IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Predicting emotion in music through audio pattern analysis

To cite this article: J M Brotzer *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **482** 012021

Predicting emotion in music through audio pattern analysis

J M Brotzer[1], **E R Mosqueda**[2], **and K Gorro**[3]

[1 2]University of San Carlos, Talamban, Cebu City, Cebu, Philippines
[3]Department of Computer and Information Sciences, University of San Carlos, Talamban, Cebu City, Cebu, Philippines

[1]jomar.brotzer@gmail.com
[2]ethanray.mosqueda@gmail.com
[3]kdgorro@usc.edu.ph

Abstract. Music plays a big part in the development of humanity and has inspired so many people to do great things. Which is why a study was conducted focusing on music and its relation to human emotions. It tackles on how an emotional response is generated from an audio pattern through the use of a model constructed by the researchers. The work done is aimed to generate ratings of the emotion conveyed in a song represented as valence and arousal. Audio feature extraction and Artificial Neural Networks played big parts in the development of the research in creating the regression model and predicting the human emotion being emphasized in an audio pattern. The findings of the research will be used to give meaning to audio patterns and how their structure and composition leads to different emotional states of man.

1. Introduction

Researchers look for different kinds of correlations between music and the feelings people experience. Although ideas have been realized, there is still room to improve in implementations to reach the desired output: emotion in audio. The application of different combinations of machine learning algorithms with the features have yet to be tested and still remain uncertain on which is the best option. This research attempted to determine whether or not using a neural network with regression would yield a better result in determining valence and arousal values as compared to other methods. Valence is defined as the measure of how positive or negative an emotion is while arousal is how calmly or exciting the emotion is felt. In regression, output will be determined using R^2 score, which is a measure in statistics that deals with how close the data points are to the regression line. The research is aimed to construct a model that would predict valence and arousal R^2 scores given an audio sample. Manipulating audio patterns to derive human emotion has proven to be a difficult task. Some methods of classification have been considered to give emotions but only to a certain extent. The study is only limited to the use of the machine learning algorithm Artificial Neural Networks. Furthermore, the type of AI that has been developed in the research will be of a narrow kind for it only outputs mood ratings. The emotion regression was done using the Thayer's model of emotion and the annotations of each song from the dataset.

2. Review of Related Literature

In doing research, it is important to find references and studies about the topic being researched. Previous studies can provide valuable input for researchers especially in terms of methodology. Recently, Neural Networks have shown promising results in complex machine learning which is why the researchers decided to explore neural networks and predicting valence and arousal values in audio patterns. The following are studies related to this research:

2.1. Feature Selection

The study [1] focused on identifying which set of features, when combined, yielded the best prediction results for valence and arousal. This study had a regression approach towards the problem. They used features extracted by Essentia and Marsyas which are both python libraries used for retrieving audio information such as Low-level, Rhythm and Tonal types. The study concluded that the use of several groups of audio features ended up giving better results compared to fewer groups.

2.2. SVM Approach

This study [2] focused more on the concept of Music Emotion Recognition (MER). They approached the problem using regression to predict the values arousal and valence scores directly. The study wished to evaluate the performance of a regression approach to predicting arousal and valence. They used an SVM as a regressor, which in turn yielded them an R^2 score of 58.3% for arousal and 28.1% for valence as their best result.

2.3. Fuzzy Approach

This study [3] focuses on the use of fuzzy logic for music emotion classification. Seeing the subjective nature of how humans perceive music to be, using fuzzy logic allows for a less binary approach to emotion analysis. Fuzzy classifiers allow for a value to be assigned based on how strong an emotion may be perceived. This approach manages to deal with the situation wherein people share different opinions regarding a certain song as well as being able to detect the variations of emotions found in said music.

2.4. Lyrical Analysis

The study [4] makes use of a bi-modal method of audio emotion classification as it deals with both audio and lyrical features, using these as basis for classifying audio to emotion. Selected machine learning algorithms are implemented in classifying to compare the each of their performances on the data.

