Audio Content Based Playlists Report - AMPLab

Sergio Cárdenas Gracia - Master's in Sound and Music Computing

Audio analysis with Essentia

The feature extraction process is designed to analyze all MP3 audio files within a specified folder, ensuring efficient processing by handling errors gracefully and resuming analysis without recomputing already processed tracks in case of interruptions. Extracted features are stored in a JSON file for further use.

The process begins with the initialization of all required models before processing starts, optimizing performance and avoiding redundant initializations. Each audio file is loaded only once using the AudioLoader. The loaded stereo audio is then converted to mono using MonoMixer, and the mono audio is resampled to 16 kHz using Resample. This ensures that the original stereo audio, the mono version, and the resampled mono audio are all available without requiring multiple file reads.

Once the audio has been preprocessed, required embeddings for each track are computed once and stored. These embeddings are reused across different classifier models, eliminating redundant computations. Classifier models operate on short audio chunks rather than entire tracks, generating multiple predictions for different chunks of the same track. The predictions are then averaged to produce a final classification result per track.

For each processed track, the results from classifiers are stored in a JSON file. The average embeddings across time are also stored, as they are necessary for computing cosine similarity in later stages. This structured storage ensures that previously processed tracks do not require reanalysis in case of process interruptions.

Music collection overview

The music collection spans across 400 possible music styles, as determined by the Discogs-Effnet model using the *discogs-effnet* embeddings. Due to the large number of styles, the focus is placed on the distribution of parent genres. As shown in *Figure 1*, rock is the most prominent genre, accounting for approximately a quarter of the dataset. Following rock, electronic music and hip hop represent the next largest portions, with several other genres being less represented in comparison.

The tempo distribution of the collection, as determined by the TempoCNN model, shows distinct peaks around 75 and 155 BPM, with most tracks falling between 60 and 130 BPM, as observed in *Figure 2*. Danceability, assessed using the Danceability classifier model based on *discogs-effnet* embeddings and shown in *Figure 3*, is predominantly close to 1 for the majority of the tracks, indicating high danceability. Tracks with danceability closer to 0 are less common, and intermediate values are barely represented.

Key analysis was conducted using the KeyExtractor algorithm, which utilized three profiles: ‘temperley’, ‘krumhansl’, and ‘edma’ (check this for reference). As shown in *Figure 4*, each profile produces different results, with ‘temperley’ identifying fewer minor keys. Given the collection's genre diversity and ‘temperley’ strong performance in identifying minor keys, it appears to be the most reliable choice. However, depending on the specific context or genre focus, ‘krumhansl’ or ‘edma’ could also be considered.

The integrated loudness of the collection, measured using the LoudnessEBUR128 algorithm and shown in *Figure 5*, primarily falls between -12 and -6 LUFS, which aligns with traditional music production standards. However, since modern streaming platforms tend to favor tracks with loudness levels between -14 and -16 LUFS, the collection's loudness values seem slightly outdated.

The emotional characteristics of the tracks are evaluated through valence and arousal values, which were predicted using a model trained on the emoMusic dataset with *msd-musicnn* embeddings. *Figure 6* demonstrates that the majority of tracks fall within a concentrated range of 4-7 for valence and 3-7 for arousal. This indicates that most of the tracks have moderate emotional valence (pleasantness) and arousal (energy).

Tracks in the collection are categorized as either instrumental or voice-based, using a model based on *discogs-effnet* embeddings. As shown in *Figure 7*, the collection includes 600 instrumental tracks and 1500 voice tracks, showing a strong emphasis on vocal music.

Overall, the collection demonstrates a rich diversity across multiple musical dimensions, such as genre, tempo, danceability, tonality, emotion, and vocal presence. While certain characteristics like rock, electronic music, and high danceability dominate the collection, there is a broad range of variation in tempos, key profiles, and emotional expression, which adds depth to the collection. However, the representation of different classes is not entirely balanced, with certain genres and characteristics being more prevalent than others.

Playlist generation

The queries app enables users to generate playlists by filtering and ranking tracks based on previously extracted descriptors, storing the results in an M3U8 playlist file. While the playlists generally meet expectations, some limitations are present, particularly in voice/instrumental classification and inconsistencies in tempo and danceability. These issues include occasional misclassifications, tempos appearing doubled or halved, and tracks not reflecting the expected danceability.

