Shahid Beheshti University
Data Science (2023-2024)

Emotion-Driven Song Recommendation System: Harmonizing Audio and Facial Expressions

Author 1
Department of Computer Science
Shahid Beheshti University
a.hassanaliaragh@cs.sbu.ac.ir
p.givehchian@cs.sbu.ac.ir

Abstract

In the realm of personalized content recommendations, our project presents an innovative approach that harmonizes audio and facial emotion analysis. Initially, we delve into audio processing to predict the emotional dimensions of songs, followed by facial emotion analysis to decipher live emotions from videos. Our ultimate goal is to seamlessly connect these two modalities, crafting a unique song recommendation system tailored to users' facial expressions.

1 Introduction

Our project blends music and facial expressions for a personalized user experience. Using the DEAM dataset with 1802 audio files, we decode emotions in songs and faces. We analyze audio features and use machine learning and deep learning for valence and arousal prediction. Simultaneously, we dive into facial emotion analysis, extracting features with computer vision algorithms.

Our innovation lies in fusing audio and visual emotion analysis, connecting emotions in both domains. The result is a recommendation engine syncing a user's facial expressions with a fitting song from the DEAM dataset. Our goal is a unique and immersive journey at the crossroads of music, emotions, and tech, offering users a truly emotive and personalized musical experience.

2 Related work/Background

Emotions can be influenced by such attributes as tempo, timbre, harmony, and loudness (to name only a few), and much prior work in Music-IR has been directed towards the development of informative acoustic features.

Although some research has focused on searching for the most informative features for emotion classification, no dominant single feature has emerged. In searching for the most informative emotion and expressive features to extract from audio, Mion and De Poli investigated a system for feature selection and demonstrated it on an initial set of single-dimensional features, including intensity and spectral shape as well as several music-theoretic features.

Their system used sequential feature selection (SFS), followed by principal component analysis (PCA) on the subset to identify and remove redundant feature dimensions. The focus of their research, however, was monophonic instrument classification across nine classes spanning emotion and expression, as opposed to musical mixtures. Of the 17 tested features the most informative overall were found to be roughness, notes per second, attack time, and peak sound level.

MacDorman et al. examined the ability of multiple acoustic features (sonogram, spectral histogram, periodicity histogram, fluctuation pattern, and mel-frequency cepstral coefficients–MFCCs) to predict pleasure and arousal ratings of music excerpts. They found all of these features to be better at predicting arousal than pleasure, and the best prediction results were when all five features were used together.

Schmidt et al. investigated the use of multiple acoustic feature domains for music clips both in terms of individual performance as well as in combination in a feature fusion system. Their feature collection consisted of MFCCs, chroma, statistical spectrum descriptors, and octave-based spectral contrast.

The highest-performing individual features were spectral contrast and MFCCs, but again the best overall results were achieved using combinations of features.

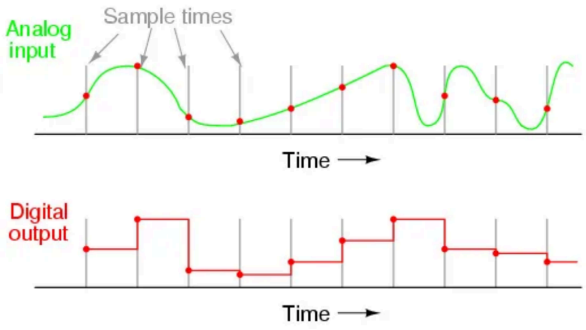
3 Proposed method

Music emotion recognition (MER)

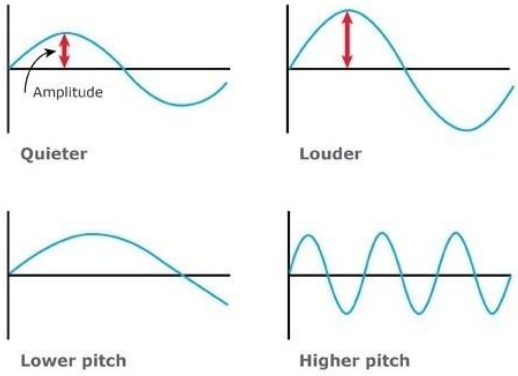
At first, we need to have a general understanding of what a sound wave is. A signal is a variation in a certain quantity over time. For audio, the quantity that varies is air pressure. Vibrations in the air carry the energy. But how can we digitize that signal? We have to sample some of these air compressions and decompressions, using equipment like a microphone.

Based on the Nyquist-sampling theorem, the number of samples should be at least twice the maximum frequency we want to capture. (F > 2*f)

Since the human ear can approximately only hear from 20Hz up to 20000Hz, a sampling rate of 44.1kHz is more common.

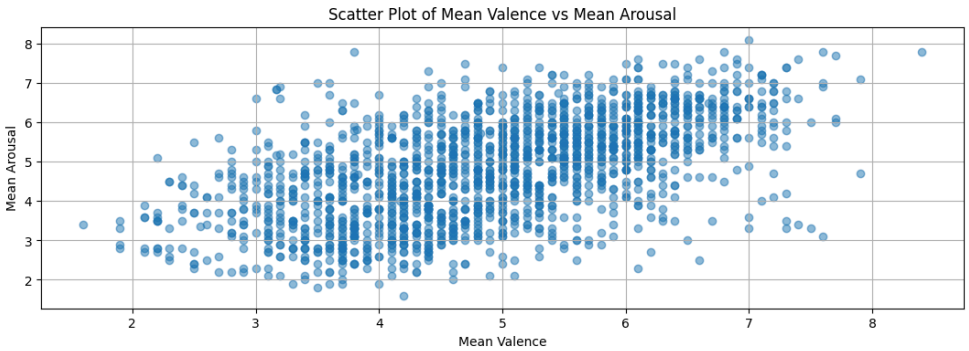


A sound wave consists of a frequency and an amplitude. The amplitude is correlated with the loudness, and the frequency is correlated with the pitch.



