Report


# Music Plagiarism

Review, Implementation and Analysis

Music, Mind and Technology

Spring'23

Team: Symphony

Harshita Gupta (2020101078)

Pratham Gupta (2020101080)

Nukit Tailor (2020114012)

# Abstract

Music plagiarism is when someone uses a significant portion of another person's music composition without permission or proper attribution. In this project, we aim to explore automated ways of detecting and understanding plagiarism in music. We review existing work on plagiarism detection in music in terms of sampling, melody, rhythm, etc., and compare the results of these models. Ultimately we provide a thorough examination of the patterns of music plagiarism among Indian composers, drawing upon the insights obtained during the initial phase of the project.

**Keywords:** Music Plagiarism, Melody, Rhythm, Sampling, Bipartite Graph Matching, Timre, Tempo

# Contents

# 1. Introduction

Music plagiarism is when someone uses a significant portion of another person's music composition without permission or proper attribution. It can also refer to the act of claiming someone else's music as one's own. Plagiarism can occur in many forms, including using the melody, chord progressions, lyrics, or other significant aspects of a song without permission or credit. In many cases, music plagiarism is illegal and can result in legal action being taken against the plagiarizer.

Each year, over 10,000 new albums of recorded music are released and over 100,000 new musical pieces are registered for copyright. However, there are no general rules that set a minimum number of similar notes or beats for music copyright infringement.

## 1.1 Milestones

Here's an outline of the milestones achieved during the course of the project:

1. **Literature review to decide on a bipartite matching-based approach for working with melody plagiarism**

   We have chosen to follow a Smith-Waterman-based bipartite matching algorithm for melody inspection after an extensive and intensive literature review of different possible plagiarism analysis methods.

2. **Generation of a dataset of 154 plagiarized Indian songs and their original versions**

   We refocus our project scope to analyze the patterns of plagiarism by popular Indian Composers typically from Hindi and Tamil languages, Hence we compile a dataset of snippets of their songs along with the original songs that were plagiarized.

3. **Building on existing paper's implementation to give similarity scores for our dataset**

   The code obtained from the research paper that does a melodic similarity sequencing based on a function of pitch, downbeat and tempo is expanded to give similarity scores for song pairs for every composer. These are recorded with the existing database.

4. **Adding genre, county of origin, and musical feature information to expand the dataset**

   The database is expanded by adding genre and region information for all original songs. Additionally, timbre, pitch, and tempo information is added for the purpose of understanding patterns of plagiarism during the analysis phase.

5. **Identifying patterns of plagiarism through statistical analysis**

   Conclusions were drawn by building multiple relevant representations of obtained correlations like a world heat map of plagiarism intensity based on cumulative

similarity scores, correlation matrices between artists and genres copied, and language and musical feature-based analytics.

# 2. Literature Review

## 2.1 Types of Music Plagiarism

### 2.1.1 Sampling

Sampling is the re-use of recorded sounds of music excerpts in another song.

**Methods**

1. The samples are often manipulated in pitch or tempo to fit the rhythm and tonality of the new song.
2. Mix additional instruments to the sample, such as additional vocals or drums.
3. Crop an excerpt of one or more bars and loop them.
4. Rearrangement and post-processing of the respective sample beyond recognition.

**Inspection Methods**

1. Due to the fact that sampling is basically the use of "a song in a song" it is related to the task of cover song detection. Cover song detection is commonly approached by chroma features, i.e., descriptor, which represents the tonal content of a musical audio signal in a condensed form that can be used for harmonicity similarity analysis.
2. Compare a time-frequency representation of both music excerpts by means of Short-Term Fourier Transform (STFT).
   (a) In order to retrieve the occurrences of Xo inside Xs, it is re-sampled both in time and frequency yielding Xo.
   (b) Each Xo is shifted frame-wise along all frames of Xs and the accumulated, absolute difference d is computed between all corresponding time-frequency tiles.
   (c) Assuming only re-sampling and looping were applied, periodic minima will occur in d. These correspond to the point, where an optimal matching can be found.

### 2.1.2 Rhythm

Rhythm is the pattern of sound, silence, and emphasis in a song. In music theory, rhythm refers to the recurrence of notes and rests (silences) in time. When a series of notes and rests repeats, it forms a rhythmic pattern. More formally, rhythm is formed by a periodical pattern of accents in the amplitude envelopes of different frequency bands. Commonly the drums make up the beat or the guitar is playing the rhythm.

**Inspection Methods**

1. Extract rhythmical features such as the beat spectrum or tempo in order to measure rhythmical similarity.
2. We assume, that the original rhythm may have undergone a number of manipulations, such as time stretching, pitch shifting, re-sampling or even shuffling of individual beats: Steps:
    - Rhythmic source separation (using NMF - clustering of the components is necessary since NMF often splits one instrument into several components. The assignment of components to each other is based on evaluating the correlation between the amplitude envelopes.)
    - Tempo alignment to compensate for the difference in the re-sampling factor.
    - Similarity of individual sources using Pearson Coefficient.

### 2.1.3 Melody

Copied melodies are less obvious than the previously explained plagiarism types. A melodic motive is considered to be identical, even if it is transposed to another key, slowed down, sped up or interpreted with different rhythmic accentuation. Thus, melody plagiarism is a grey area, where it is hard to discern copying from the citation.

**Inspection Methods**

- In the MIR literature, a closely related task is Query-by-Humming (QbH). QbH can be used to retrieve songs from a database by letting the user hum or sing the respective melody. Melody plagiarism inspection can be done with basically the same approach since means to identify and evaluate melodic similarity are required. The main difference is, that QbH searches across extensive databases while plagiarism detection concentrates on one single comparison, which has to be more precise.
- Sequence Alignment - relies on the Smith-Waterman algorithm to find a local alignment between symbol sequences. The algorithm tries to identify subsequences of symbols, which encode intervals between consecutive notes in the MIDI transcription. On execution, each of these melody fragments is compared to the entire suspect sequence. The resulting scores are ordered in descending and presented via the graphical user interface.