The similarities app enables users to select a query track from the collection and create two lists of 10 most similar tracks using *effnet-discogs* and *msd-musicnn* embeddings. After testing, the system produces generally good results. The choice between embeddings depends on personal preference; however, based on my experience, *effnet-discogs* embeddings appear more reliable, especially in terms of genre, likely due to their training and stronger alignment with metadata-based genre classification. This genre consistency enhances the coherence between the query track and the retrieved tracks.

Figures

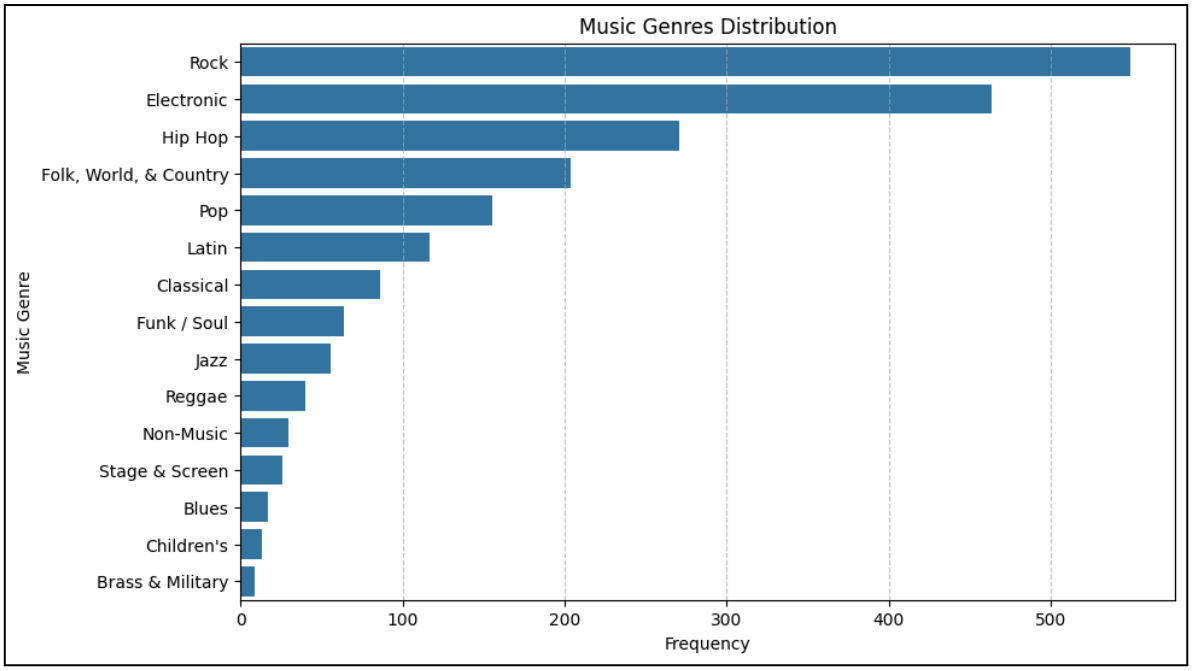