Analyzing emotions from pitch, frequency, and amplitude in audio involves understanding how these features contribute to the overall emotional tone of music. The pitch, whether high or low, can convey feelings of excitement or calmness. Frequency distribution, spanning from low bass to high treble, influences the emotional depth and brightness of a piece. Amplitude, indicating loudness or softness, contributes to the intensity or subtlety perceived in the music. By examining changes and patterns in these features, one can discern the emotional dynamics of a song. Integrating machine learning techniques and training models on annotated emotional data enables the extraction of emotional cues from audio signals.

Our dataset has 4 target values for each song, the mean_valence, mean_arousal, std_arousal, and std_valence.



Valence refers to the pleasantness or unpleasantness of an emotion. Emotions can be classified on a scale from positive to negative valence. For example, joy and love are considered positive valence emotions, while anger and sadness are negative valence emotions.

Arousal refers to the level of physiological activation or intensity associated with an emotion. Emotions can range from low arousal (calm or relaxed) to high arousal (excited or agitated). For instance, calmness and contentment are low-arousal emotions, while fear and excitement are high-arousal emotions.

To enhance the interpretability of existing labels, we leverage statistical measures such as mean (μ) and standard deviation (σ) to establish ranges for both valence and arousal, assuming a normal distribution. This process involves creating three versions of the range: one standard deviation (1σ), two standard deviations (2σ), and three standard deviations (3σ), encompassing 68%, 95%, and 99% of the data, respectively (refer to Figure 1).

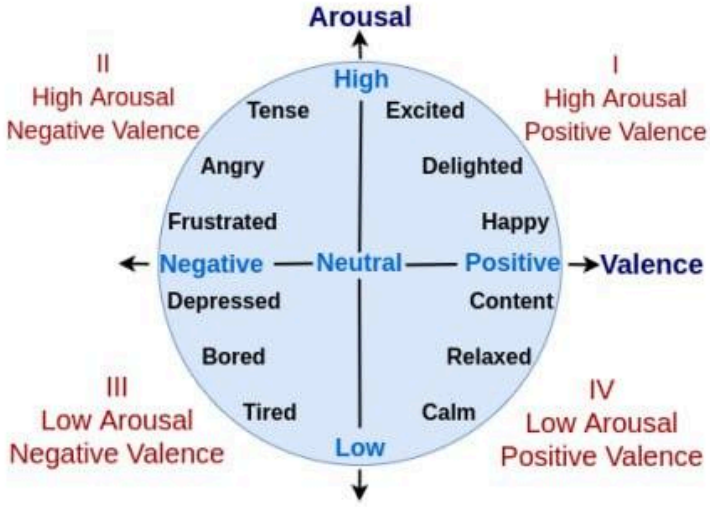
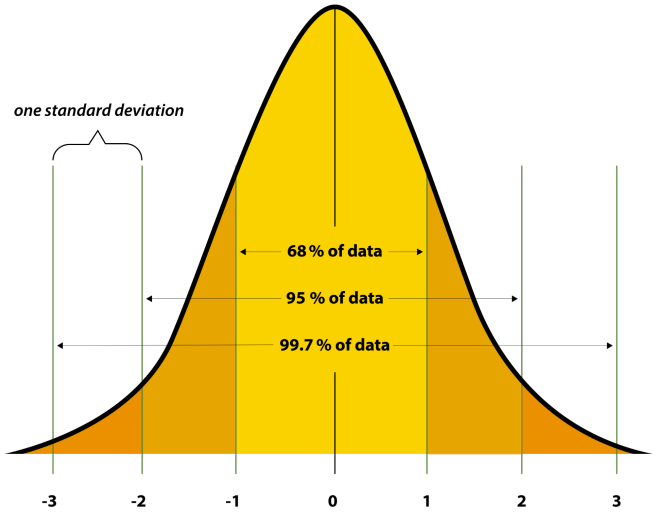
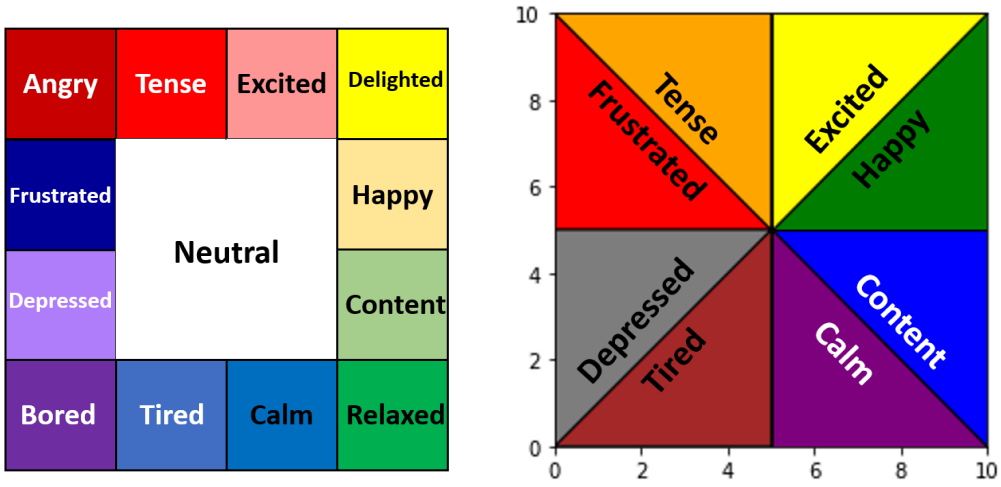


Fig. 1: The valence-arousal space. Valence denotes the range of emotions from being very sad (negative) to very happy (positive) and arousal reflects the energy or intensity of emotions from very passive to very active.

In each version, we expand the features to include minimum and maximum values for both valence and arousal—represented as min_valence, max_valence, min_arousal, and max_arousal. These four points collectively define an area within the entire spectrum of potential values, creating a 10x10 square in this context. Our goal is to determine which emotions are encompassed by this area within the valence-arousal space, as illustrated in Figure 7.



We approach the allocation of emotions to the 10x10 square from two perspectives. In the first perspective, we partition the square into 16 smaller squares by introducing four horizontal and four vertical lines. The central four mini-squares are assigned to the neutral class, while the surrounding squares follow the order depicted in Figure 8.



