From Acoustic to Emotions: Audio Emotion Visualizer

Anishka Sachdeva (2018101112)
Harshika Jain (2018101115)
Subodh Sondkar (2018101064)

Aim

- How different kinds of emotions (happiness, sadness, anger, etc.) can be extracted/predicted from a piece of music using its acoustic features. The project aims to build a machine learning model to predict perceived emotions based upon a set of features extracted from the music/audio.
- This project is motivated and driven by an already existing study, and the end goal is to build the audio emotion visualizer in Python. The visualizer will help understand the emotional trajectory of any musical composition.

Background

- Content-based prediction of musical emotions and moods has a large number of exciting applications in Music Information Retrieval.
- The two-dimensional circumplex model proposes that all affective states arise from two independent neurophysiological systems: valence (a pleasure-displeasure continuum) and activity (activation-deactivation) [1].
- The two-dimensional models have been criticized for their lack of differentiation when it comes to emotions that are close neighbours in the valence-activation space, such as anger and fear [2].
- 3-dimensional variant containing valence, energy arousal and tension arousal has given better empirical results [3].

Motivation

Because it is important to understand how content-based prediction of musical emotions and moods can be used to build exciting and robust applications in Music Information Retrieval. The music itself expresses emotions, which can be highly subjective and difficult to quantify. Automatic recognition of emotions (or mood) in music. is still in its early stages, though it has gained attention recently. Various features, such as harmony, timbre, interpretation, and lyrics, affect emotion, and the mood of a piece may also change over its entire duration. But in developing robust automated systems to organize music in terms of emotional content, we face issues that often lack well-defined solutions; there may be considerable disagreement in the interpretation and perception of the emotions of a song. There might also exist ambiguity within the musical piece itself concerning the emotional content. Therefore, the project will help understand the emotional trajectory of any musical composition. It will also help understand where acoustic features might lack in emotion recognition and give direction to future prospects. It will help us appreciate the fact that emotion is not entirely encapsulated within the audio alone (for example, social context or setting etc., also plays an important role) and how incorporating music metadata, such as tags and lyrics, can be helpful.

Work Done

Dataset

In their work, both discrete and dimensional models of emotions were simultaneously investigated t clarify their mutual relationship and applicability to music and emotions. The three-dimensional model i used to collect data regarding the dimensional approach as it encompasses both lower-dimensiona models. A selection of film soundtracks was used to obtain a large sample of unknown yet emotionall stimulating musical examples. Soundtracks are composed to convey powerful emotional cues and ma serve as a relatively 'neutral' musical material in terms of music preferences and familiarity. A three-par selection process was used. Initially, 12 experts chose 360 excerpts representing happy, sad, tender, scary, and angry emotions and different quadrants in the 3D affect space. The expert panel consisting of musi students with extensive musical backgrounds rated the examples on Likert scales using basic emotion concepts and dimensional ratings. Then a sampling of the 360 excerpts was carried out using both conceptual frameworks.

- The excerpts were categorized and ranked according to the primary emotion concept tha received the highest rating for the basic emotion examples. From these ranked lists, the top 5 examples and 5 moderately high examples were chosen for each primary emotion (happiness sadness, tenderness, anger, and fear), yielding 50 basic emotion examples ([5 top + 5 moderate × 5 categories).
- For the dimensional model, each dimension was sampled at 4 percentiles along its axis while th other two dimensions were kept constant, resulting in 60 audio examples covering the affec space.
- The mean duration of these 110 excerpts was 15.3 seconds (SD 1.9 s). In the next phase, 11 university students aged 18-42 years rated the 110 tracks using both 3D set and basic emotion (on Likert scales). For the ensuing analyses, the means of the ratings across the participant were used as high consensus existed (Cronbach α > .99 for each concept).