Figure 1: Music genres distribution plot

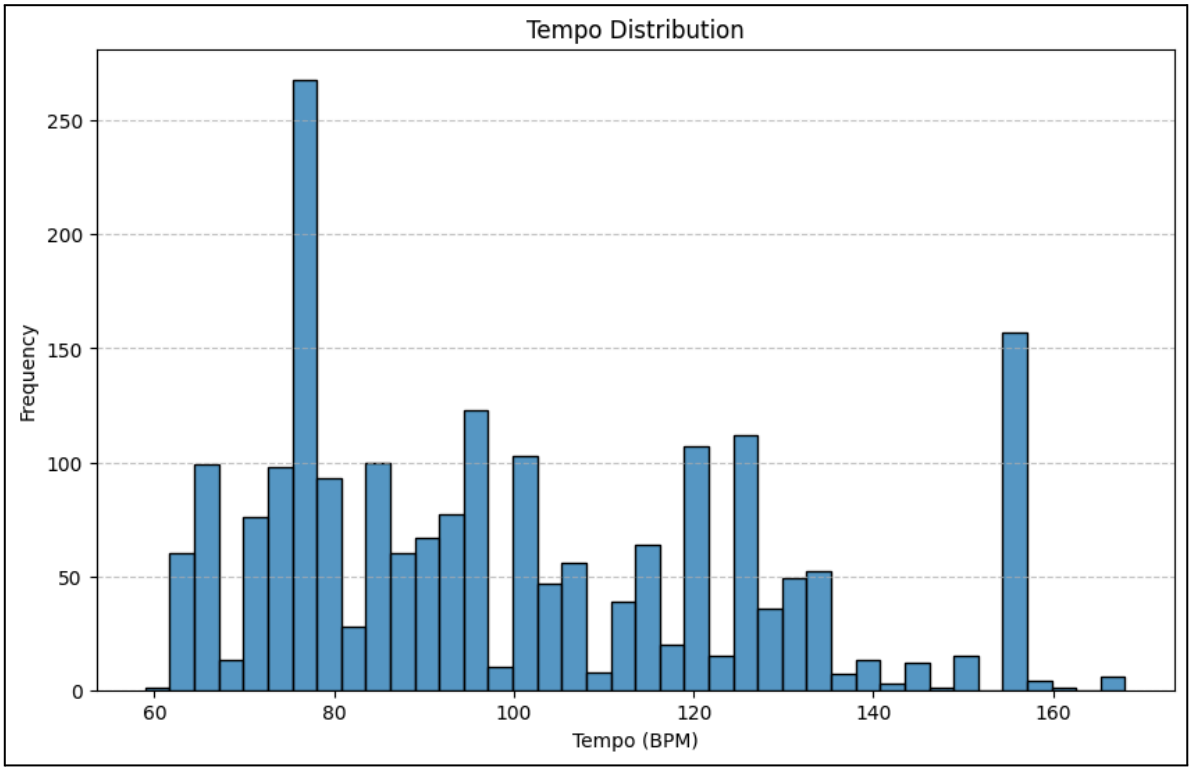

Figure 2: Tempo distribution plot

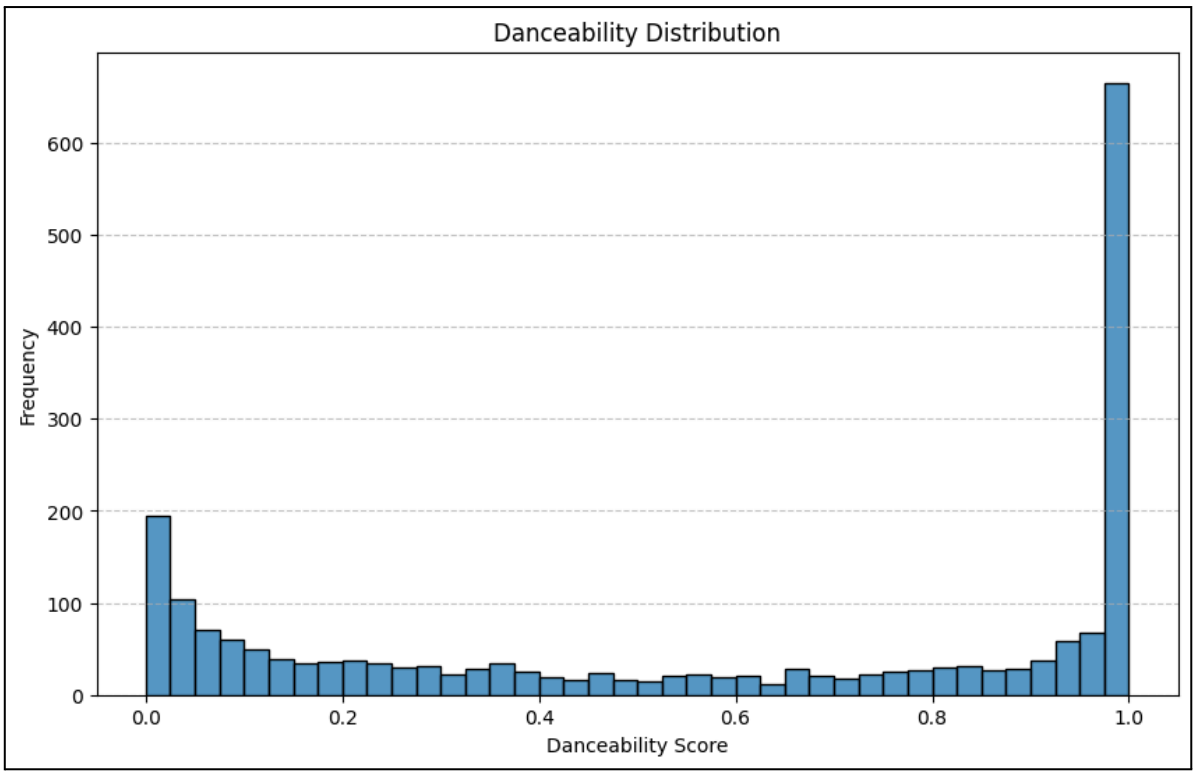

Figure 3: Danceability distribution plot

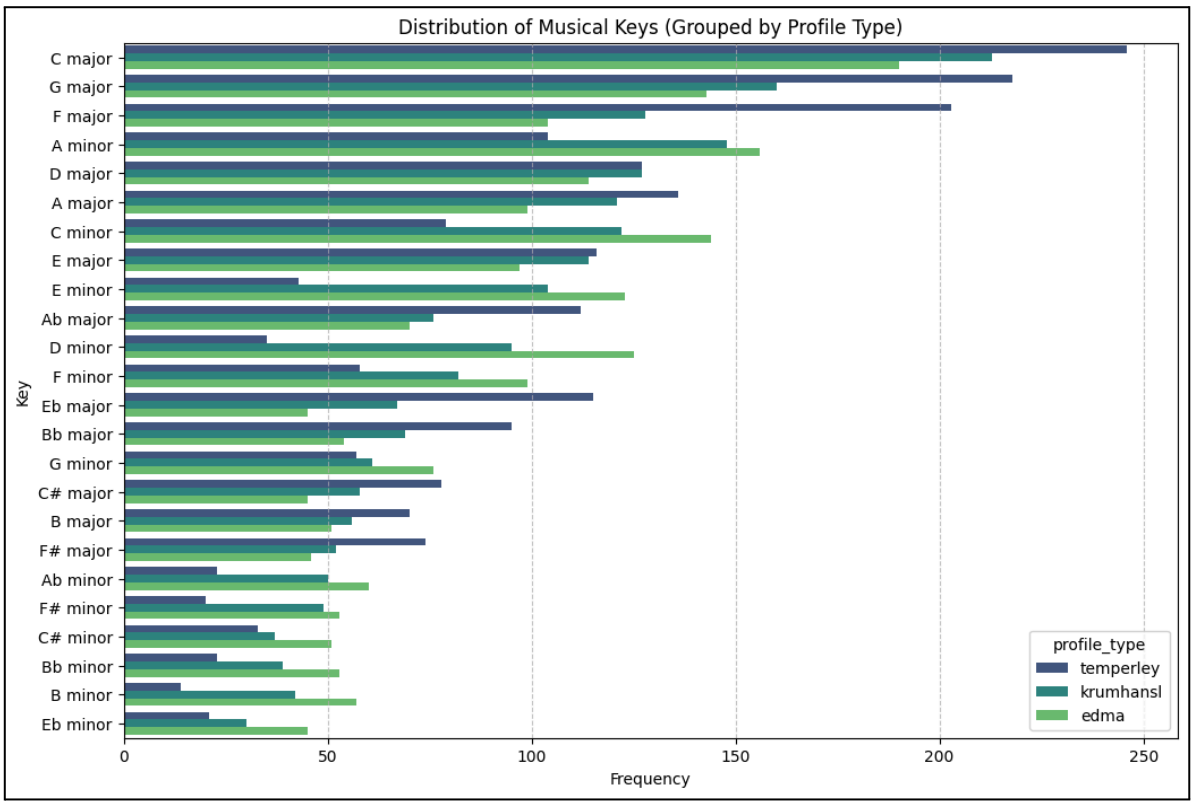

Figure 4: Musical keys distribution plot (grouped by profile type)