## 2.2 Plagiarism or inspiration?

We came across an article on the relevance of melody as a marker for plagiarism in Pop and Rock music. “Where lies the threshold of musical plagiarism?” Can we say that the similarities are coincidental which comes down to plagiarism or to something in between that may be labelled ”inspiration? How the concepts of “idea” and “inspiration” translates to the language of music? Musical plagiarism is primarily a matter of a close similarity of a rather long melody. Inspiration concerns more subtle melodic similarities, chord progressions, special effects, style of arrangement, lyrical themes, and so on. Common chord progressions and generic rhythmical patterns cannot be copyright-protected. We don’t know about any cases of plagiarism focusing primarily

on rhythm. The rhythmical similarity is not even an easy-to-notice phenomenon unless you are focusing on it. When comparing melodies, a perfect matching means both, identical pitch and identical rhythm (timing). Melody is usually the deciding element in cases of plagiarism. Similar but not identical fragments can instigate a sense of similarity.

## 2.3 Related Work

### 2.3.1 An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering

The paper tries to identify similarities between melodies of pop music. Their goal was to build an automated system able to take as input two melodies, as MusicXMLfiles, and provide an indication of their similarity (a percentage). The paper presents 2 methods- a text similarity-based method (in which they convert the music sheets into text strings using a technique called PINL representation. This representation uses symbols for pauses, intervals, and base note lengths to represent the melody as a text string.) and a clustering-based method (in which each melody in the knowledge base is first converted into text using the PINL representation and then into a vector of real numbers using the char2vec technique. The char2vec technique is particularly suitable for embedding music representations, and different vector sizes are used for the experiments) and further combine to get an improved hybrid method. To assess the effectiveness of the proposed methods the authors performed tests on a large dataset of ascertained plagiarism and non-plagiarism cases.

### 2.3.2 Identification and Detection of Plagiarism in Music using Machine Learning Algorithms

The authors first perform feature extraction to extract the note or the chord progression then, the harmonic reduction is performed to understand the structure of the music and then using Word2Vec model is applied to get the relationship between similar chords to perform chord substitution which will be the final data that is extracted for the classier models(KNN, Logistic Regression, Random Forest, DecisionTree, Gaussian Naïve Bayes) to predict plagiarism and the results were obtained.

### 2.3.3 Music Plagiarism Detection via Bipartite Graph Matching

This paper proposes a new method (MESMF) which reduces the music plagiarism detection problem into a bipartite graph maximum matching task. The authors designed several kinds of melody representations and similarity computation methods.

#### Audio-based and sheet-based methods

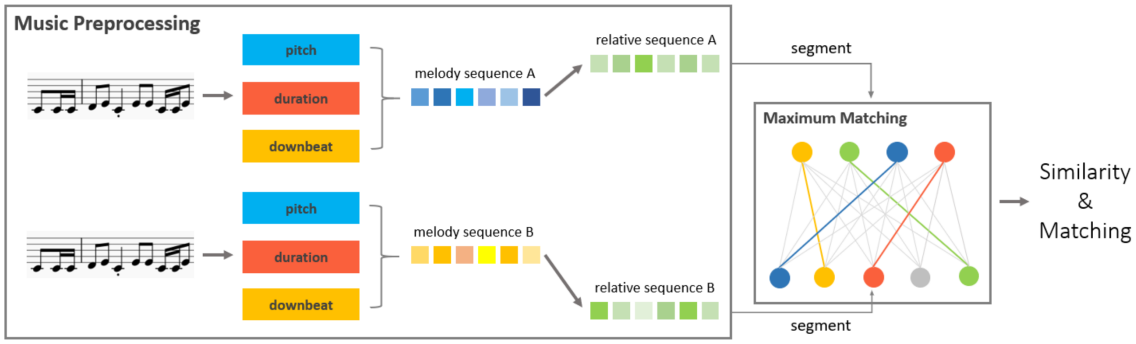
Audio-based methods inspect music similarity using time-frequency representation or low-level features such as cepstrum coefficients. The whole song may be a copy of another one, but if the order of the notes is shifted then the audio-based algorithms cannot detect this situation. Sheet-based methods measure symbolic melodic similarity which plays a crucial role in MIR. Symbolic melodic similarity evaluates the degree of similarity as human listeners can do.

**MESMF (Maximum weight matching and edit distances model applied on the Sequences of Melodic Features)**

Every note can be represented as a pair i.e. it's (pitch and duration). Given two songs A and B, the algorithm extracts their melodic features and transforms them into sequences.

$A_s = (a_1, a_2, \ldots a_n)$, where the ith note of song A, $a_i$ is represented as $ai = (pitch_a i, duration_a i, downbeat_a i)$, where $downbeat_a i$ denotes whether this note is downbeat. The sequences are transformed to relative form by subtraction of pitch and division of duration between neighbouring elements. The two sequences $A_s$ and $B_s$ are divided into pieces with the same length l and overlapping rate r and are treated as nodes in the bipartite graph. We perform the maximum weight matching algorithm to get the final similarity.

The edge between two nodes has a weight equal to the function of the edit distance. The dataset used in the paper consists of some well-known music plagiarism cases.




**Pros**

- The proposed methods deals with shift, swapping, transposition, and tempo variance problems in music plagiarism.
  - Transposition means increasing the pitches of all the notes to the same degree (relative pitch).
  - Tempo invariance means that speeding up or slowing down the musical pieces will not influence the duration sequences we extract from them (relative duration).
- It can also effectively pick out the local similar regions from two musical pieces with relatively low global similarity.

# 3. Data

During our literature review, we discovered a lack of datasets available for music plagiarism, which prompted us to create our own. Our dataset includes over 150 song pairs in Hindi and Tamil, consisting of snippets from original songs and their closest copies. This collection can be used to develop, test, and evaluate music plagiarism

detection algorithms. Additionally, it can provide insight into patterns of music plagiarism, helping to prevent such unethical practices in the music industry. We also included information about the region, artist, and genre of the original songs, shedding light on the cultural and musical influences on the composers. Access to the dataset, containing over 300 music files in .rm formats, can be found here [1], while the dataset itself can be accessed here [2]. Our contribution to the field of music information retrieval can help promote originality in the music industry and encourage ethical practices in music composition.