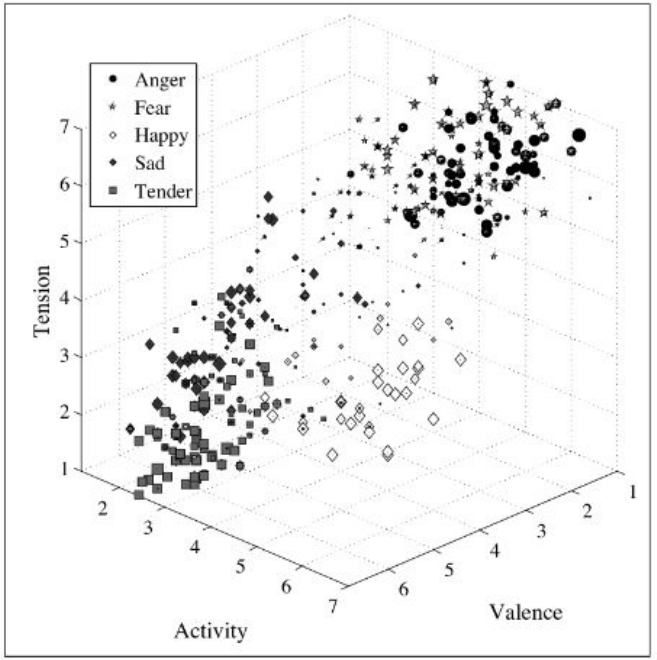
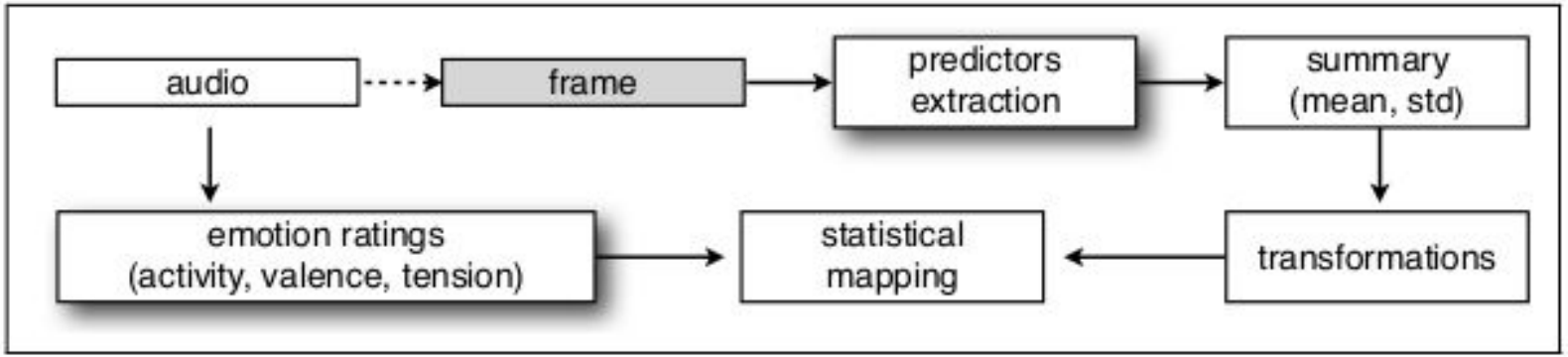


Figure 1. Average ratings of the three dimensions and basic emotions for the 360 soundtrack excerpts.

Research Design

- Quantitative
- This is because we'll split the data into training and testing sets and then try to fit a machine learning model (Regression Model) on our data and analyze the correctness/accuracy of the model by using various performance evaluation metrics. It may involve certain numerical analysis such as accuracy measuring, comparisons based on applying different ML model parameters/weights, etc., and hence the research design is quantitative in nature.

General design of the methodology



Feature Extraction

Extraction of features is a very important part in analyzing and finding relations between different things. The data provided of audio cannot be understood by the models directly to convert them into an understandable format feature extraction is used. It is a process that explains most of the data but in an understandable way. Feature extraction is required for classification, prediction and recommendation algorithms.

Features Extracted

['tempo', 'total_beats', 'average_beats', 'chroma_stft', 'chroma_cq', 'chroma_cens', 'melspectrogram', 'mfcc', 'mfcc_delta', 'rms', 'spectral_centroid', 'spectral_spread', 'spectral_contrast', 'rolloff', 'entropy_fft', 'entropy_welch', 'spectral_novelty', 'poly', 'tonnetz', 'zero crossing rate', harmonic']

Data Transformation

- PLS and other methods take in normalised data as input.
- Hence, the paper suggested using Box-Cox transformation method to normalise the data.
- Box-Cox transformation has many limitations, such as the data should be positive only, etc.
- Hence, we used the Yeo-Johnson transformation method.
- Yeo-Johnson transformation normalises the data and also tries to introduce symmetry.

