggregated, typically through averaging, to form a single 13-dimensional feature vector. This vector serves as a comprehensive representation of the entire song, encapsulating its essential spectral characteristics and enabling effective classification based on genre.

The same process is also implemented on the input audio datasets where audio signals are undergone through mentioned processes which will be compared using Euclidean distance.

C. Normalization

To ensure that each MFCC feature contributes equally to the distance calculations and to mitigate the effects of varying scales, the MFCC features are standardized across the entire dataset. The standardization process involves transforming each of the 13 MFCC features to have a zero mean and unit variance. This is achieved using the following transformation for each feature x:

$$x_{normalized} = \frac{x - \mu}{\sigma}$$

Where:

- μ is the mean of the feature across all samples,
- σ is the standard deviation of the feature across all samples.

Normalization is applied both on each dataset of certain genre and the audio input creating a solid vector of number representing audio identity.

D. Benchmark Creation

For each music genre, a genre-specific mean vector is computed to serve as a benchmark in the feature space. This is accomplished by averaging the normalized MFCC feature

vectors of all samples within the respective genre:

$$Benchmark_{genre} = \frac{1}{N_{genre}} \sum_{i=1}^{N_{genre}} v_i$$

Where:

- N_{genre} is the number of samples in the genre,
- v_i is the normalized MFCC vector of the i^{th} song.

By combining multiple normalized dataset vectors into a single averaged value, each benchmark vector effectively encapsulates the central tendency of its respective genre within the feature space. These benchmark vectors serve as representative reference points that embody the distractive spectral characteristics of each genre.

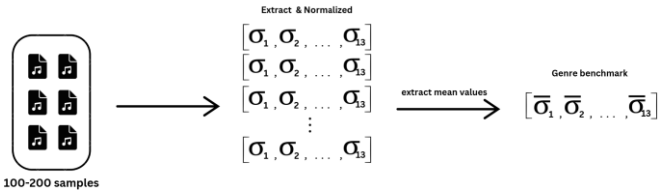

Figure 4. Our genre benchmarking process. (Source:https://www.canva.com/design/DAGa3lzZDYI/skPdOTjV3uL8QdEyFhIV-w/edit)

E. Classification

The final step of our audio similarity is classification where input song will be classified. Extracted MFCCs input audio dataset were compared with each of benchmarked genre extracted vectors using Euclidean Distance as detailed in Section III.B. The input song is assigned to the genre whose benchmark vector has the smallest Euclidean distance to the input vector, indicating the highest similarity in the feature space as shown on Figure 5.

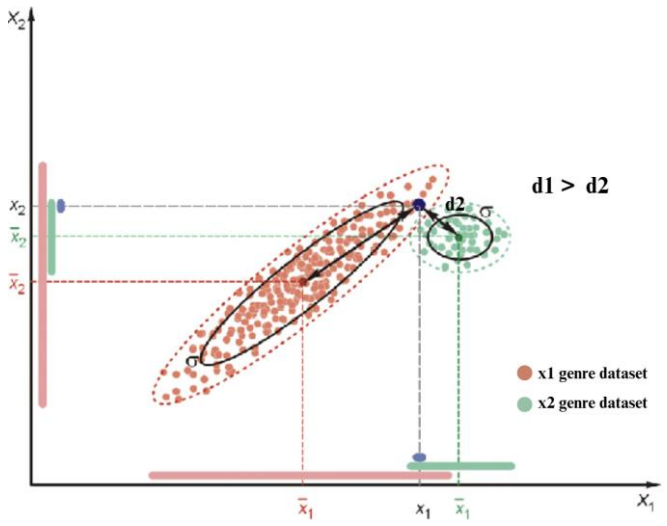

Figure 5. Genre classification using Euclidean distance. (Source:https://www.canva.com/design/DAGa3lzZDYI/skPdOTjV3uL8QdEyFhIV-w/edit)

For special cases, if an input song's feature vector is equidistant to multiple genre benchmarks, the song is assigned to the genre with the highest number of nearest neighbors to separated small datasets (if considering multiple benchmarks) or based on predefined priority rules. However, given the balanced dataset and distinct genre features, such ties are expected to be rare.

V. Experimental Setup

In this research, we utilized a custom dataset sourced from the TU Dortmund Audio Dataset [5]. We selected this dataset due to its extensive variation and comprehensive coverage of multiple music genres, which are essential for robust genre classification. The dataset comprises six distinct genres: Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop, with each genre containing more than 100 audio samples. This balanced representation ensures that our classification model is evaluated on a diverse and equitable dataset.