### 3.1 Metadata

- The plagiarised song is the song that is suspected of being a copy.
- The original song is the song that is believed to have been copied.
- The original composer is the artist or musician who created the original song.
- Plagiarising composer is the one who is suspected of having plagiarised the original song.
- The language is the language of the original song, Hindi or Tamil in our data.
- The genre(s) of the original song provide information about the cultural and musical influences that the original composer drew upon when creating the song.
- The country of origin can provide insight into the where a particular aritst tries to plagiarize from.
- The tempo difference between the original and plagiarised song is a measure of the difference in the underlying rhythm and pace of the two songs.
- The timbral similarity measures the similarity between the tone color or texture of the two songs.
- The melodic similarity score is obtained using bipartite graph matching to measure the degree of similarity between the melodic motifs of the two songs.

# 4. Methods

- **Ground Truth:** The website https://www.itwofs.com/hindi-am.html was used as the reference or "ground truth" for our analysis.
- **Data Collection** We collected the audio rm files, genre, region, and artist information for each song.
- **Midi File** The process involved taking pairs of audio files in the .rm format (original and plagiarized versions) and using a Python library called pydub.AudioSegment to convert them into the more common mp3 format. Each mp3 file was then

[1] https://drive.google.com/drive/u/0/folders/1BnZRpFOcsmUxKKBVGVtinUoPZKtYlHA
[2] https://docs.google.com/spreadsheets/d/162GKmlOb5febj9z7Z6ifFXloxHbUgMqfHPGcTouXNAE/

shortened to a duration of 60 seconds. To further analyze the mp3 files, they were converted into midi files using Spotify's basic_pitch library.

- **Melodic Similarity Score** To determine the melodic similarity between the two midi files, we used the Music Plagiarism Detection via the Bipartite Graph Matching algorithm.
- **Tempo Difference** To determine the tempo of each audio file, we utilized a Python library called librosa.beat.beat_track. Afterwards, we computed the absolute difference between the tempos of the two audio files.
- **Timbral Similarity** To calculate the timbral similarity, we computed the mel-spectrogram and chroma features for each file. We flatten the feature matrices into 1D vectors and concatenated them into a single feature vector. Finally, we computed the cosine distance between the two feature vectors to determine the timbral similarity between the two audio files. Mel-spectrograms are a type of audio feature that represent the power spectral density of a signal, while chroma features represent the pitch content of a signal. The librosa.feature.melspectrogram() function takes an audio signal y and a sampling rate sr as input and returns a matrix of mel-spectrogram coefficients. This matrix contains one row for each mel frequency band and one column for each frame of the audio signal. librosa.feature.melspectrogram() function also takes another parameter n_mels, which specifies the number of mel frequency bands to generate in the spectrogram. In our case, n_mels is set to 128, which means that the mel-spectrogram will have 128 frequency bands. However, after computing the mel-spectrogram, we flatten the resulting matrix into a 1D vector using the reshape() function. This means that each mel-spectrogram contains 128 coefficients, but the resulting feature vector contains 128 times the number of frames in the original mel-spectrogram.
- **Geographical Heatmap** We visualized the region-wise cumulative similarity scores using Geopandas. This helped us understand the distribution of plagiarism cases across different regions.

# 5. Analysis

## 5.1 Region

The melodic similarity scores were cumulated, country wise and then represented on the world map for identifying patterns of plagiarism in terms of region of original music.