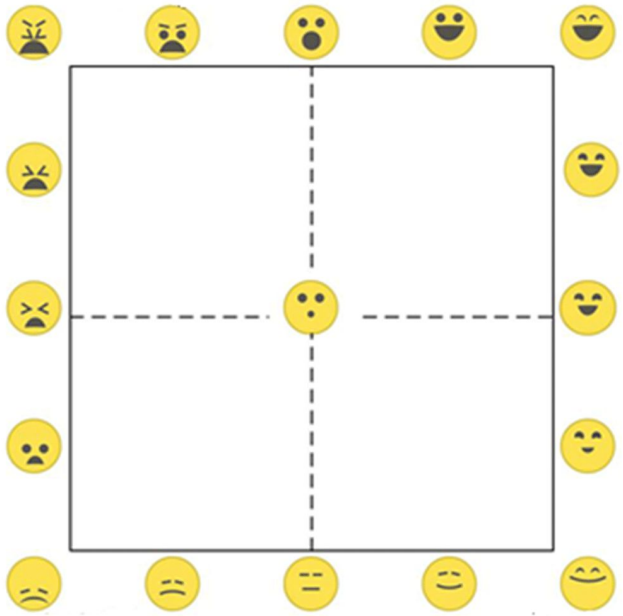
Regression models

- **Multiple Linear Regression (MLR):** Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable, i.e., It is used to estimate the relationship between two or more independent variables and one dependent variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.
- **PCA:** Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed. It is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- **Partial Least Squares (PLS):** Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space.

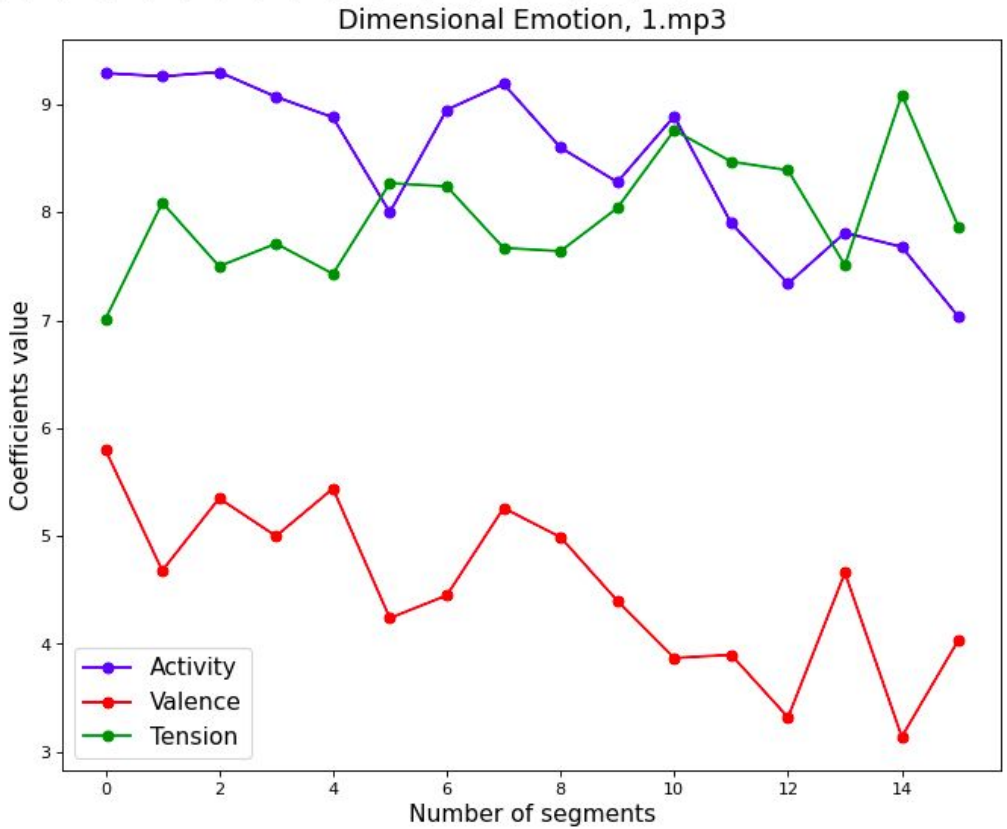
Audio Emotion Visualiser

Built an audio emotion visualiser in python. The tool displays the emotional trajectory based on the three dimensional affective space, dimensions being valence, energy and tension.

