Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Instrumental Music Emotion Recognition with MFCC and KNN Algorithm

Tiron Dutono[1], Frida Magna Nuriyah[1], Tri Budi Santoso[2]
titon@pens.ac.id, tribudi@pens@ac.id
[1]Electrical Department, Politeknik Elektronika Negeri Surabaya
[2]Multimedia Creative Department, Politeknik Elektronika Negeri Surabaya

Article Information

Submitted : 3 Feb 2023
Reviewed: 10 Feb 2023
Accepted : 26 Feb 2023

Keywords

Music signal; MFCC;
Emotion Classification;
K-Nearest Neighbor.

Abstract

Every piece of music contains emotion in every sound presented. Detection of the music emotion is quite difficult to do because the emotions felt are subjective. Based on this problem, it is necessary to have an automatic classification system to detect the emotions produced in music. In this paper, an explanation of the result to develop an emotional classification system of instrumental music. This system described the process starting with the receiving an input in the form of a music file in the format wav. Furthermore, the feature extraction process is carried out using Mel-Frequency Cepstral Coefficients (MFCC). The result of the extraction of such features are used in the classification process using the K-Nearest Neighbor (K-NN). The system produced output in the form of happy, relaxed, and sad emotions. The output of the system has a classification achieved an accuracy of 97.5% for the value of *k* = 1, reaching an accuracy of 95% for the value of *k* = 2.95% and for *k* = 3, reaching an accuracy of up to 90%.

A. Introduction

Every human being has a wide variety of emotions. Emotions are a person's feelings towards something that affects behavior. There are several ways of expressing emotions. One example of expressing emotions verbally is from the music being listened to.

It is indisputable that music is closely related to emotions. Music can create various emotions for the listener. For example, music that has a slow tempo, the listener may feel calming emotions, sadness, or happiness [1]. Therefore, the music is also often used as a support for the atmosphere for an activity.

A music classification system based on mood parameters with the K-Nearest Neighbor classification method and Self Organizing Map has been presented [2]-[4]. The mood parameters used are based on robert Thayer's energy-stress model.
A music classification system based on mood parameters with the K-Nearest Neighbor classification method and Self Organizing Map has been presented [5] and [6]. The mood parameters used are based on robert Thayer's energy-stress model.

In this paper, we presented a system of classifying emotions towards instrumental music using the K-Nearest Neighbor (K-NN) Algorithm. The objective of this paper is proposed a system to find and classify emotions in an instrumental music automatically. This system begins with the pre-processing of musical instrument features by using the standard audio processing, namely MFCC. The feature classification process is carried out using the KNN algorithm. Through the combination of the two will give a result that will be closer to real conditions.

B. Understanding Emotion in Music

Emotion and Mood in a Music

The emotions are such complex set of interactions between objective and subjective factors, mediated by the nervous system, which can give rise to affective experiences such as feeling of passion and pleasure. Emotion able generate cognitive process such as relevant perceptual effects, activate psychologically widespread assessments with the condition as well as desire. In this case emotions also cause frequent but not always, expressive, goal oriented and adaptive behaviors. While mood is a relatively long-lasting emotional state [7] and [8]. Conceptually some moods can be displayed in four different classes. Mood has a negative valence (dimensional tensions), such as “sad” or “angry”, or positive valences such as “relaxed” or “happy”. The desire elemen, related to the energy dimension, on the y-axis distinguishes different moods from calm. But calm behavior at the bottom to intense and strong at the top as shown in Figure 1.

Music Information Retrieval

Music Information Retrieval (MIR) is a science to understand and process the information from a music files in the form of metadata or the content. The MIR system assists user in searching the music based on its content information. *Music Classification* is one of the popular topics in MIR. The classifications carried out are various, based on the genre of music, singer, and emotion/mood. The objective of Music summary is to find the most representative part in music. This is often assumed in the intro and refrain as the most frequently repeated part of a piece of music.