2.5. Neural Network Architectures

The article [5] provides information on artificial neural networks. Neural networks have different kinds of applications and design architectures. The article mentioned proves to be insightful for researches involving neural networks giving an in depth understanding of the concept.

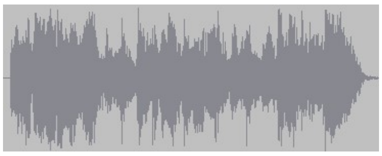
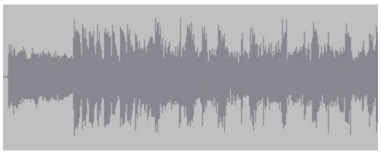
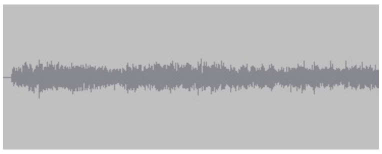
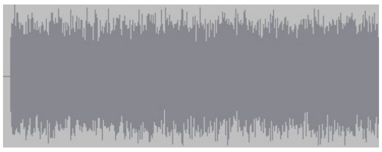
3. Methodology

3.1. Data Gathering

3.1.1. Selecting Data. The data used in the study was taken from the DEAM (Database for Emotional Analysis in Music) Dataset, consisting of 1802 audio files along with valence and arousal values annotated per second and averaged per song. For the study, the researchers used the annotations averaged per song. These values are real numbers that range from 1 to 9. These values are plotted in a plane using Thayer’s Model of Emotion where Valence is the range in x axis and Arousal in the y axis. In Thayer’s model, there are four quadrants with different ranges of valence and arousal. Table 1 shows a sample of songs in the dataset and their corresponding placement in the plane.

The emotion ratings were scaled down to -0.2 to 0.2 to have a zero-centred axis and give lenience to error for the model in predicting values. This gives the neural network more room to train and react on the weight changes as the model is subjected to the data. In figure 1, it is clear that the data is condensed at the centre, in the positive and the negative axes. Basing on the Thayer’s Model, songs that have mostly negative valence with positive arousal and songs that have positive valence with negative arousal are lacking in numbers. This means songs that are intensely unpleasant and calmingly pleasant aren't prevalent in the dataset.

Table 1. Sample Audio of DEAM Dataset. The table shows the varying of structure depending on where they are located in the Valence and Arousal plane.

Audio Pattern	Information	Averaged Song Ratings
	File: 115.mp3 (Quadrant 1) High Valence High Arousal	Valence: 8.4 Arousal: 7.8
	File: 2057.mp3 (Quadrant 2) Low Valence High Arousal	Valence: 3.2 Arousal: 6.8
	File: 488.mp3 (Quadrant 3) Low Valence Low Arousal	Valence: 1.6 Arousal: 3.4
	File: 977.mp3 (Quadrant 4) High Valence Low Arousal	Valence: 7.3 Arousal: 3.5