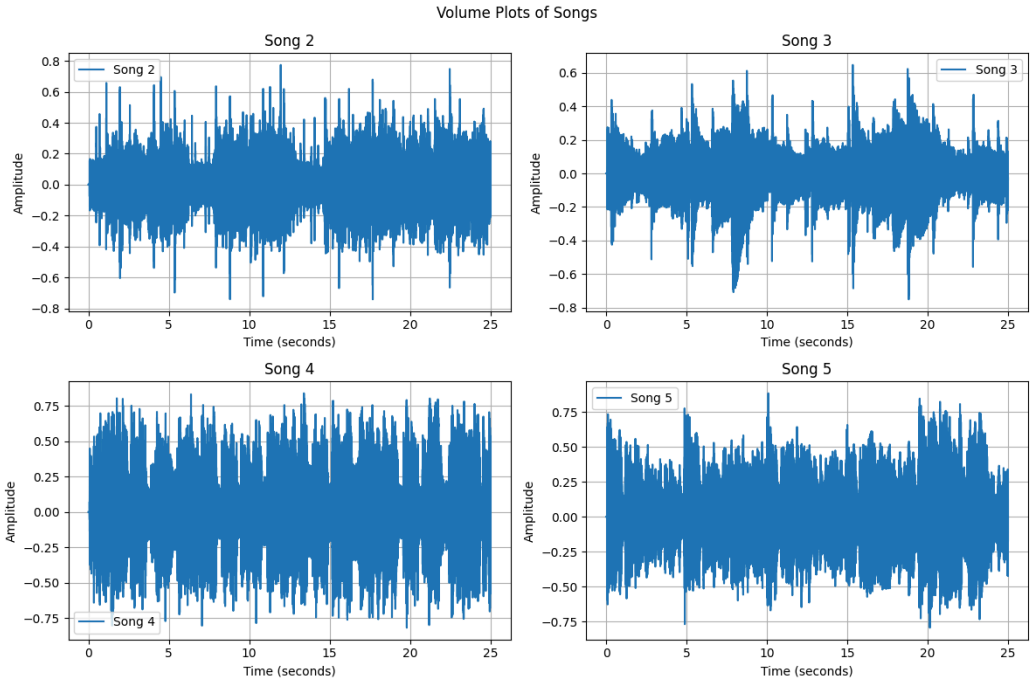
Alternatively, the second perspective involves dividing the main square into eight sections using one vertical line, one horizontal line, and two diagonals.

To attribute emotions to each song, we undertook the task of quantifying the contribution of specific segments within the arousal and valence ranges. By determining the percentage of the entire area covered by these segments, we assigned corresponding percentages to each emotion, resulting in a vector of emotions with scores ranging from zero to 100. The cumulative sum of all emotion scores equals 100.

To facilitate the application of classification models, we converted this emotion vector into a binary format using various methods. A straightforward approach involves designating the emotion with the highest score as one and the rest as zeros. For a more nuanced labeling system, we introduced a threshold, converting scores above it to one. We set the threshold at 100 divided by the length of the vector for this purpose.

A volume plot, often referred to as a loudness plot or amplitude plot, is a graphical representation that illustrates the variations in sound intensity or volume over time in an audio signal. It typically plots the amplitude of the audio signal on the y-axis against the time on the x-axis.

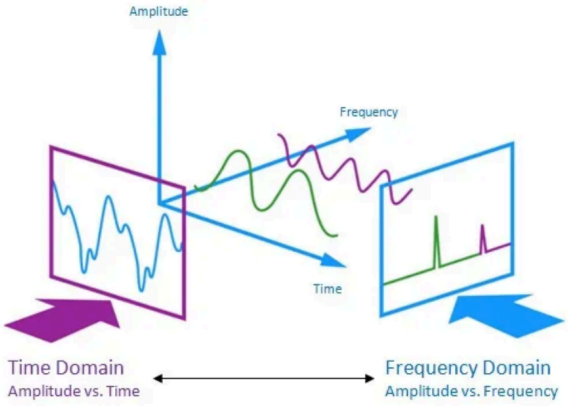
In simpler terms, the volume plot gives a visual depiction of how loud or soft the audio is at different points throughout its duration. Peaks and valleys on the plot indicate changes in volume, providing insights into the dynamic range and intensity of the audio.



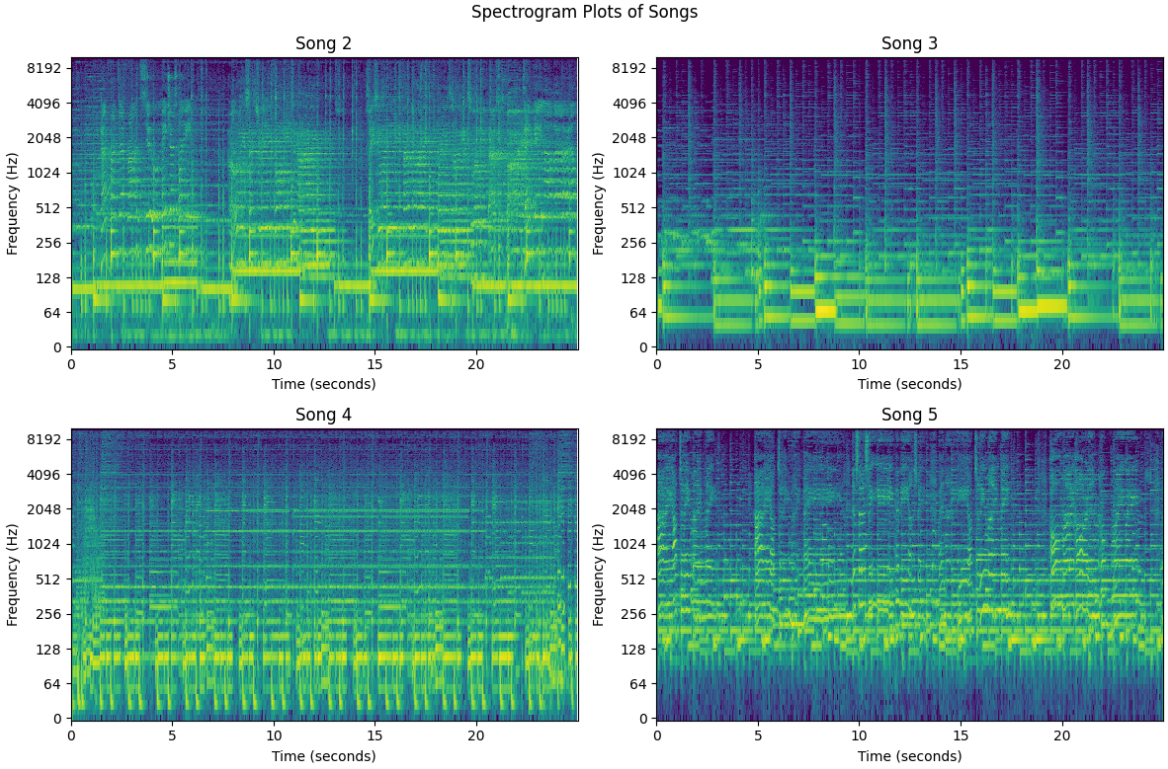
Volume plots, depicting amplitude variations over time, offer a basic representation of audio loudness but may fall short in capturing the intricate features crucial for emotion classification. Emotions in speech or music are often conveyed through nuanced temporal dynamics, intricate frequency content, and specific spectral features that volume plots alone cannot adequately portray. Complex patterns, such as pitch variations, rhythmic structures, and the distribution of energy across different frequency bands, are essential for understanding emotional expression.