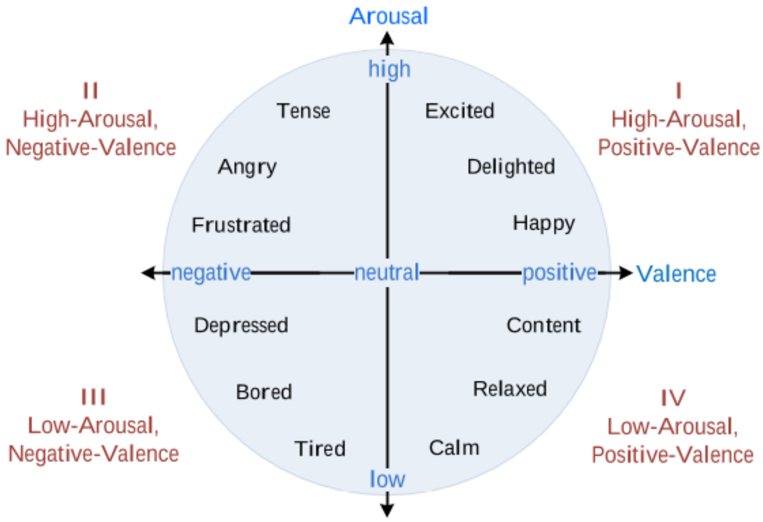

Figure1. Emotions model by Thayer's *arousal-valence*

Music Elements

Music has several constituent elements which are divided into two parts, segmental features, and suprasegmental features. Segmental features are individual sounds or notes that make up music such as duration, amplitude, and pitch. For suprasegmental features are the basic structures of a work, such as melody, tempo, and rhythm. There are certain musical features that are strongly associated with certain emotions [7]. Tempo is usually considered the most important, but several other factors, such as mode, loudness, and melody, also affect the emotional valence of the piece [8].

Mell Scale Cepstrum Coefficient

Mel-Frequency Cepstral Coefficients (MFCC) is one of the most common feature vectors used in speech and music related pattern recognition applications. MFCC has revealed its outstanding performance in speech and music emotion recognition [9] and [10]. The way MFCC works is based on the difference in frequency that is captured by the human ear so that it can represent the characteristics of sound signals as humans represent them. The block diagram of the MFCC as in the Figure 2.

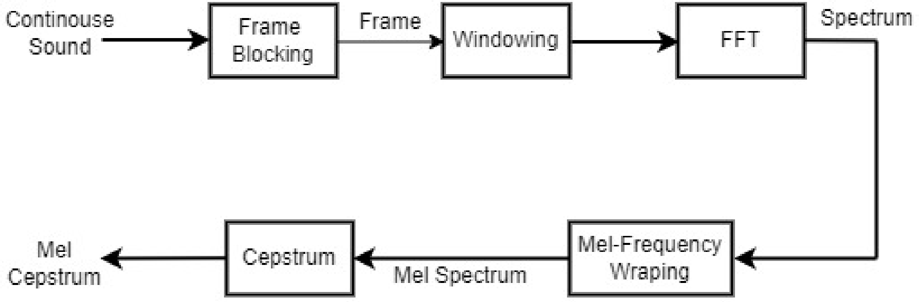

Figure 2. MFCC Block diagram

One of the characteristics of MFCC is frequency scaling using Me-Frequency Warping. To get the Mel Frequency Scale, wrapping is done on the spectrum generated from the FFT. Then the mel scales are grouped using a filter bank. After the spectra are computed, the data are mapped to Mel Scale using an overlapping triangular filter (filter bank). The following is an equation for calculating the Mel Scale:

$$Mel\,(f) = 2595\,\log_{10}(1+f/700) \tag{1}$$

Where is Mel (*f*) is a *Mel Scale* function, and *f* is frekuensi

C. System Design

The design of the emotion recognition system in music can be presented simply as shown in Figure 3. The system will receive input in the form of a music file in mono .wav format, which will then enter the Pre-Processing stage which is the initial stage of signal processing. The next stage of the signal will conduct feature extraction. From the results of the feature extraction, it will be classified using K-Nearest Neighbor (K-NN). Furthermore, the system produces output in the form of emotional results.