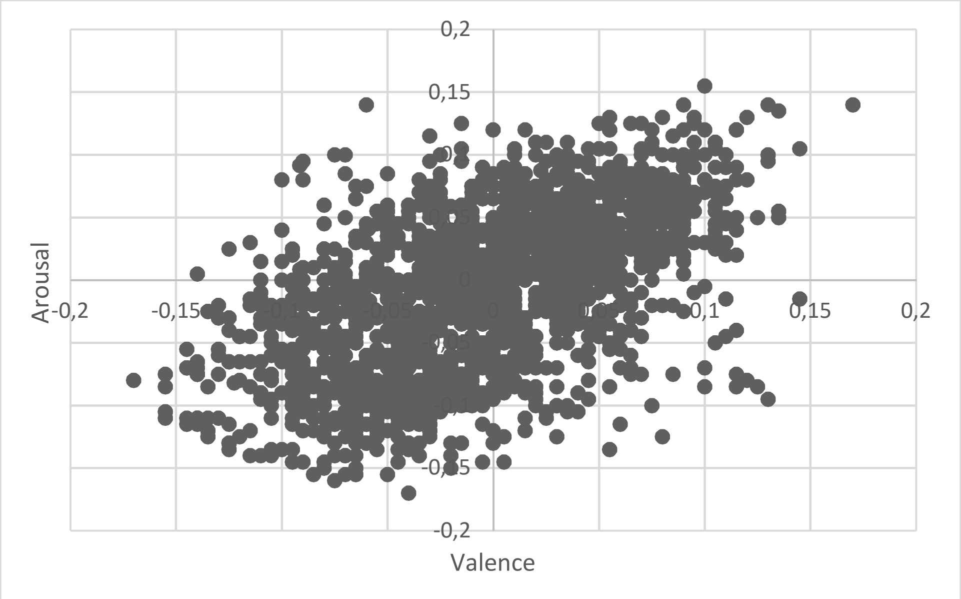


Figure 1. This is a figure showing data points of the valence and arousal labels.

3.1.2. Audio Feature Extraction. For processing the data, the researchers have used LibROSA, a python audio analysis library to extract audio features and NumPy to structure the data into a two-dimensional numpy array of this shape: (1802, 193). The labels of each audio file are likewise stored in the same fashion with a shape: (1802,). These are the selected audio features extracted based on a study on classifying bird songs [6]. The features extracted are already proportioned by LibROSA so that the length of the audio would not affect the shape of the features obtained. Table 2 shows the features extracted with LibROSA.

Table 2. Extracted Audio Features from LibROSA.

Audio Features	Definition	Outputted Shape
Mel-frequency cepstral coefficients (MFCC)	Coefficients that collectively make up a Mel Frequency Cepstrum.	(40,)
Chromagram of a Short Time Fourier Transform	Values that relate to the twelve different pitch classes.	(12,)
Mel-based Power Spectrogram	A Spectrum representation that indicates the fluctuations of the intensity of frequencies	(128,)
Octave-based Spectral Contrast	Decibel difference between peaks and valleys of a spectrum	(7,)
Tonnetz	(German: tone-network) is a conceptual lattice diagram representing tonal space	(6,)
Total Features:		(193,)

Another group of audio features was also taken into consideration from a certain research [2] that gathered information on which audio features contributed greatly to the prediction of valence and arousal. To get the features, the researchers used Essentia to extract the features in table 3 and also shaped them in the same fashion as the audio features extracted from LibROSA apart from the number of features to be inputted to the model

Table 3. Extracted Audio Features from Essentia. (38 features in total)

Audio Features	Feature Type
Melbands Kurtosis	Low-level
Melbands Skewness	Low-level
Spectral Energy	Low-level
Beats Loudness Band Ratio	Rhythm
Key Strength	Tonal
Chords Histogram	Tonal
Chords Strength	Tonal
Harmonic Pitch Class Profile Entropy (HPCP)	Tonal

3.2. Neural Network Construction and Training

Two neural network configurations were created using the libraries: Tensorflow and Keras. One neural network having 1-1-1 NN layer layout (Input, Hidden, and Output) and the other having 1-6-1. The type of neural networks developed were feed-forward; one complex and the other simple. A feed-forward neural network is a kind of neural network where the nodes do not form a cycle and only move in one direction, which is forward and through the nodes. An example of a feed forward neural network is shown in figure 2. The 1-1-1 architecture was to see if a simple NN architecture would be sufficient enough to predict valence and arousal through the use of the different audio features obtained. Otherwise, the researchers would turn to the use of a more complex feed-forward structure.