### 5.1.1 Anu Malik

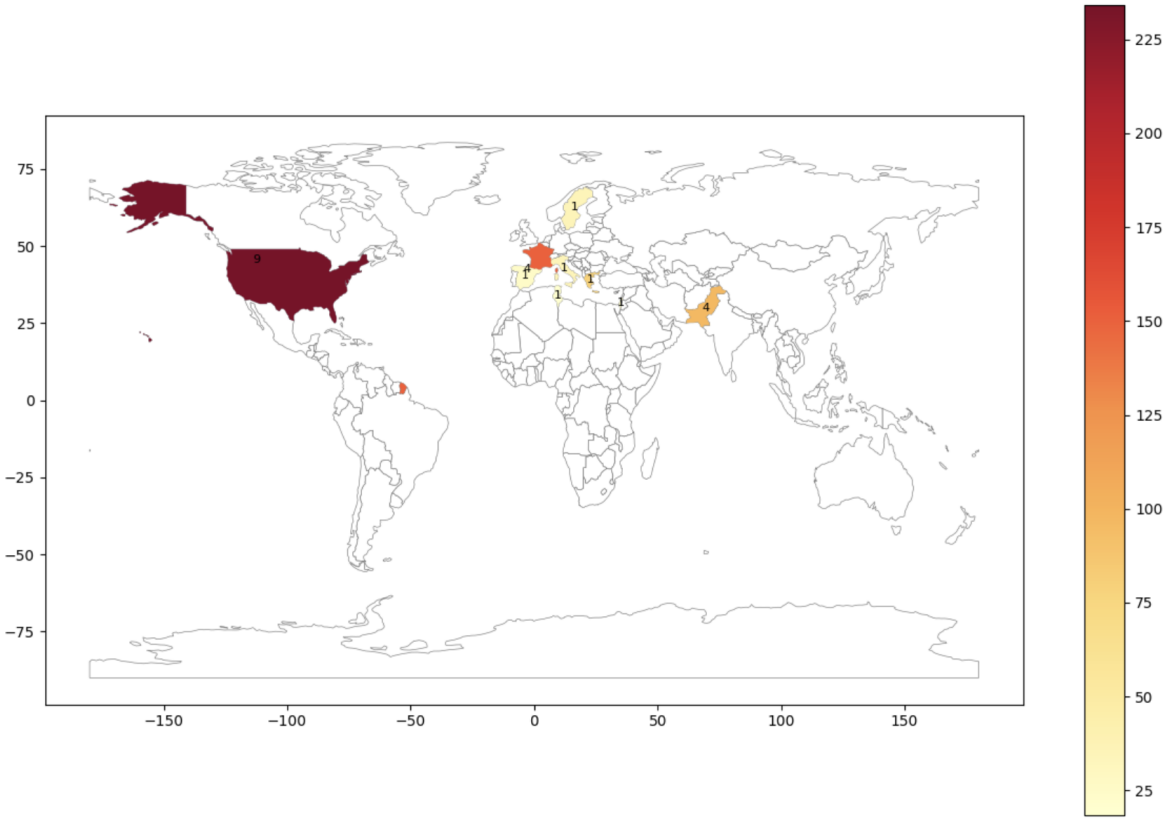



Figure 1: Regional analysis - Anu Malik

Figure 1 marks the countries from which songs have been copied by Anu Malik. A majority of songs have been plagiarized from original American music with intense melodic similarities. Overall very high similarity scores were obtained for most songs indicating minimum effort ripping off of original compositions by Anu Malik.

### 5.1.2 Pritam

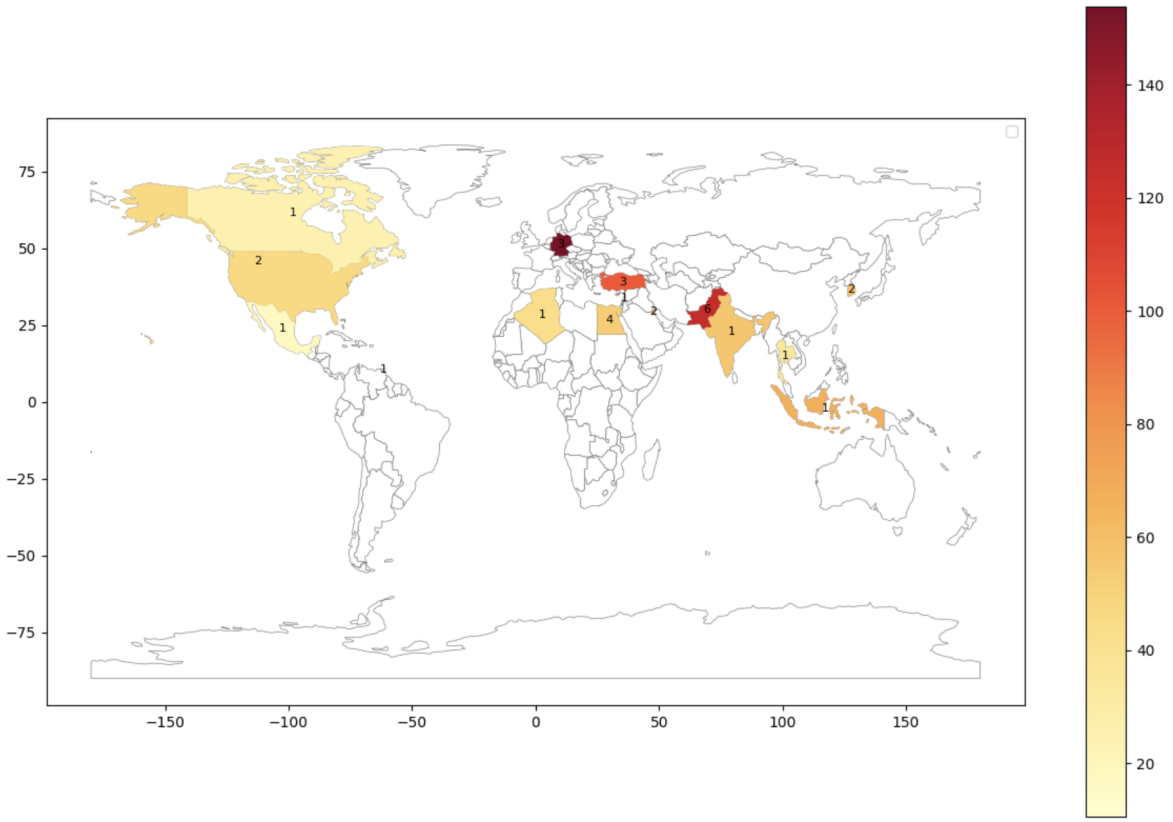



Figure 2: Regional analysis - Pritam

Figure 2 marks the countries from which songs have been copied by Pritam. The results clearly indicate a high usage of German music as shown by intense melodic similarity between original and copied songs. Overall we observe a high influence of German, Middle-eastern and Pakistani music in Pritam's music composition.