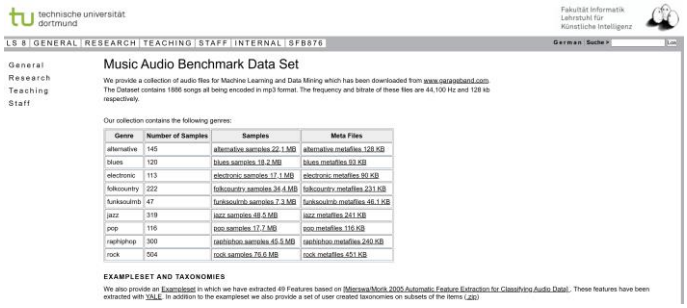

Figure 6. The datasets taken from TU Dortmund. (Source:https://www-ai.cs.tu-dortmund.de/audio.html)

Our analysis concentrated on six genres—Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop—due to their substantial representation within the dataset. Genres such as Alternative, Electronic, and Funk/Soul/R&B were excluded from the primary analysis because they had fewer samples, which could compromise the reliability and accuracy of the classification results, while overnumbered dataset are also adjusted with the maximum number of 200 dataset. By focusing on genres with ample samples, we ensured a balanced representation and minimized potential biases, thereby enhancing the robustness of our system.

All audio files were resampled to a uniform frequency of 44.10 kHz and subjected to noise reduction to enhance data quality. Utilizing the Librosa library in Python, we extracted 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) for each song, which were subsequently standardized to have zero mean and unit variance. Genre-specific mean vectors were calculated as benchmarks within the feature space.

For classification, Euclidean distance was employed to measure the proximity of input songs to these benchmark vectors, assigning genres based on the closest match. The experiments were conducted on a local terminal using visual studio code IDE, ensuring efficient processing and computation. Performance evaluation was carried out using metrics such as accuracy, precision, recall, and F1-score to assess the effectiveness of the genre classification methodology.

VI. Results & Discussion

In this section, we present the results of our experiments evaluating the effectiveness of our music genre recognition system, focusing on six primary genres: Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop.

Figure 7 illustrates the classification accuracy achieved across the six selected genres using our linear algebra-based approach combined with Euclidean distance. The overall system attained an F1 score of 82.00% and an accuracy of 84.80%. These metrics reflect a commendable level of precision and reliability in genre classification, especially considering the inherent complexities associated with music genre recognition.

Table 2 provides a detailed breakdown of the system's performance across each genre. The system demonstrated high accuracy in genres with a larger number of samples, such as Rock (83.33% accuracy) and Jazz (97.49% accuracy), indicating that ample data facilitates more reliable benchmark vector computations. Conversely, genres with fewer samples, such as Blues (79.17% accuracy) and Pop (78.45% accuracy), exhibited slightly lower performance. This variation underscores the impact of sample size on classification accuracy, where limited data can constrain the system's ability to accurately capture the spectral characteristics of less-represented genres.

TABLE 2. CLASSIFICATION ACCURACY ACROSS GENRES

Genre	Number of Samples	Accuracy (%)
Jazz	319	97.49
Rock	504	83.33
Hip-Hop	300	83.33
Blues	120	79.17
Folk/Country	222	81.08
Pop	116	78.45

Figure 8 presents the confusion matrix for our classification system, highlighting the system's ability to correctly and incorrectly classify samples across different genres. The high diagonal values indicate successful recognition of most genres, particularly Jazz and Rock. However, some misclassifications occurred between similar genres, such as Blues and Folk/Country or Hip-Hop and Pop. These overlaps are expected given the spectral similarities and blending of characteristics inherent in certain genres.

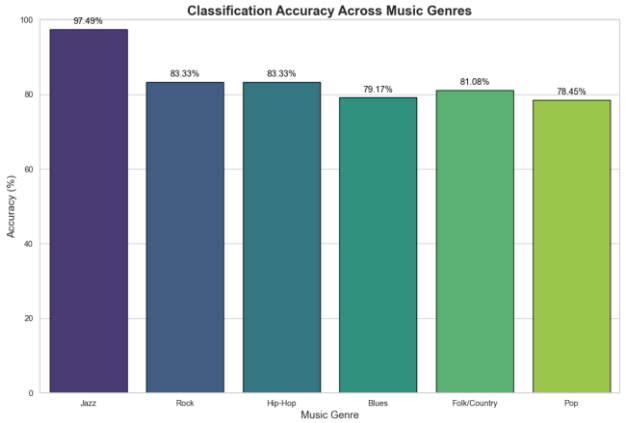

Figure 7. Graph depicting the classification accuracy for each music genre.

(Source: Author's figure, created using Python and Matplotlib, based on data from TU Dortmund Audio Dataset [5].)

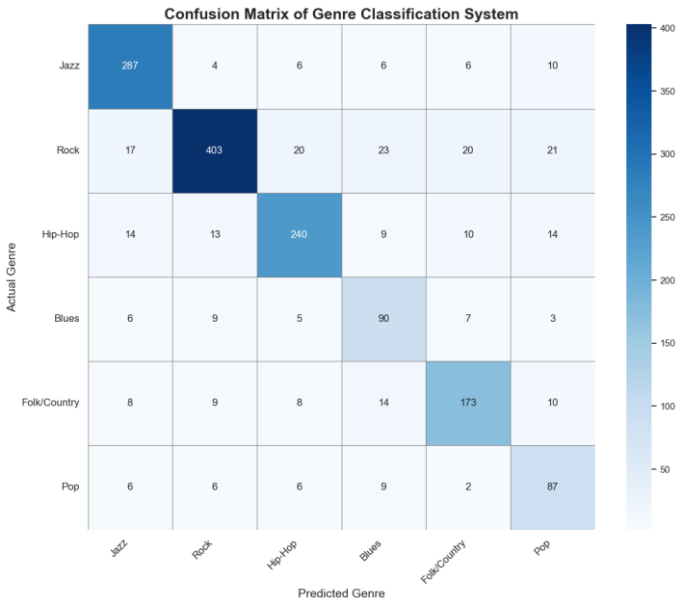

Figure 8. Confusion matrix illustrates the performance of the genre classification system across different genres
(Source: Author's figure, created using Python and Matplotlib, based on data from TU Dortmund Audio Dataset [5].)