Spectrograms are preferred over simple volume plots for tasks like emotion classification due to their ability to provide a comprehensive two-dimensional representation of audio signals, capturing both frequency and temporal dynamics. They are visual representations of the frequency content of an audio signal over time.

Spectrograms work by applying Fourier transformations to audio signals. Fourier transformations decompose a time-domain signal into its constituent frequency components, revealing the spectrum of frequencies present. In the context of spectrograms, Short-Time Fourier Transform (STFT) is commonly used. It involves dividing the signal into short, overlapping time windows, and then performing Fourier transformations on each window to obtain frequency information.



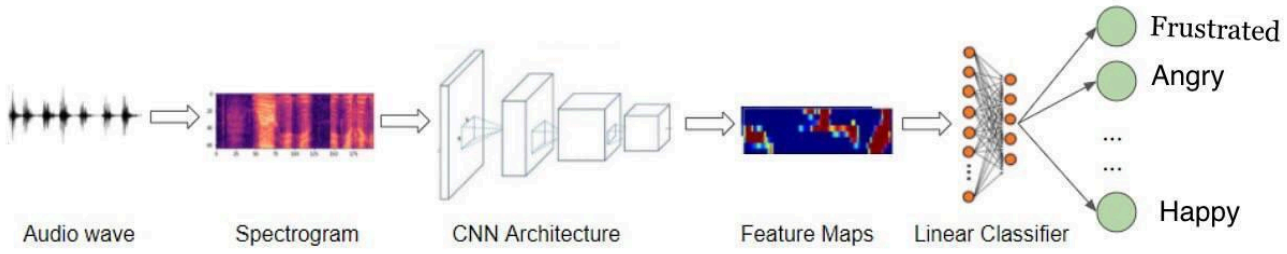
The result is a series of spectra over time, forming a spectrogram. They provide a detailed, two-dimensional view, where the x-axis represents time, the y-axis represents frequency, and the color intensity indicates the amplitude or energy of each frequency component.



Unlike volume plots, spectrograms offer insights into the distribution of energy across different frequency bands, enabling the model to discern tonal characteristics, pitch variations, and rhythmic patterns associated with emotional expression. Spectrograms are compatible with Convolutional Neural Networks (CNNs), allowing these models to automatically learn hierarchical representations of features in both time and frequency domains. This richer information, extracted from spectrograms, enhances the model's capability to discern complex

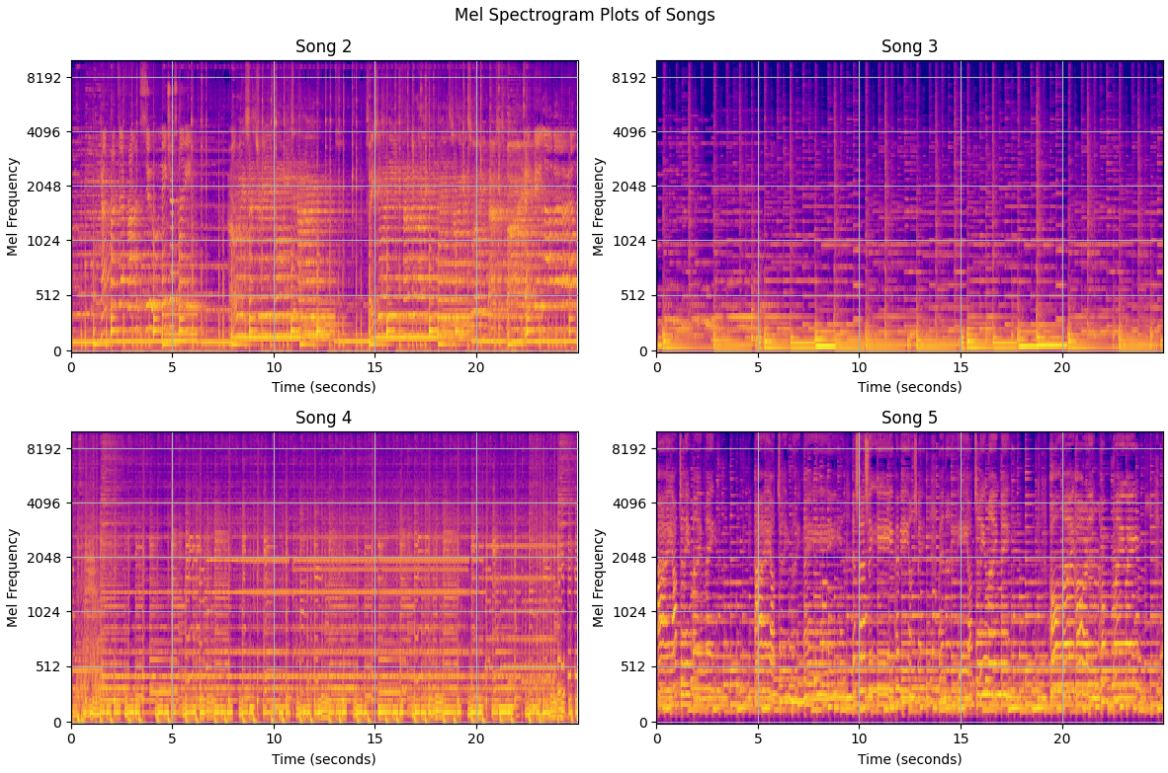
patterns and improves its robustness to background noise, making it a more effective tool for analyzing and classifying emotions in speech or music.