### 5.1.3 R D Burman

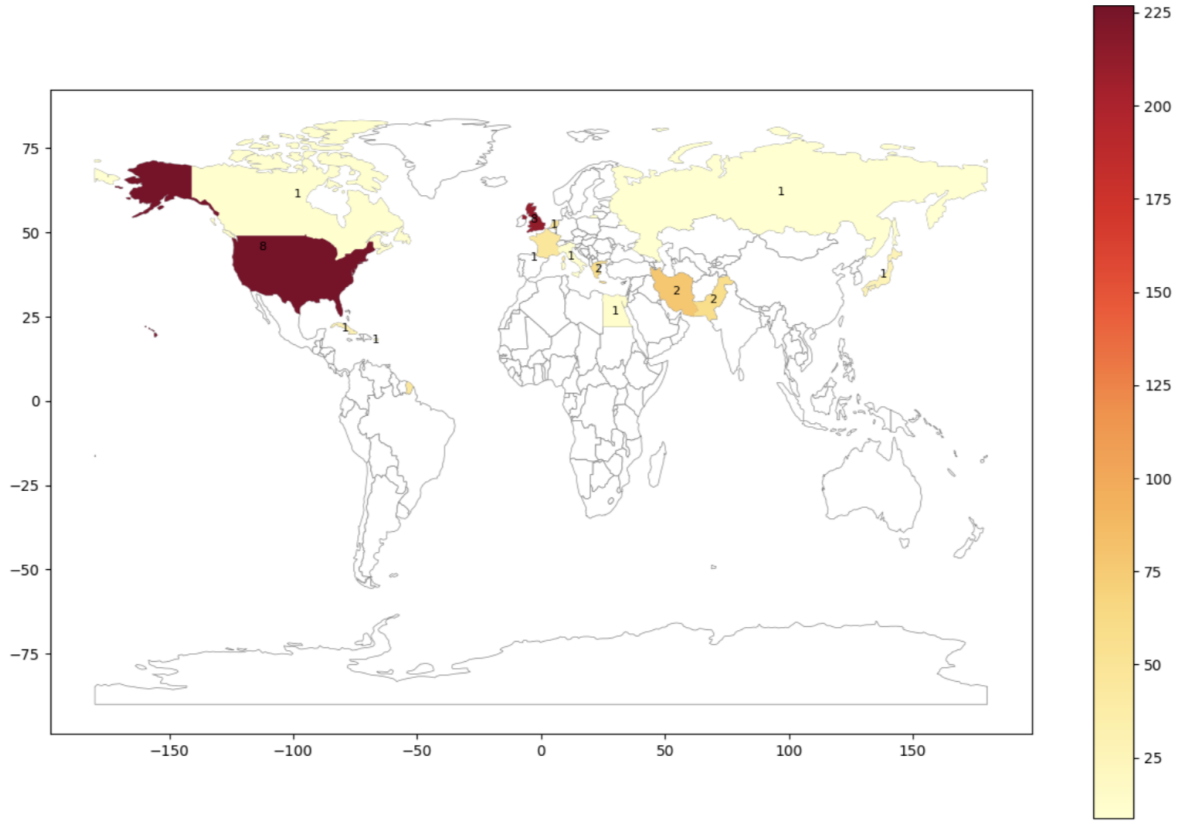



Figure 3: Regional analysis - RD Burman

Figure 3 marks the countries from which songs have been copied by RD Burman. A majority of songs have been plagiarized from English and American music with intense melodic similarities. Additionally music influence of Middle Eastern states like Iran and Egypt and Eastern Europe has been observed.

### 5.1.4 A R Rahman

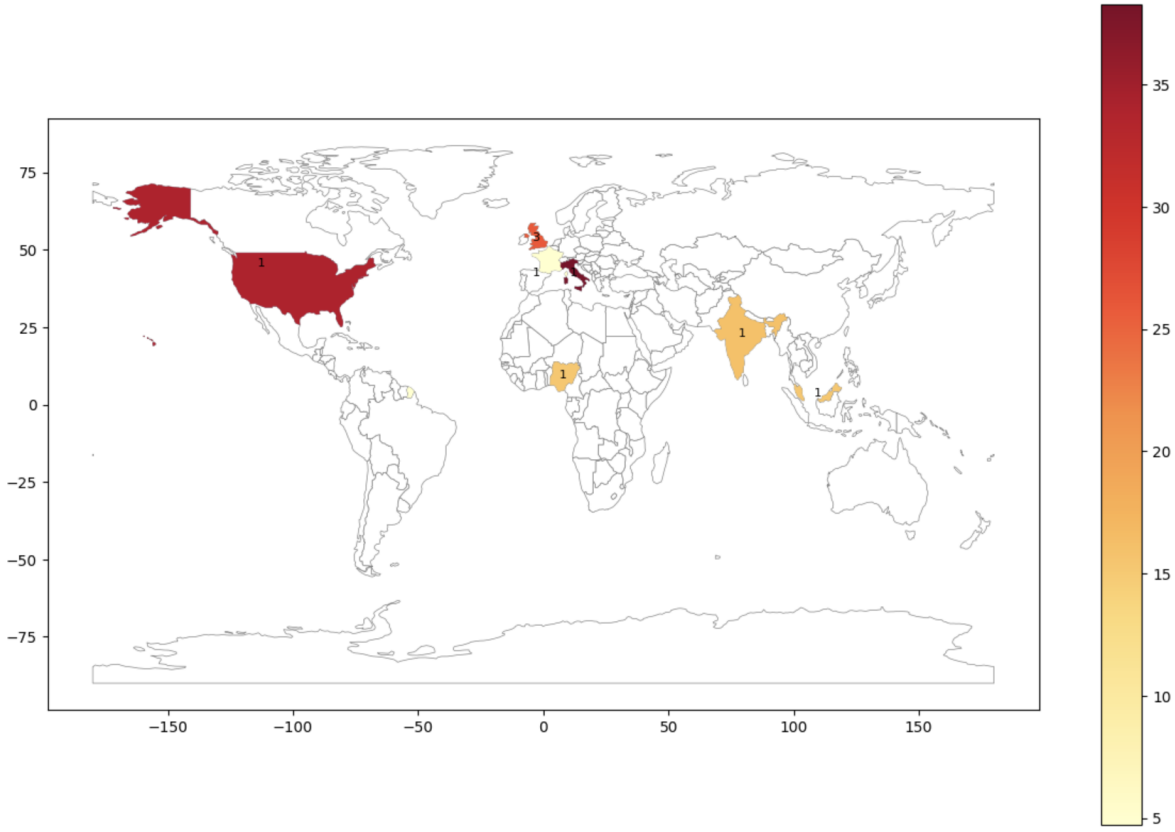



Figure 4: Regional analysis - A R Rehman

Figure 4 marks the countries from which songs have been copied by AR Rahman. A majority of songs have been plagiarized from English and American music with intense melodic similarities.

