Audio and Music Processing Lab

Report - Audio Content Based playlists

UPF Barcelona, SMC 2024-2025

Madhav Jaideep

Audio Analysis and feature extraction:

- Python script (AnalyzeAudio.py) to process and analyse a collection of tracks using Essentia library
- **Dataset :** MusAV – contains around 2,100 tracks of 30 seconds each with over 1404 genres
- **Analysis and feature extraction using essential:**
    - **Tempo:** Extracted using **RhythmExtractor2013** to get BPM
    - **Key:** Three profiles (temperley, krumhansl, and edma) extracted using KeyExtractor
    - **Loudness:** Integrated loudness calculated using the **LoudnessEBUR128** algorithm
    - **Danceability:** Using the **Danceability** feature to assess the danceability of the audio.
    - **Embeddings :** Discogs-Effnet, MusicCNN
    - **Music styles :** Classified using a model trained on Discogs-effnet (**genre_discogs400-discogs-effnet**)
    - **Voice/Instrumental:** Classification distinguishing between vocal and instrumental parts of the audio (**voice_instrumental-discogs-effnet-1**)
    - **Emotion (Arousal and Valence):** Analysis of arousal and valence using musicnn model **emomusic-msd-musicnn-2**
- The **analyze_collection** function processes a collection of mp3 files, analyses and stores the results in a JSON file, where each extracted feature is stored as a dictionary.
- Analysis results are saved periodically to avoid data loss.

Music collection Overview:

- Python notebook (CollectionAnalyze.ipynb) Includes plots and statistical reports about all the features and analysis.
- **Music styles:**
    - There are a total of 400 music styles extracted, which are categorized under their broad parent categories given. From the plot distribution, we can note that the **Rock** genre has the highest number of songs (more than 500 tracks), and then the others are lesser with electronic, hip hop and so on. The

collection seems to be diverse in terms of genres but with more inclinations towards rock and electronic music taking up almost half of the total collection.

- **Tempo and Danceability:**
    - The plot distribution for tempo shows that the music collection is diverse around a significant range of tempos from 60 to around 185 with a higher concentration of tracks around a tempo of 130 range. Overall, the collection varies widely regarding the tempo of the songs.
    - For Danceability, the output range is ideally within 0-3 (higher being more danceable), but the distribution shows there are some outliers as well. From the report, we see that a high number of tracks are around the 1-2 range, showing that most of the tracks have moderate to good energy and songs of lower energy and danceability is comparatively lesser in number.
- **Key/Scale Analysis:**
    - The key/Scale extraction was done for three different profiles (temperley, krumhansl, and edma). We can see that the scale estimation for all three are different, with temperly classifying mostly as major scales, krumhansl is slightly more balanced with more minor scale classifications and finally the edma profile has classified the major and minor scales well and is more balanced. For they key estimations, krumhansl and edma have very similar key classifications, Temperley varies from the other two more. Across all three profiles, we see that C major is the one that is classified best and has the highest percentage. From this analysis, overall it seems like the edma profile is best for being used in analysis and processing of key and scales.
- **Loudness analysis:**
    - The integrated loudness in LUFS standards currently differ based on for which platform the song is mastered for and usually ranges from around -14 to -10 for most streaming platforms. Rock and pop songs usually lie around -11 to -9 LUFS and electronic music usually pushes it to around even -6 or -5 LUFS sometimes. From the **plot distribution** we can see that the highest concentration of music is around -11 to -8 LUFS which makes sense as almost half of our collection consists of tracks that are rock or electronic, which are usually mastered to be loud. Similarly there is a number of tracks that go past -8 LUFS to around -6 and even -5 which could be due to the high number of electronic music in the collection.
- **Arousal/Valence (Music Emotion):**
    - The report shows that a high number of tracks are of 5-6 arousal and 5-6 valence, which usually means that they are more exciting and happy/uplifting. A lot of the songs are more energetic and in a joyful mood but is neither overly energetic or too calm.
- **Voice – Instrumental distribution:**

    - The results show that more than 1400 tracks are with vocals and around 600 tracks are instrumental. The collection is more inclined towards having songs with vocals on them.

Playlist generation

- Created two apps, one app for generating playlists using descriptor queries using extracted features. Second app is for generating playlists using track similarity (given a query track)
- **Playlist generation using descriptor queries:**
    - (app_descriptors.py, run using run1.sh)
    - The tempo range slider is set from 50 to 190 as most songs usually lie in that range
    - Edma profile was used for key/scale estimation as it gave the most balanced results from analysing them. The key strength threshold is set to a default of 0.7 as below that value, there are chances of misclassification of the songs.
    - The valence and arousal automatically takes the minimum and maximum ranges from the collection of both to be set in the slider, as it will be easier to identify and choose the values which we know are present in the dataset.
- **Playlist generation using similarity:**
    - (app_similarity.py, run using run2.sh)
    - From the generated playlists, personally I felt the Discogs-effnet model gave more accuracy and was better in finding tracks that were similar. MusicCNN performed good as well by generating the songs in the correct genre, but the energy and style of the songs differed in some cases. The Discogs model mostly identified the correct styles and type of songs and matched the energy and emotions of the query track given.

Conclusion:

Overall the system built is generally good in classifying and analysing a given collection of music, extracting various features. The playlist generation on descriptors and similarity is efficient, with Discogs-effnet proving to be more effective and accurate. While the system performs well, there's room for improvement maybe by using more low level feature extractions. In some cases, the styles had misclassifications and were mixed up. Similarly in classifying valence and arousal there were giving slightly more misclassifications than ideal, I noticed some high energy tracks that had unique elements being misclassified as a track with low valence and arousal. Danceability and tempo estimations were mostly accurate by giving the correct type of songs. Despite these areas for enhancement, the system provides a strong foundation for music analysis and playlist generation.