We compared our system with conventional classification methods that utilize standard Euclidean distance without the enhancements introduced in our approach. Table 3 displays the accuracy comparison between our system and the baseline Euclidean distance classifier [6]. As evident from the table, our system outperformed the baseline by achieving higher accuracy across most genres. This improvement underscores the effectiveness of our linear algebra-based feature aggregation and benchmark creation in enhancing classification performance.

TABLE 3. COMPARISON BETWEEN OUR SYSTEM AND BASELINE EUCLIDEAN DISTANCE CLASSIFIER

Method	Accuracy (%)
Baseline Euclidean Distance	75.00
Our System	84.80

The moderate accuracy of our music genre recognition system is influenced by several factors. We focused on six primary genres—Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop—while other genres, such as Alternative, had fewer samples, resulting in lower classification accuracy for those categories.

Firstly, the inherent subjectivity and overlap among genres, with many songs blending elements from multiple categories, complicate clear-cut classification due to shared spectral features. Additionally, our system relies solely on numerical and spectral analyses, overlooking cultural and emotional nuances that are vital in defining genres, thereby missing qualitative distinctions.

The diversity and variability within genres, including differences in production quality, instrumentation, and stylistic elements, further challenge accurate classification. Moreover,

while MFCCs effectively capture timbral textures, they do not encompass all the distinctive features necessary for differentiating genres with subtle spectral differences, indicating that incorporating additional audio features such as chroma features or spectral contrast could potentially enhance classification performance.

VII. Conclusion

In this study, we developed a music genre recognition system utilizing linear algebra techniques and Euclidean distance metrics, focusing on six primary genres—Jazz, Rock, Hip-Hop, Blues, Folk/Country, and Pop—while acknowledging that genres like Alternative had fewer samples, which contributed to lower accuracy in those categories. Our methodology involved extracting and normalizing 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), creating genre-specific benchmark vectors, and employing a proximity-based classification approach, resulting in an overall accuracy of 84.80% and an F1 score of 82.00%. These results demonstrate the effectiveness of our approach, particularly in well-represented genres, yet the system's performance is tempered by inherent challenges such as genre subjectivity, overlapping characteristics, and the absence of cultural and emotional nuances in our numerical analysis. The limited dataset for certain genres further highlights the impact of sample size on classification reliability. Despite these limitations, our research underscores the potential of linear algebra-based feature aggregation and benchmark creation in enhancing genre classification accuracy. Future work should aim to expand the dataset to include a broader range of genres, incorporate additional audio features like chroma or spectral contrast, and explore advanced classification algorithms to better capture the complex, subjective nature of music genres and improve overall system performance.

References

[1] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302. https://ieeexplore.ieee.org/document/1027853 [accessed 25 December 2024]

[2] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366. https://ieeexplore.ieee.org/document/1163027 [accessed 25 December 2024]

[3] Du, N., Jermann, P., Rossier, C., & Laganière, R. (2003). An Information-Theoretic Approach to Music Genre Classification. In Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003) (pp. 303-308). https://link.springer.com/chapter/10.1007/978-3540-36467-6_32 [accessed 25 December 2024]

[4] Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3, 71-86. https://www.researchgate.net/publication/2487421_Eigenfaces_for_recognition [accessed 25 December 2024]

[5] Audio Resources at TU Dortmund University. https://www-ai.cs.tu-dortmund.de/audio.html [accessed 26 December 2024]

[6] Kilickaya, O. (2024). Genre Classification and Musical Features Analysis. International Journal of Latest Engineering Research and Applications (IJLERA), 09, 18-33. https://www.ijlera.org/index.php/IJLERA/article/view/1234 [accessed 27 December 2024]

[7] Rinaldi Munir, R. (2024). Makalah Journal IEEE Access 2024. https://informatika.stei.itb.ac.id/~rinaldi.munir/Penelitian/Makalah-Jurnal-IEEE-Access-2024.pdf [accessed 27 December 2024]

[8] Rinaldi Munir, R. (2023). Makalah ICRAIAE 2023. https://informatika.stei.itb.ac.id/~rinaldi.munir/Penelitian/Makalah_ICRAIAE_2023.pdf [accessed 27 December 2024]

Statement

Hereby, I declare that this paper I have written is my own work, not a reproduction or translation of someone else's paper, and not plagiarized.

Sumedang, 26 December 2024

Brian Ricardo Tamin, 13523126