### 5.1.5 Tamil music

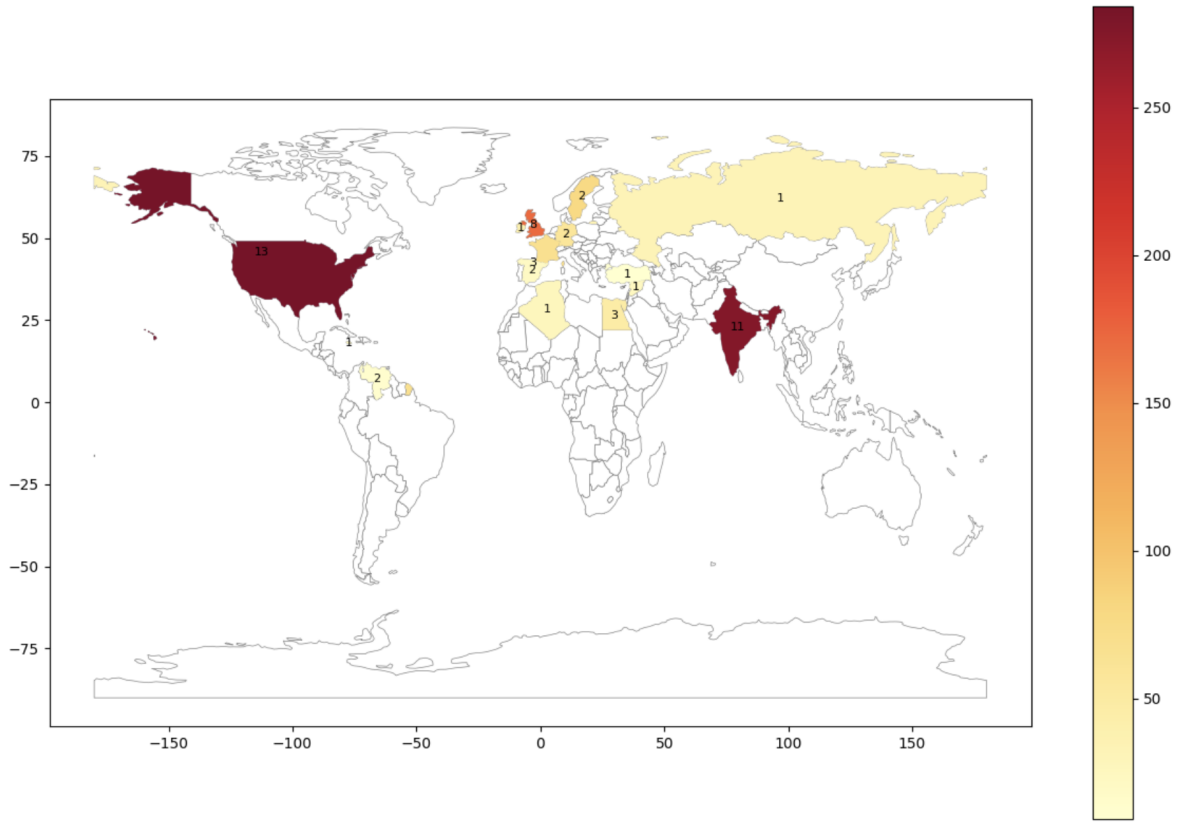



Figure 5: Regional analysis - Tamil music

Figure 5 represents the influence of various countries on Tamil music. A majority of songs plagiarized from foreign music copy English language songs, predominantly from the UK and US. Additionally a high percentage of Tamil music has in fact been copied from ghazals, Indian music originally composed for Hindi films, or folk/classical music from other languages.

## 5.2 Genre

The dataset was annotated with Genres for all original music including their specific regional or niche subclasses. These were used to visualize the prominence of every genre in the extent of its plagiarism. Refer figure 6.

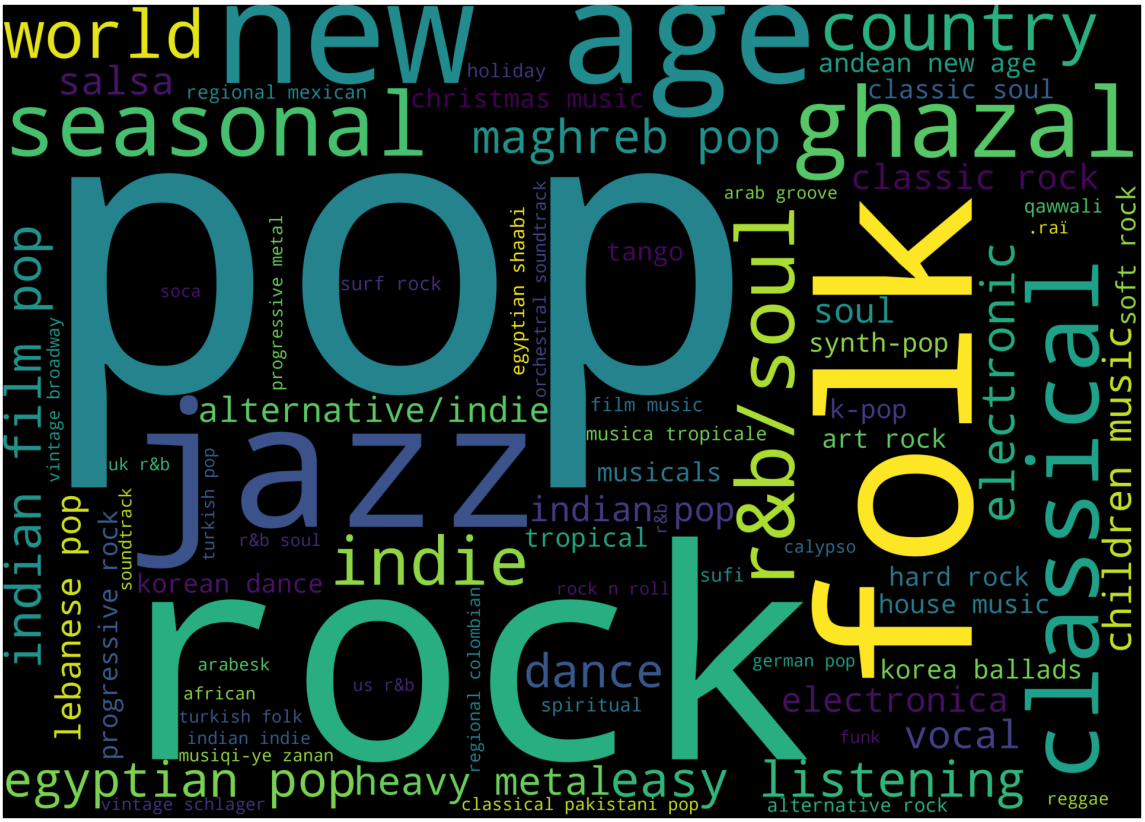



Figure 6: Genre prominence in plagiarism

Additionally, all categories were binned into 10 major genres. The similarity scores were then cumulated genre-wise which after normalization gave a fine insight into what genres the Indian composers were ripping music off the most. Refer figure 7.