The visualiser initially asks the user to input the song. The tool plays the audio in the background and shows a 3D cartesian plane with a point moving in the plane depicting the trajectory. The audio is divided into segments of 15 seconds. The point moves after every 15 seconds, depicting the change in emotional trajectory.



Graph plotted based on the dimensional emotion prediction from the python tool for one of the songs

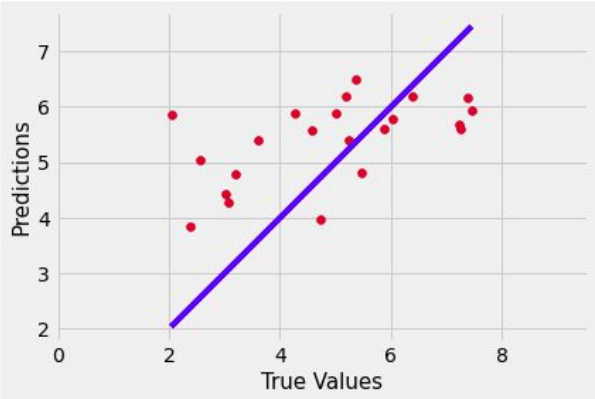


Results

	Prediction rate (R^2)		
ML Model	**Valence**	**Energy**	**Tension**
MLR (PCA)	0.29	0.52	0.50
PLS	0.71	0.87	0.77
MLR (PCA(λ))	0.12	0.59	0.36
PLS(λ)	0.65	0.88	0.76

	Prediction rate (R^2)				
ML Model	**Angry**	**Scary**	**Happy**	**Sad**	**Tender**
MLR (PCA)	0.38	0.34	0.19	0.31	0.40
PLS	0.81	0.74	0.66	0.67	0.70
MLR (PCA(λ))	0.34	0.005	0.19	0.44	0.38
PLS(λ)	0.75	0.70	0.56	0.64	0.74

MLR (without transformation)



Valence

Energy

Tension

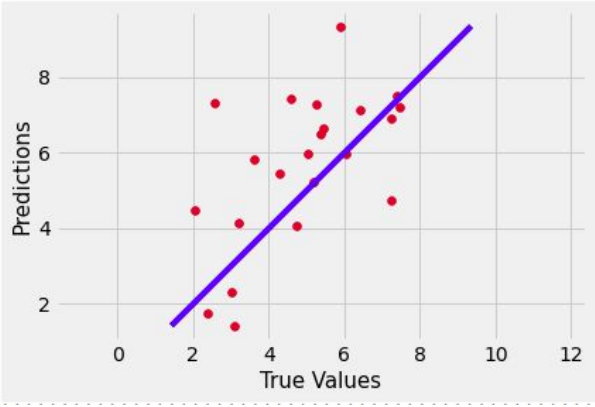
MLR (with transformation)

Valence

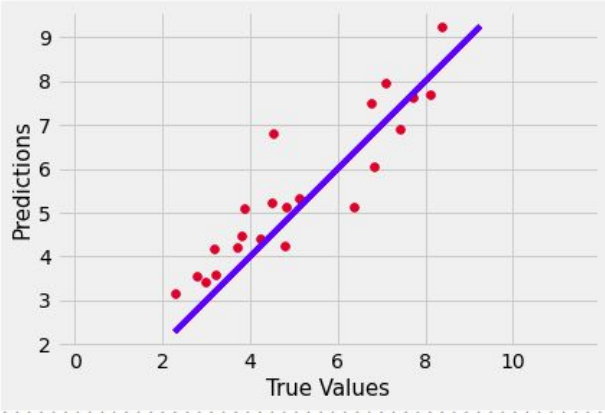
Energy

Tension

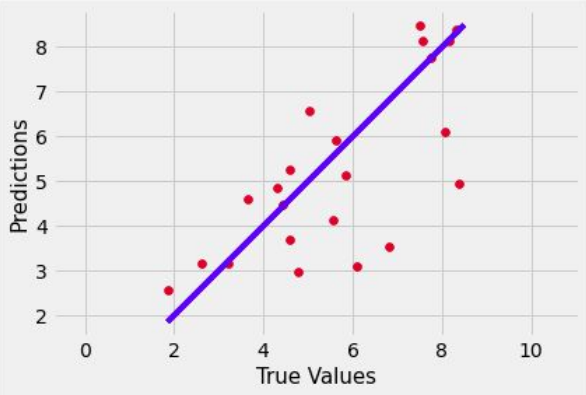
PLS (without transformation)