This Spectrogram is given to a CNN, to find the hidden patterns in the deeper dimensions.



This was done in 2 different ways, one was saving the plots as PNG files and giving those RGBA images with 4 channels to the CNN, and the other was getting the spectrograms data and using that as input.
We tested both of these methods, but since the results didn't differ much, and using the data directly instead of the images was much more computationally efficient, we did it the second way.

The issue with the spectrogram was that it didn't perceive the frequencies the same as the human ear. Using a Mel spectrogram, derived from the Mel-frequency cepstral coefficients (MFCCs), is advantageous over a standard spectrogram in audio signal processing due to its alignment with human auditory perception. The Mel scale, mimicking the non-linear response of the human ear to different frequencies, emphasizes features that are more relevant to how we perceive sound. This perceptual relevance aids in capturing acoustic characteristics more effectively, particularly in tasks like speech and music analysis. By compressing lower frequencies and expanding higher frequencies, Mel spectrograms enhance discrimination in tonal characteristics and contribute to a more compact representation.



We gave this mel spectrogram as input to the model and gained better results since the model could perceive the sounds the way the human ear does.

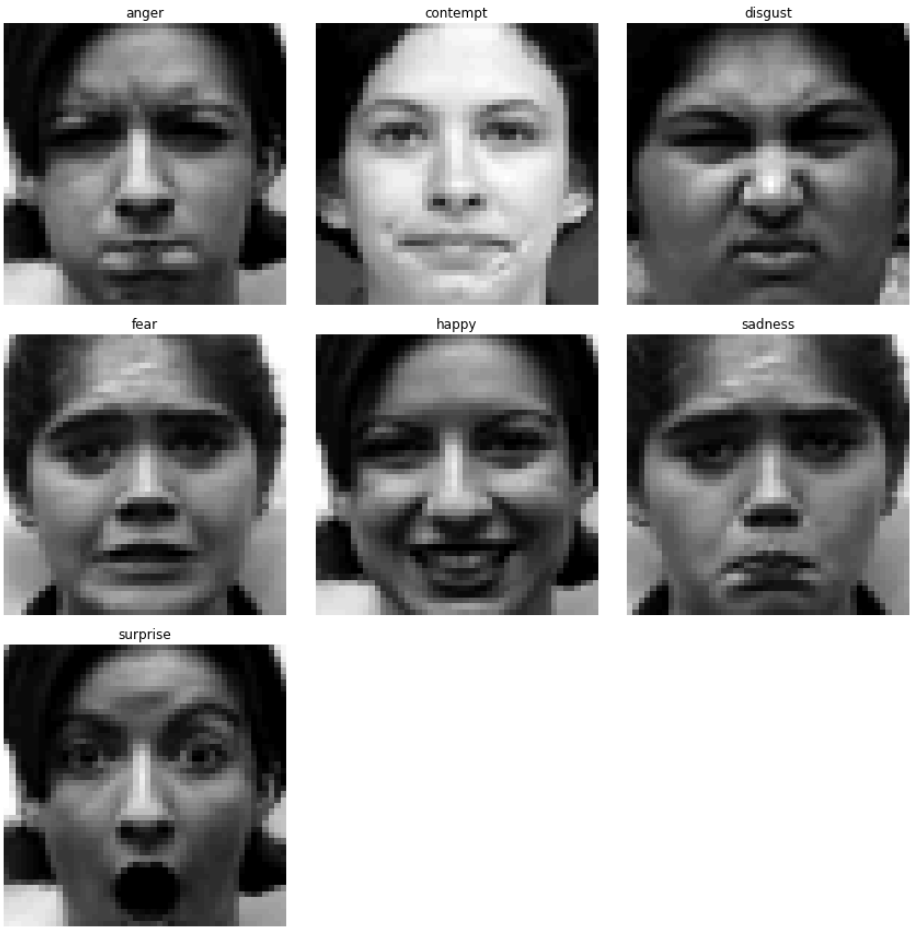
We also tried a Convolutional Neural Network (CNN) coupled with a Bidirectional Long Short-Term Memory (LSTM) layer, tailored for song emotion classification. The initial CNN layers, comprising two convolutional and max-pooling stages, are designed to hierarchically extract spatial features from Mel spectrograms, providing a compact representation of audio signals. The subsequent Bidirectional LSTM layer adds a temporal dimension, capturing sequential dependencies crucial for understanding the emotional dynamics of songs. The 'ReLU' activation functions introduce non-linearity throughout the model. After reshaping and flattening the output, two fully connected dense layers follow, with the first acting as a feature extractor and the second employing a linear activation, indicating a regression setup for predicting continuous values representing the percentages of each of the 8 emotion categories. This architecture combines spatial and temporal processing, making it well-suited for song emotion classification by capturing both local and sequential patterns inherent in audio data.
The problem was that this model was too complex for the problem would overfit our training data, and was also very computationally expensive.