| | Pop | Electronic | Soundtrack | Classic | Country | Folk | R&B | Rock | New Age | Jazz | Ghazal | Indie | Heavy Metal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anu Malik | 35.70 | 6.81 | 0.00 | 3.00 | 2.06 | 4.31 | 7.73 | 21.88 | 2.10 | 3.43 | 2.24 | 4.56 | 6.17 |
| Pritam | 38.33 | 0.00 | 0.00 | 12.43 | 2.00 | 9.99 | 0.00 | 15.67 | 9.40 | 0.00 | 3.03 | 6.87 | 2.28 |
| RD Burman | 35.30 | 2.44 | 0.71 | 9.58 | 5.38 | 6.71 | 4.28 | 19.89 | 2.53 | 7.49 | 4.04 | 1.67 | 0.00 |
| AR Rehman | 30.48 | 34.68 | 8.24 | 0.00 | 0.00 | 0.00 | 5.45 | 0.00 | 0.00 | 18.37 | 0.00 | 2.78 | 0.00 |
| Tamil | 48.50 | 8.23 | 8.67 | 3.87 | 0.00 | 14.10 | 1.08 | 11.57 | 0.00 | 0.00 | 3.98 | 0.00 | 0.00 |

Figure 7: Artist-Genre prominence correlation

1. Pop and Electronic are the most common genres for plagiarism across all artists, except for AR Rahman who seems to borrow more from Jazz.
2. Pritam has a higher percentage of plagiarized songs in Classic, Country, Folk, and New Age genres compared to other artists.
3. Anu Malik has a high percentage of plagiarized songs in the Rock, R&B, and Indie genres.
4. RD Burman seems to borrow more from Classic, Country, Jazz, and Ghazal genres compared to other artists.

5. Tamil seems to borrow more from Pop, Electronic, and Folk genres.

## 5.3 Timbral Similarity

Timbral similarity refers to the degree of similarity or resemblance between the sound qualities (characteristic tone colour or texture) of two audio signals. The timbral similarity is assessed by analyzing the spectral content, temporal characteristics, and perceptual features of the audio signals. We observe a positive correlation between the melodic similarity scores derived from the bipartite graph matching of midi files (acoustic features) and the timbral similarity extracted from the .rm files (perceptual features). From this, we can reason that there is no significant loss of information when we convert rm to midi files and the Bipartite Graph Matching algorithm to detect music plagiarism is in line with how people perceive audio similarity. Figure 8

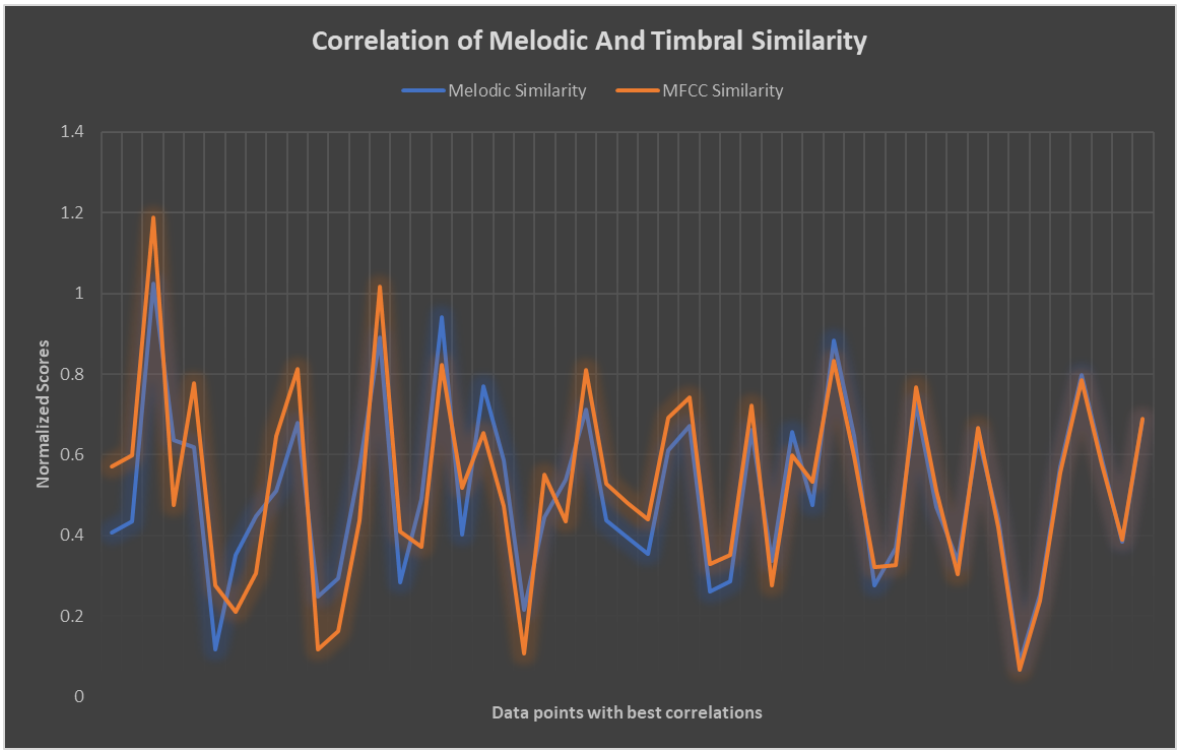



Figure 8: Bipartite Graph matching based Melodic Similarity shows sound correlation with MFCC based Timbral similarity