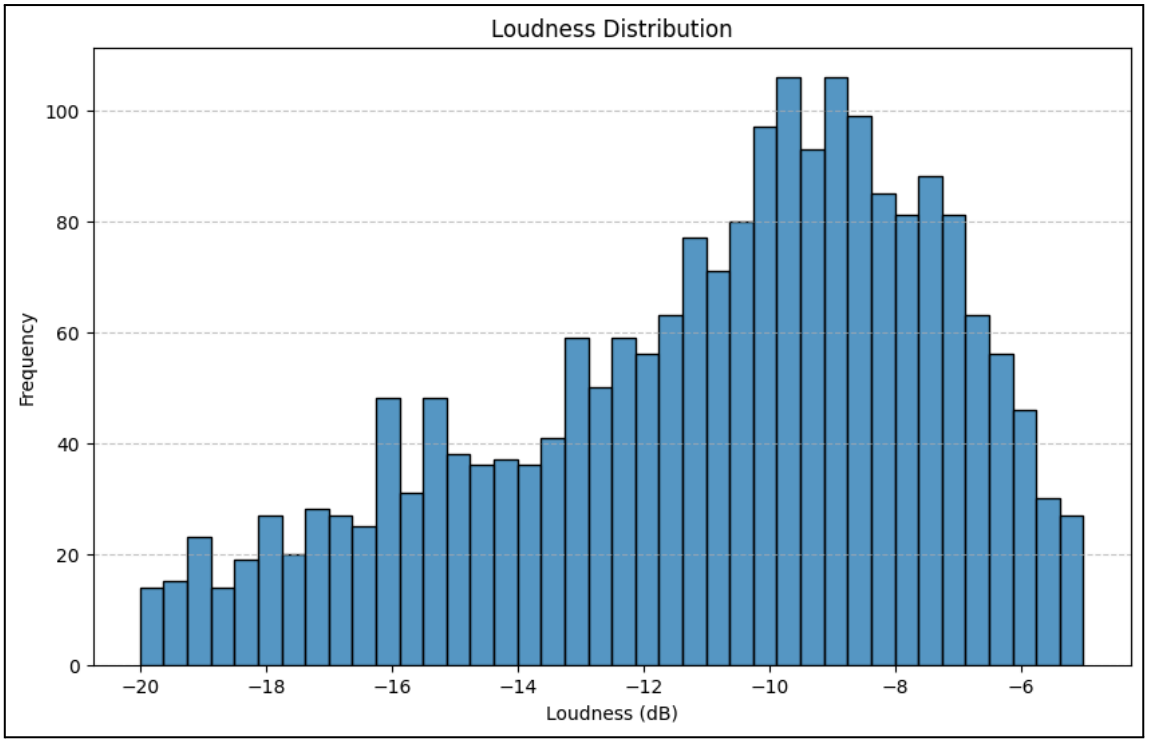

Figure 5: Loudness distribution plot

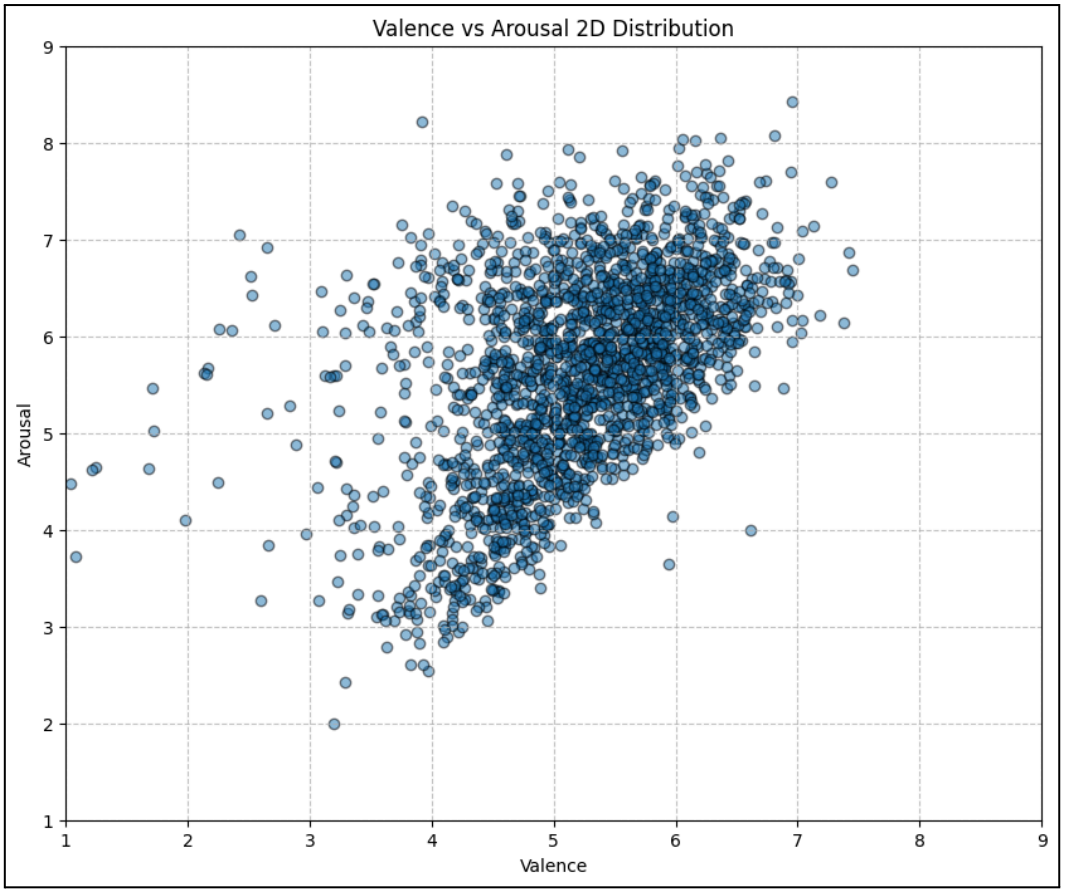