Face emotion recognition (FER)

In order to extract emotion from the face, we used one of the main datasets for this purpose, the Cohn-Kanade Dataset (CK+) which contains 920 individual facial expressions.
Here is a brief explanation of the dataset:
It contains adapted data up to 920 images from 920 original CK+ datasets
Data is already reshaped to 48x48 pixels, in grayscale format and face cropped using haarcascade_frontalface_default.
Noisy (based on room light/hair format/skin color) images were adapted to be clearly identified using the Haar classifier.
Columns from file are defined as emotion/pixels/Usage

Emotions are defined as determined index below:

- **0** : Anger (45 samples)
- **1** : Disgust (59 samples)
- **2** : Fear (25 samples)
- **3** : Happiness (69 samples)
- **4** : Sadness (28 samples)
- **5** : Surprise (83 samples)
- **6** : Neutral (593 samples)
- **7** : Contempt (18 samples)



We used CNN to do the prediction on this task

Model Architecture

The CNN model was constructed using the Keras Sequential API. It consisted of three convolutional layers with max-pooling layers in between, facilitating the extraction of hierarchical features from the facial images. A Flatten layer was employed to transform the 2D feature maps into a 1D array. Subsequently, dense layers with varying numbers of neurons and ReLU activation functions were added to further extract high-level features. The output layer comprised seven neurons with softmax activation, aligning with the seven emotion classes in the dataset.

Model Compilation and Training

The model was compiled using the Adam optimizer, categorical cross-entropy loss function, and accuracy as the evaluation metric. The training was performed for 25 epochs with a batch size of 32, and the dataset was split into training and validation sets to monitor the model's performance. The training process involved adjusting the weights and biases of the neural network to minimize the loss function.

In order to use it in real-time applications we need real-time face detection to extract the image in each frame and use the model we trained to predict emotion in each frame the use of the OpenCV library with the Haar Cascade classifier significantly simplifies this process. initializing a CascadeClassifier object with the pre-trained Haar Cascade classifier for frontal face detection. This classifier, 'haarcascade_frontalface_default.xml,' contains features that are effective in identifying facial features based on patterns, contrasts, and edges. Once the classifier is set up, it can be applied to images or video frames to locate and outline frontal faces within the provided data. The versatility and efficiency of Haar Cascade classifiers make them widely employed in various

applications, such as facial recognition, emotion detection, and overall human-computer interaction.

Recommendation System based on facial expressions

In the third and final part we should connect FER and MER, the final application of this purpose is designed in this order:

Recommendation based on facial expressions involves two main phases: observation and recommendation.

Observation Phase:
During a 15-second observation, each emotion expressed by the user is quantified on a scale of 0 to 15, corresponding to the seconds spent on that emotion.

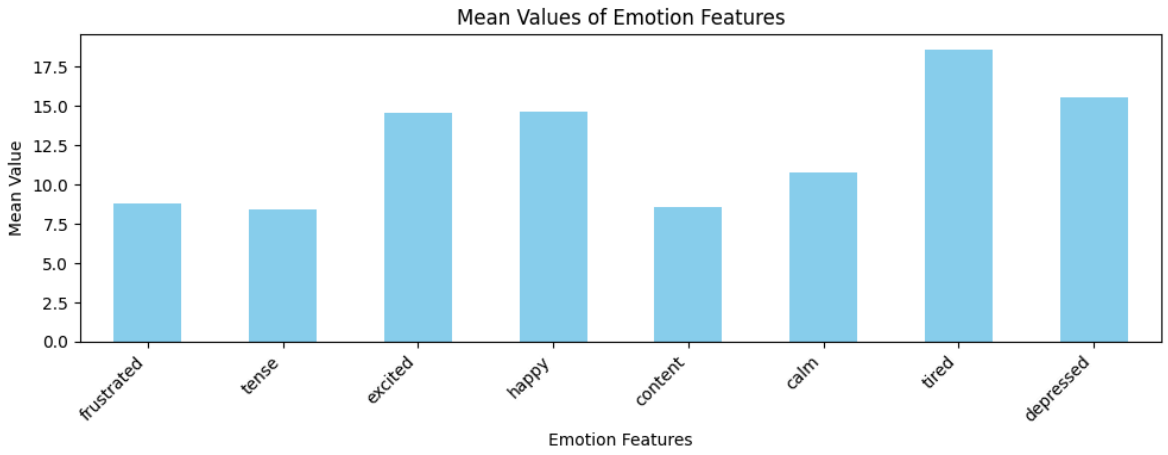
Recommendation Phase:
Utilizing cosine similarity, the distance between the vector of facial expressions and the emotions associated with songs is calculated. The top 5 similar songs are then recommended, with the first song automatically played.

4 Results

The dataset we used was the DEAM dataset - The MediaEval Database for Emotional Analysis of Music. DEAM dataset consists of 1802 excerpts and full songs annotated with valence and arousal values both continuously (per second) and over the whole song.

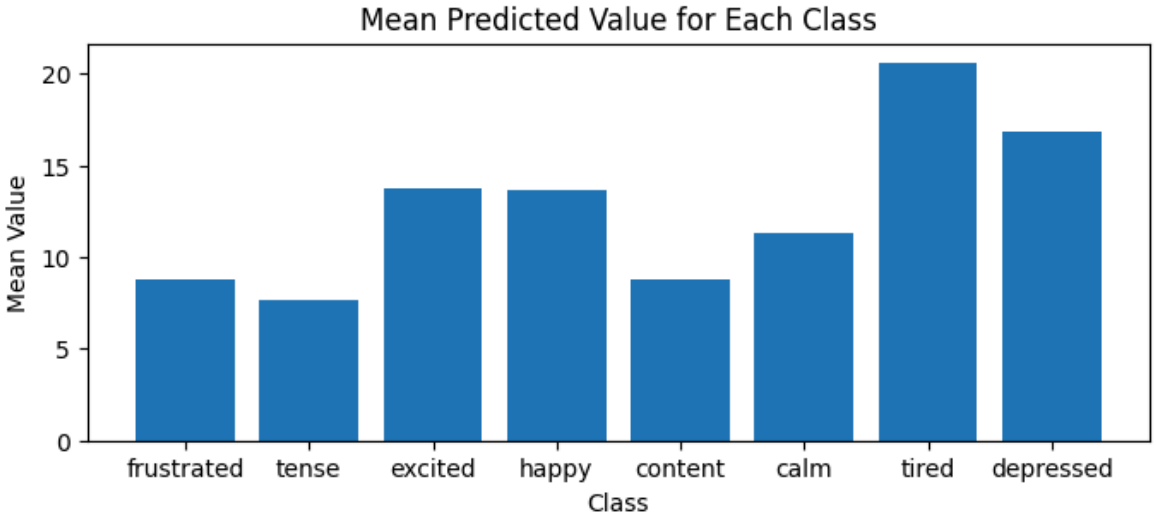
We only worked with the std and mean arousal and valence throughout the whole song.