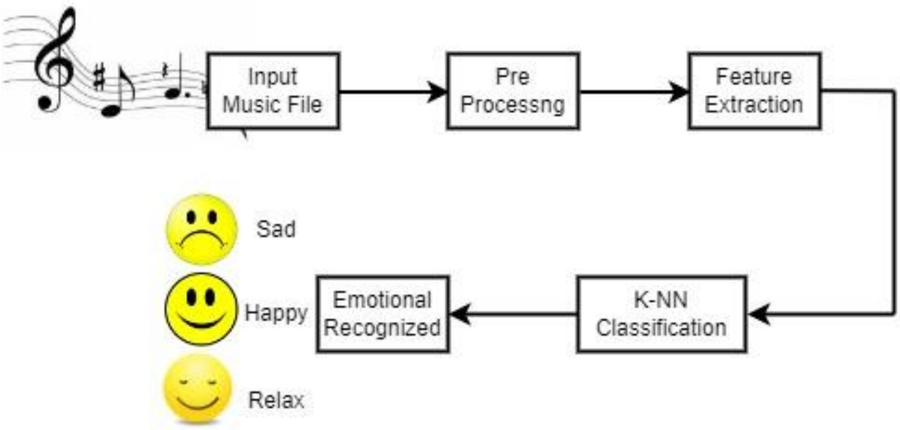

Figure 3. Emotional Recognition for instrumental music

Data Set Collection

The data used is a various of instrumental music which has a duration of 10s upto 15s and a sampling rate of 22050 KHz. The data is taken from the Musical Emotion Classification dataset. At the data collection stage, it is divided into 3 emotions. The total data used in this study is 300 offline data, that is, using data that has been previously stored in the '*.wav' format.

From the collection of music data, obtained as many as 300 data consisting of emotions Happy, Sad, and Relax. The data will then be divided into 80:20 to be used for the training process and the testing process on the classification. Next given the labeling of the music files for each of these folders. The labeling format follows the label of the Musical Emotion Classification dataset.

Feature Extraction

At this stage, several music feature extractions are used, including tempo, beats, spectral centroid, chroma, rolloff and MFCC extraction with 13 coefficients.

For feature extraction, this is done in each folder using a library that is already available in Python.

In the feature extraction process, it begins with structuring the sampled signal into short time frames of about 10s-15s. Then do some extraction for the tempo which is used to determine the speed of the beat of the music. Beats are used to determine the fast and slow rhythm in a song that can affect an emotion. The last feature used is MFCC using 13 coefficients with the Sr value used is 22050 KHz. From the 5 features, the feature set value will be obtained, which value will be loaded in the training database.

The K-Nearest Neighbor algorithm or commonly called as K-NN isa data classification method that work relatively in a simpler way compared to the other classification methods. This algorithm tries to classify a new data whose class is not yet known by selecting a numbers of k data that are closest to the new data. Near or far neighbors are usually calculated based on the Euclidean distance by using this equation:

$$d(x,y)=(\sum(X_i - Y_i)^2)^{1/2} \quad (2)$$

where:

d = distance,
X_i = i-th data training,
Y_i = data testing,
i = record (line i) from the table, with $i = 0 \ldots n$
n = total of data training

The dataset used for the classification process is feature extraction data that has been saved in *.csv file format. The dataset is called which can then be processed for classification. Furthermore, the dataset is divided into training data and testing data, where the data is divided into 80:20 proportions.

The next stage is scaling, namely the standardization of the dataset. The purpose of this scaling is that the range of values used is between $-1 < x < 1$. After the data is divided into training and testing data, the training process for K-NN is carried out.

The next stage is testing to predict the testing data. After obtaining the value of *k* with the level of accuracy. Next is the error rate *k* value, to find out the best *k* with a low error value. There is none exactly method to find the best value for *k*. We need to find out with various values by trial and error and assuming that training data is unknown. The smaller values for K can be noisy and will have a higher influence on the result. The other site, the larger values of *k* will have smoother decision boundaries but increased bias, and computationally expensive. In general, practice, choosing the value of *k* is:

$$k = (N)^{1/2} \quad (3)$$

where *N* is the number of samples in the training dataset.

Another way to choose *k* is though cross-validation. One way to select the cross-validation dataset from the training dataset. Take the small portion from the training dataset and call it a validation dataset, and then use the same to evaluate different possible values of k. We predict the label for every instance in the validation set using with k = 1, k = 2, etc. Then choose the value of k that give the best performance on the validation set and then we can take that value and use that as the final setting of the algorithm to minimizing the validation error.