Valence

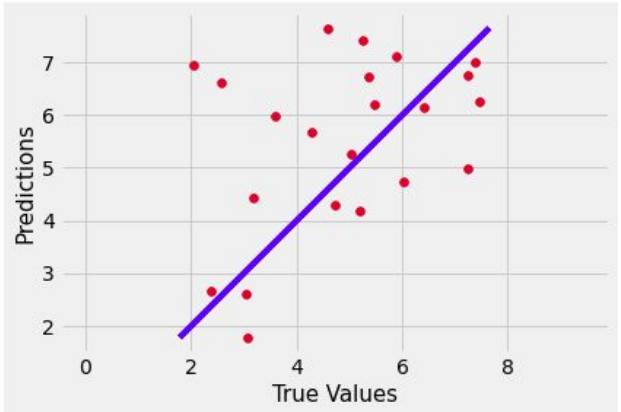


Energy

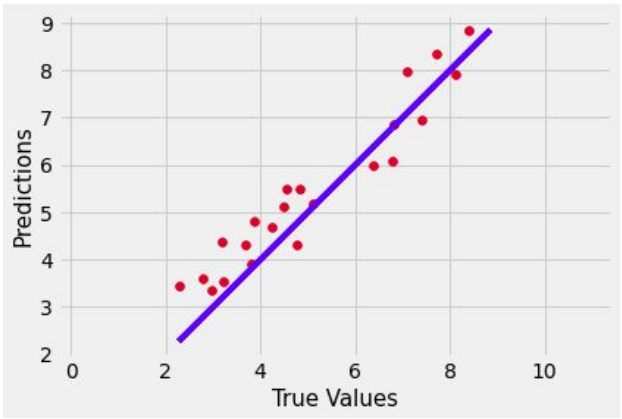


Tension

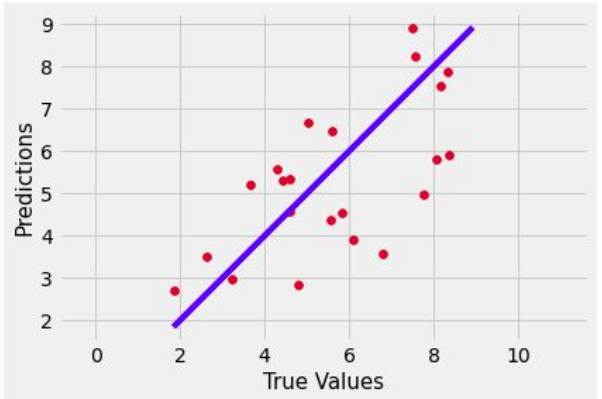
PLS (with transformation)



Valence



Energy



Tension

Inferences and Learnings

We learnt:

- Audio and music feature extraction
- Statistical approaches for excess input variables

Best Regression Model: PLS. This is because PLS assumes covariance exists between features and creates a regression model based on that. PCA, on the other hand, only tries to find a hyperplane which maximises inter-variance.

Limitations

Size of Dataset

- Limited amount of data points (110 in total).
- A larger dataset will provide consistent results and decrease the bias in the model.

Length of Audio Clip

- The length of all the audio clips were 15 seconds.
- This can be reduced to the order of 1 second for effective emotion visualisation of samples of small size.

Future Work

1. Results can be improved by weighting:
    - Music Lyrics
    - Music Genre
    - Music Tags
2. Visualizing the emotional trajectory for smaller audio segments.

Reference:

[1] J. A. Russell. A circumplex model of affect. J*ournal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[2] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8), 2005.

[3] U. Schimmack and R. Reisenzein. Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, 2(4):412– 417, 2002.

[4] Eerola, T., Lartillot, O., & Toiviainen, P. (2009, October). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *Ismir* (pp. 621-626).


Thank You!