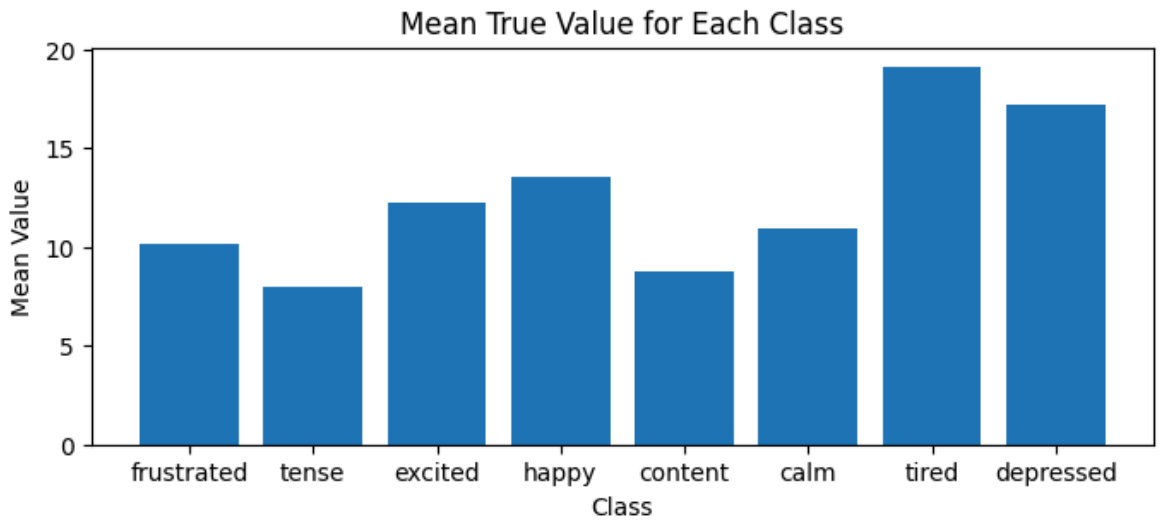
In stage 1, the task was a regression task, where we tried to predict the percentage of each the 8 emotions in the songs.



We can see that in our data, tired has the highest percentage. Now using a CNN and the spectrogram as input, with 3 convolutional layers, a linear activation function, and the Adam optimizer, and MSE as our loss function we can reach a Mean Average Error (MAE) of 10.34 on the testing sample.
But then we try this using the mel-spectrogram, which reaches an MAE of 11.08 without scaling, 10.86 with a simpler CNN and only 2 convolutional layers, and 9.36 after scaling.
In general, the simpler CNN had better results.
The combination of CNN and RNN only reached an MAE of 9.76, so the simpler CNN model seems to be working better.

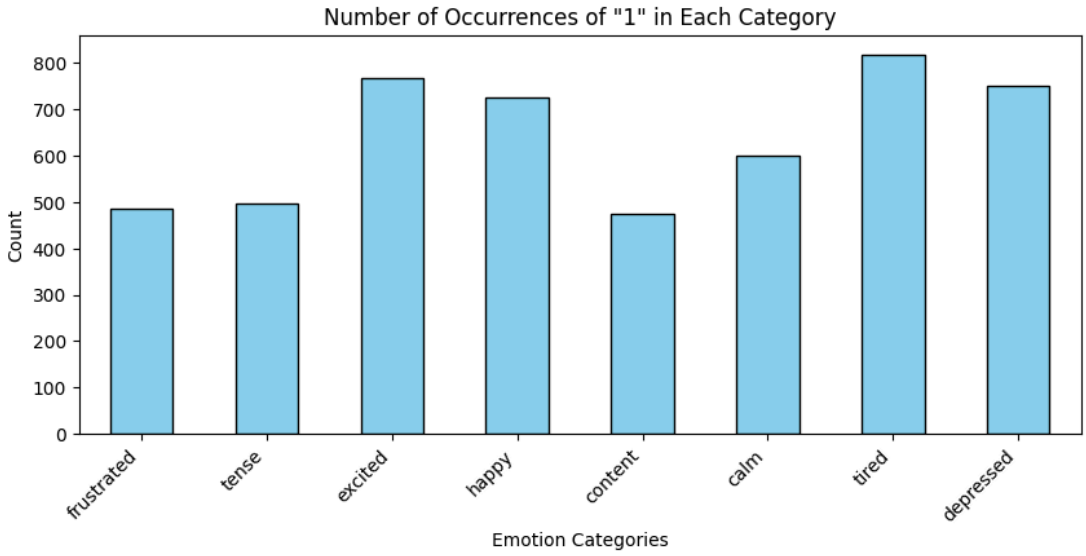
Results for the simple CNN with MAE of 9.36 (testing samples):