Figure 6: Valence vs arousal 2D distribution plot

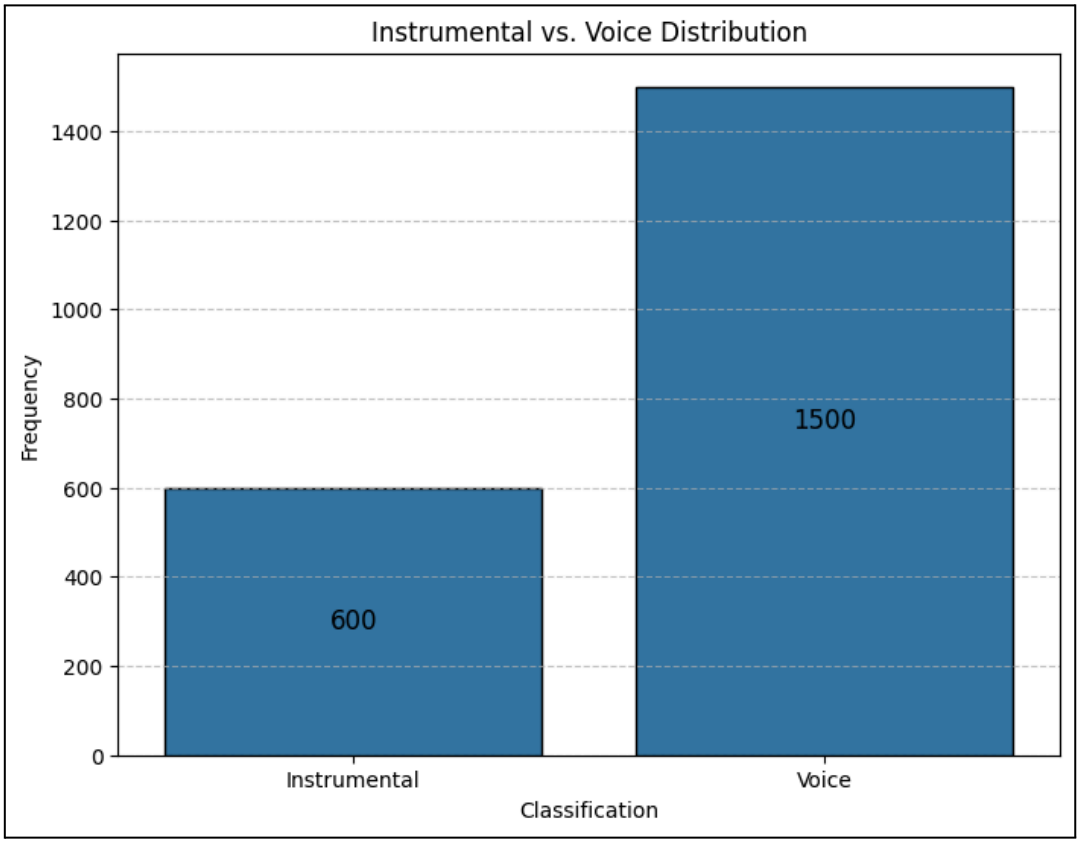

Figure 7: Instrumental vs voice distribution plot