## 5.4 Tempo Difference

Tempo is the speed or pace of a musical composition. We observe a negative correlation between the melodic similarity scores and tempo difference between the two audio files. From this, we can infer that the way in which we perceive and process tempo similarity is an important aspect of how we relate to music. Figure 9

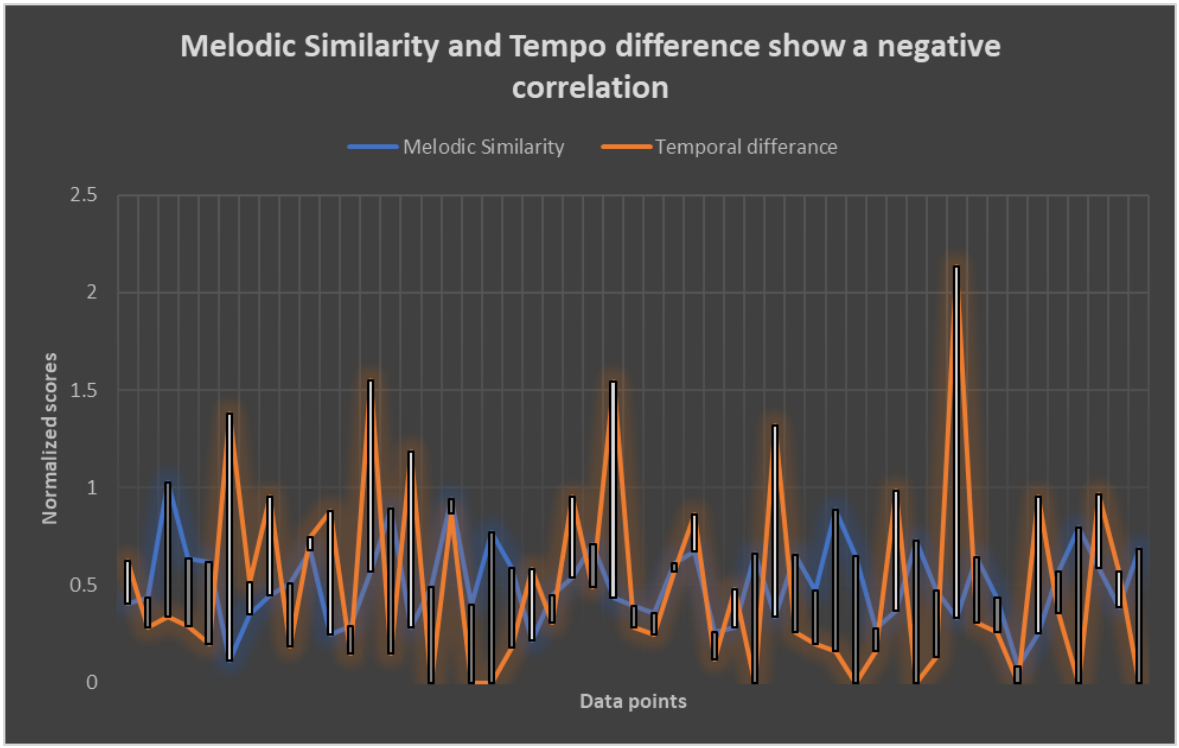



Figure 9: Bipartite Graph matching based Melodic Similarity is inversely dependent on the tempo distance of the two songs

# 6. Conclusion

- Prevalence: our analysis revealed instances of plagiarism where the songs share significant similarities in pitch, tempo and timbre.
- Pop music is a genre that is known for its catchy melodies and memorable hooks, and it is often targeted by artists looking to replicate its success. Rock, electronic, and jazz are other common genres for plagiarism.
- Plagiarism can occur across different genres. Our analysis has revealed instances of plagiarism occurring across different genres, such as Bollywood songs copying from Western pop or Indian classical music.
- Finally, the Smith-Waterman algorithm using tempo, pitch and downbeat for calculation of melodic similarity is an efficient tool not only for the purpose of plagiarism inspection but also mapping timbral similarity.

# 7. Future Work

The data can be expanded to add more artists and languages. This could potentially reveal more instances of plagiarism and provide a more comprehensive picture of the prevalence of plagiarism in the music industry. We can also explore and evaluate other

automated ways of music plagiarism detection using the data. With regard to acoustic features, we explored only timbre and tempo, thus there is scope to explore other audio features such as rhythm and harmony. We can also work to find out how plagiarism is perceived and dealt with in different musical traditions and cultures around the world.

# 8. My Contribution

1. Literature Review – Read and annotated multiple papers that presented different methods of plagiarism inspection in music in terms of Sampling, Melody and Rhythm.
2. Generation of the dataset – Added a significant number of music file pairs as data points for the dataset.
3. Dataset Annotation – Added features like Genres and Region of Origin to expand the database.
4. Pattern Identification – Did analytical tasks like the normalization of data to build empirical correlations viz:
   (a) Genre Prominence in composer's plagiarism patterns
   (b) Timbre similarity and Melodic similarity

# 9. References


1. https://lawyerdrummer.com/2017/03/music-plagiarism-2/
2. https://link.springer.com/article/10.1007/s10618-022-00835-2
3. https://arxiv.org/pdf/2107.09889.pdf
4. https://norma.ncirl.ie/5202/1/rajeshramachandrannair.pdf
5. https://www.iosrjournals.org/iosr-jce/papers/Conf.17013-2017/Volume-1/3.%2008-12.pdf?id=7557
6. https://basicpitch.spotify.com/
7. https://www.chosic.com/music-genre-finder/
8. https://link.springer.com/article/10.1007/s10805-020-09360-7
9. https://www.itwofs.com/hindi-am.html
10. https://www.eurasip.org/Proceedings/Eusipco/Eusipco2012/Conference/papers/1569556475.pdf