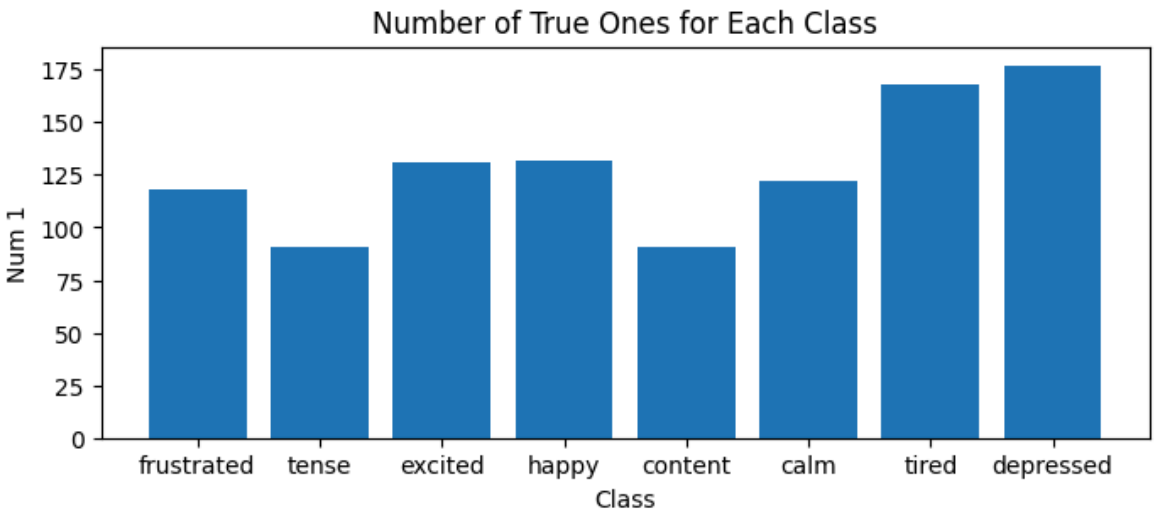


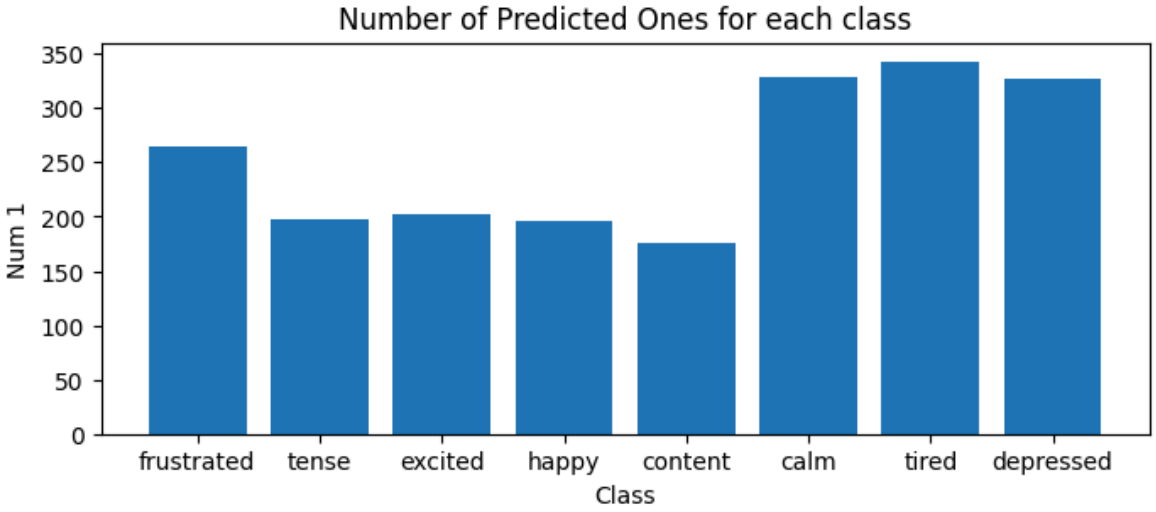
We can see that the mean predicted and the mean true values are very close.

In the next stage we are going to treat the emotion percentages as categorical values, so if an emotion has more than 13 percent, it is labeled as 1, and else as 0.
This is a multi-label classification task, since each song can have more than one emotion.



We used a CNN with 2 convolutional layers, a sigmoid activation function, and binary cross-entropy as the loss, but the model would overfit our training data.
so we added 3 Dropout layers and were able to reach a loss of 0.56 on our testing set.
This model returned an array with a probability for each class. We found the best threshold that gave us the best F1 score to be 0.4, so all probabilities above 0.4 were classified as 1, and the rest were 0.





Again we can see that the distributions of the count of true ones and predicted ones are very close. The final F11-score is 0.59, and tired has the highest recall, and excited the highest precision. Here is the classification Report of our model:

	precision	recall	f1-score	support
frustrated	0.34	0.76	0.47	118
tense	0.39	0.85	0.53	91
excited	0.58	0.90	0.71	131
happy	0.57	0.84	0.68	132
content	0.36	0.70	0.48	91
calm	0.34	0.91	0.49	122
tired	0.48	0.98	0.64	168
depressed	0.51	0.94	0.66	177
micro avg	0.44	0.87	0.59	1030
macro avg	0.45	0.86	0.58	1030
weighted avg	0.46	0.87	0.60	1030
samples avg	0.49	0.88	0.60	1030

5 Discussion

We use cnn for both MER and FER and these are some of the general advantages

Hierarchical Feature Extraction:

CNNs excel in capturing hierarchical features, making them ideal for tasks like MER where intricate patterns in frequency, pitch, and amplitude contribute to emotional content.
The convolutional layers automatically learn relevant spatial features within the mel spectrograms, providing a comprehensive understanding of audio signals.

Spatial and Temporal Processing:

The architecture combines spatial and temporal processing efficiently, allowing the model to capture both local features and sequential dependencies inherent in music.
This dual processing capability enables CNNs to discern complex patterns, such as tonal characteristics, pitch variations, and rhythmic structures, crucial for emotion classification.

In the task MER we used spectrograms as input for our model, below are some reasons

Compatibility with Spectrograms:

Spectrograms, representing the frequency and temporal dynamics of audio signals, align well with CNNs.

Robustness to Background Noise:

CNNs enhance robustness to background noise, a common challenge in audio processing, by automatically learning features that are relevant to emotion classification.
The model's ability to extract complex patterns from spectrograms contributes to improved performance even in the presence of noise.

Effective in Image-Based Input:

The transformation of mel spectrograms into image-like representations allows CNNs to leverage their success in image classification tasks.
CNNs can effectively identify patterns in these image-like representations, translating well to the task of recognizing emotional cues in music.

The combination of mel spectrograms and CNNs provides a comprehensive representation of audio signals, capturing both frequency and temporal dynamics.
This comprehensive representation enhances the model's ability to understand the emotional nuances present in music.

Mel Spectrogram Advantage:

The use of Mel spectrograms, derived from the Mel-frequency cepstral coefficients (MFCCs), aligns with human auditory perception, enhancing discrimination in tonal characteristics.
This perceptual relevance contributes to a more compact representation, allowing the model to perceive sounds similar to the human ear.

The advantages of mel-spectrogram in comparison to spectrum, spectogram and volume plots were explained above.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.