To evaluate the level of accuracy of the system in predicting emotions data, a test is carried out on the predicted data. The test is carried out using the confusion matrix test method. The confusion matrix visualizes the accuracy of the classification model by making comparisons between the actual and predicted classes. *Biner Classifier* predicted all data instance from the data set testing as a positive or negative. This classification release 4. This classification produces four results, namely True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN).

From the confusion matrix model, measurements can be made to determine the performance of the classification model used. The Performance parameters are accuracy, precision, recall, and f-measure.

- *Accuracy* is a calculation to measure the performance of the classification method.

  Acc. = 100% x (TP+TN) / (TP+FP+TN+FN)
- *Presicion* is useful for measuring the level between the information requested by the user and the answer given by the systems.

  Prec. = 100% x (TP) / (TP+FP)
- *Recall* is useful for measuring the success rate of the system in rediscovering an information.

  Recall = 100% x (TP) / (TP + FN)
- *F-measure* (F_1) to represent the combination of precision and recall. The system will be considered good if has a high *f-measure* value.

  F-measure = 2 x 100% x (Prec. x Recall) / (Prec. + Recall)

D. Result and Discussion

Feature Extraction Analysis

The data used in this study is instrumental music data. Where the data will be analyzed for features consisting of tempo, beats, centroid, chroma, roll off, and MFCC 13 coefficient.

From the dataset from the feature extraction, it looked that the tempo indicates the speed of the music. Where for Happy has a higher tempo value that is above 100, for Sad emotion it has a slower tempo with a value below 100, and for Relax it has a tempo value that is in the middle of the two. The same can be seen in the emotional beats of Happy, Sad, and Relax. From this, it looked that the higher the tempo value, the higher the beats value obtained. The last feature is MFCC with 13 Coefficient. The cepstral Mel-frequency (MFCC) is a coefficient commonly used to represent the texture or timbre of a sound. Next is to extract the top 13 Mel-frequency cepstral coefficients. In MFCC, the signal passes through a pre-emphasis filter, then it is sliced into frames and a function window is applied to each frame. Fourier transform is active every frame and the power spectrum are calculated and then the filter bank is calculated.

Apart from the dataset, the feature analysis can also be done with a plot using the '*sns.boxplot'*, namely the tempo shown in Figure 4. The results in the figure show that the tempo with Happy emotions is higher than Sad emotions, while Relax emotions are in the middle.

The extraction based on bats, in a similar way to the step of drawing based on tempo properties, shows that the average beats of Happy's emotion is higher than that of Sad's emotion. This is indicated by the minimum value of Happy emotional beats, which is 110, while the minimum beat of Sad emotion is 70. Relax emotions have a minimum value of 150 and a maximum of 220.

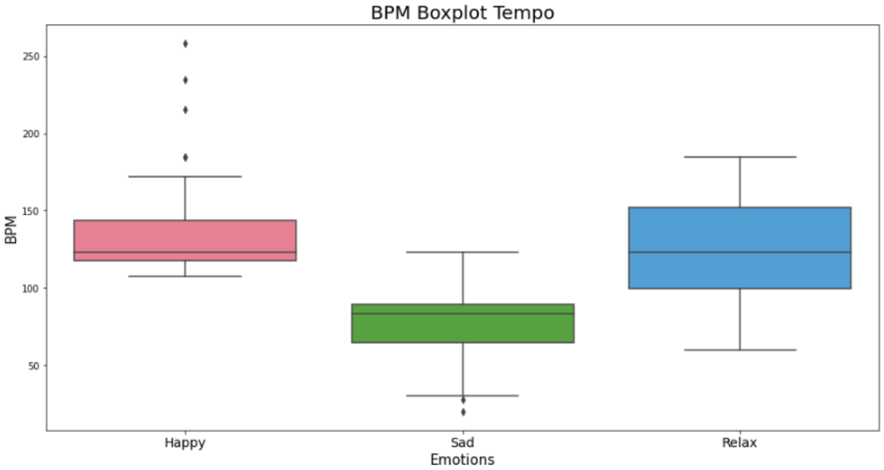

Figure 4. Tempo plot in the Boxplot

Classification Result Analysis

Feature extraction data are classified into Emotions Happy, Sad, or Relax by using the KneighborsClassifier function. This classification process is carried out with the proportion of training data as much as 80% and testing data as much as 20%.