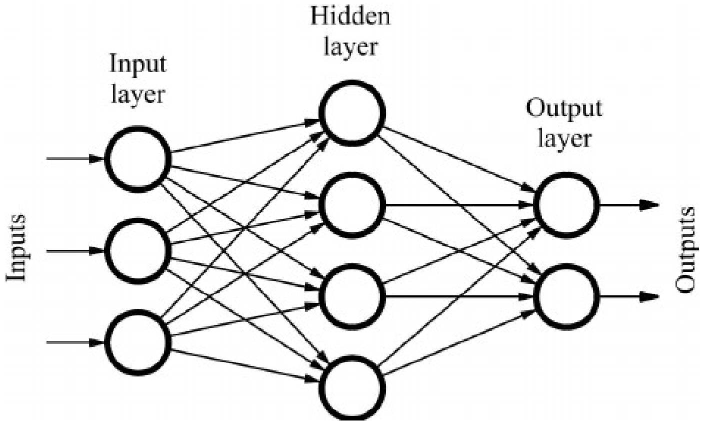


Figure 2. A Simple Feed Forward Neural Network

3.3. Evaluating Models

The models were subjected to a 10-fold cross validation where the dataset was shuffled and per iteration was subjected to training and prediction. Each iteration reported on different regression metrics available from Sci-Kit Learn. This researched used explained variance score, mean absolute error, mean squared error, median absolute error and r^2 score for the verification

4. Analysis of Results

Using the 38 features from Essentia library yielded dissatisfying results when paired with the labels during the train and test process. The 193 features from LibROSA had better results compared to the ones from Essentia but were also ineffective in prediction. So the researchers incrementally added data rows for the train and test process to see if the model would produce better results as the number of data rows increased or decreased. The scaling of the values did not greatly affect the r-squared scores but lessened the mean absolute error. Table 4 displays some of the expected emotion ratings and the predictions made by the model.

Table 4. Arousal and Valence Predictions.

Arousal Predictions		Valence Predictions	
Expected	Predicted	Expected	Predicted
0.0000	0.0166	0.0500	0.0502
-0.1000	-0.0659	-0.0500	-0.0639
0.0500	0.0492	0.0500	0.0421
-0.1000	-0.1058	-0.1000	-0.0886
-0.1000	-0.0562	0.0000	-0.0659

Several number of nodes were tested alongside it but the result from the simple feed forward network that had one hidden layer was not successful compared to the complex network. The best result obtained was from the neural network configuration that had 6 hidden layers each having 500 nodes with rectified linear unit activators and an output node with a linear activation, so the researchers used these values for further testing. Table 5 indicates the best results attained from the model.

Table 5. 10 – Fold Cross Validation Regression Metric Results.

Seed = 7, EPOCHS = 50, Audio Features = 193, Data Rows = 1802.

	Arousal Model	Valence Model
Explained Variance Score	0.3737	0.3285
Mean Absolute Error (MAE)	0.0419	0.0400
Mean Squared Error (MSE)	0.0028	0.0025
Median Absolute Error	0.0355	0.0340
R^2 Score	0.3572	0.3132

5. Conclusion

The neural network architecture developed is experimental and needs refinement. A different type of neural network may prove to be more efficient than a feed-forward network for this kind of problem. A recurrent neural network works well with speech recognition and would be a better option in proceeding with emotion prediction of audio.

The improvement of data would be greatly welcomed as time progresses music continues to evolve. More data rows and different kinds of datasets would better develop the field and lead to better and more accurate models. As of now, the data related to audio sentiment is limited and researchers struggle to get accurate data that would be viable for usage.

It is valid to predict emotion in music because music affects humanity and it is beneficial to know which types of music negatively or positively affects a person. Creating a model that can deduce an audio pattern's emotional rating provides a buffer for people to decide whether or not to listen or create music based on the rating.

References

[1] Grekow J 2017 Audio features dedicated to the detection of arousal and valence in music recordings *IEEE Intl. Conf. on Innov. in Intel. Sys. and App. (INISTA)* pp 40-4

[2] Yang Y, Lin Y, Su Y and Chen H 2008 A regression approach to music emotion recognition *IEEE Transac. on Audio, Speech, and Lang. Proc.* **16(2)** pp 448-57

[3] Yang Y, Chu Liu C and Chen H 2006 Music emotion classification: a fuzzy approach *MM '06 Proc. of the 14th ACM Intl. Conf. on Multi.* pp 81–4

[4] Malheiro R, Panda R, Gomes P and Paiva R 2013 Music emotion recognition from lyrics: a comparative study *6th Intl. Work. on Mach. Learn. and Music, Prague*

[5] Stergiou C and Siganos D 1996 *Neural Networks* online https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[6] Collis J 2016 *Machine Listening: Using Deep Learning to Build an Audio Classifier To Recognise Birdsong*