In the *K*-NN classification process with different *k* values. The goal is to find out the Accuracy value for every *k*. In this research, the values of *k* are = 1, 3, 5, 7, and 13. For the value of *k* = 1, it looked that there are 8 prediction errors with details where Happy data is predicted to be Relaxed to be 2, Relax is predicted to be Sad there are 2 data, Sad is predicted to be Relax is 3 data, and Happy is predicted to be Sad is 1 data. This can happen because there are similarities in the tone of Relax with Happy and Sad. In addition, it can also be caused by the small amount of Relax data, so that the Relax emotion classification has not been maximized. The next process is a confusion matrix to evaluate the system. Example output for the value of *k* = 1 as shown in the following table.

Table 1. Confusion Matrix of k =1

Confusion Matrix		**Original Version**		
		Happy	**Relax**	**Sad**
Prediction	*Happy*	16	2	1
	Relax	0	10	2
	Sad	0	3	13

Evaluation of the results of the classification process that has been obtained previously using the confusion matrix. This performance can be seen from several parameters of precision, recall, and f-measure for each emotion.Based on the guideline of the classification results parameters, the results of the classification test are in Table 2 where the overall k value for the classification of happy and sad emotions is good, which is in the range of 80% - 90%. However, Relax's emotions are classified as very bad. This is because the emotional data is still relatively small, causing the classification process to be not optimal.

E. Conclusion

From the observations during the design, implementation, testing and analysis stages that have been carried out, we have the following conclusions:

1. The Mel Frequency Cepstral Coefficient (MFCC) method combined with the KNN algorithm was successfully used to obtain feature extraction data.
2. This classification process consists of 2 emotions, namely Happy and Sad with the achievement of 80% to 90%, but for Relax it still reaches the value of 60%-70% so it is still in the poor category.
3. The effect of the value of k on accuracy is that the greater the value of k used, the lower the accuracy obtained.

F. Acknowledgment

The author would like to thank The Ministry of Research and the Higher Education Republic of Indonesia for supporting this work via the fiscal year 2022 with EEPIS Local research grant, also thank the laboratory member for providing support while carrying out experiments.

G. References

[1] S. Swaminathan dan E. G. Schellenberg, “Current Emotion Research in Music Psychology,” *Emotion Review,* vol. 7(2), pp. 189-197, 2015.

[2] Richard Orjesek, et. all, "DNN Based Music Emotion Recognition from Raw Audio Signal", 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16-18 April 2019.

[3] Zijing Gao, et.all., "A Novel Music Emotion Recognition Model for Scratch-generated Music", 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15-19 June 2020.

[4] Na He and Sam Ferguson,"Multi-view Neural Networks for Raw Audio-based Music Emotion Recognition", 2020 IEEE International Symposium on Multimedia (ISM), Naples, Italy, 02-04 December 2020..

[5] U.S. Yeşim, and V. Asaf, “In-Depth Analysis of Speech Production, Auditory System, Emotion Theories and Emotion Recognition”, 2020 8th International Symposium on Digital Forensics and Security (ISDFS), Beirut, Lebanon, 2020.

[6] R. E. Thayer, The biopsychology of mood and arousal, Oxford University Press, 1990.

[7] K. R. Scherer dan M. R. Zentner, “Emotional effects of music: production rules,” Music and Emotion: Theory and Research, pp. 361-387, 2001.

[8] A. Gabrielle dan E. Stromboli, “The influence of musical structure on emotional expression,” Music and Emotion: Theory and Research, pp. 223-243, 2001.

[9] W. Chai, “Automated Analysis of Musical Structure,” Master of Science Thesis in Media Arts and Sciences, Massachusetts Institute of Technology, 2005.

[10] Subhasish Ghosh, Md. Omar Faruk Riad, “ Attention-based CNN-BiGRU for Bengali Music Emotion Classification”, Indonesian Journal of Computer Science, Vol. 11, No. 3, pp. 801-815, 2022.