Indonesian Music Emotion Recognition Using Rhythm and Timbre Features

Julius Bata*[1], Dominikus Jarvis[1]

[1]Department of Information System, Atma Jaya Catholic University of Indonesia, Indonesia
**julius.victor@atmajaya.ac.id*

Abstract

This research aims to recognize and classify emotion in Indonesian music. Two types of features were extracted: rhythm and timbre. These features are used to train a classifier based on Support vector machines. The dataset used in this study consists of 115 Indonesian popular songs and divided into four emotion class: happy, sad, angry, and relax, which derived from the Thayer's emotion model. A series of experiments were conducted with ten repetitions of 10-fold cross-validation and measured by accuracy. The results show that timbre and rhythm are better for arousal classification than valence classification.

Keywords: *We Music emotion recognition, Rhythm, Timbre, Valence-arousal model, Indonesian music classification*

1. Introduction

Music has become part of daily human activities. The bigger capacity of storage in mobile devices and faster internet access make it easy to access music whenever and wherever. Nowadays, there are a large number of songs with online or personal digital collections. The problem is how to access the song collections effectively and efficiently [1][2]. Research in the field of Music Information Retrieval (MIR) is a study focused on solving problems in managing and accessing a collection of digital music [1][3]. One of the important topics of MIR research is the automatic song classification.

Classification of songs refers to a process of labeling (tagging) songs based on certain information. Traditionally songs are classified based on the artists, band (group), year, and album. In addition, there is other information used to classify songs, namely similarity, style genre, and emotion. Although the classification is initially often done based on genres, emotions are also begun to be used in classification. Music is an expression of feelings that contains emotions. One reason to listen to music is that music has a relationship with emotions [4]. Moreover, emotions are also one of the information used to search for music and organize it in the playlists. It is, therefore, necessary to have a system that can recognize and classify emotions of pieces of music [5].

Various studies have been conducted to recognize and classify songs based on emotion. Most studies modeled emotion recognition as a supervised machine-learning task [3][5], where a number of features are extracted from song audio data, then this feature is used to train a classifier. There are various types of features used, such as timbre, rhythm, harmony, and dynamics [5]. Several machine learning methods have been used to recognize song emotion, such as Support Vector Machines [6], k-Nearest Neighbors, ID3 [7]. Previous studies show those timbre and rhythm features are the two most popular features in music emotion recognition and classification [8][9]. Therefore, in this study, the features used were timbre and rhythm. Both features were used to train the SVM-based classification model.

Studies on song classification based on emotion have received a lot of attention in recent years. A number of studies have been conducted and yielded satisfying results (e.g.,[2][6]). However, there are only a few studies on Indonesian songs. Several attempts have been made to classify Indonesian Song based on emotion. Indonesian song emotion classification study was conducted by [10]. The study in [10] attempted to classify kid's songs based on mood. The authors used the rhythm pattern as features. The classification was performed using k-Nearest Neighbor (k-NN) and self-organizing map (SOM). The result shows that SOM is better than k-NN.

Different from [10], Tomo et al. [11] worked on recognition of emotions in Indonesian traditional music. They created a Wayang robot that can move autonomously to follow the pattern of gamelan music. To achieve this, they developed a gamelan music emotional recognition system. In this study, the feature used is Mel-frequency Cepstrum Coefficients (MFCC), which is the Spectral features. A total of 800 gamelan music data are used to train classification models based on neural-networks. In this study, three types of emotional labels were used: delighted, afraid, relaxed, and added 1 label noise. The system is able to recognize emotional types by more than 90% except for relaxed.

In another study, Binanto [12] performed a mood classification using the keroncong song. This study uses two features, namely intensity, and tempo. Based on these two features, 78 keroncong songs are grouped into two groups. Each group consists of two types of mood, group 1 consists of contentment and depression, while group 2 consists of exuberance and anxious/frantic. They use the median method for classification.

In this study, we perform emotion classification and recognition of Indonesian songs. Contrast to previous studies, this study uses Indonesian pop songs. The difference also lies in the features and classification methods used. The main purpose of this study was searching the most influential feature for emotion classification and recognition. Moreover, this study also conducted classification based on the types of valence and arousal of a song. This study follows and based on [13], which considering the classification task as a binary classification.

The rest of the paper is organized as follows: the research methodology used in this paper is described in section 2. The results of this study are presented and discussed in section 3. Section 4 concludes the paper and proposes future work.

2. Methodology

This study was aimed to classify Indonesian songs based on emotion. The method consists of several steps: data preparation, data pre-processing, feature extraction, model training, and classification, as shown in Figure 1.

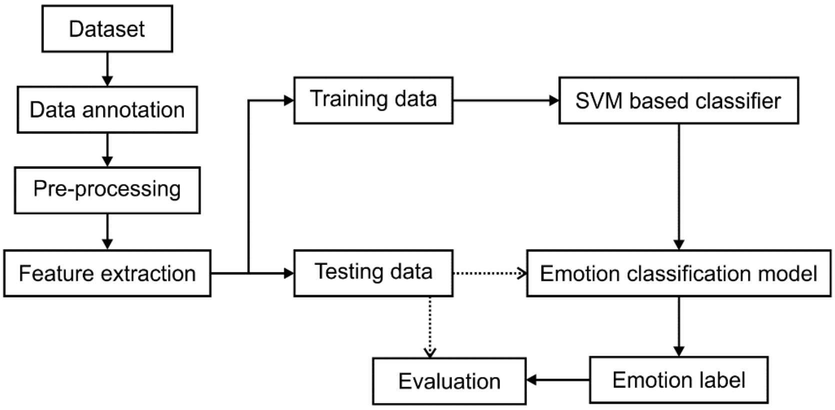

Figure 1. Music Emotion Classification Process

2.1. Dataset

The first stage of this study was collected MP3 file data. The data used consisted of 115 Indonesian pop songs, which were downloaded from several websites. Every song is labeled with the corresponding emotion type. This process was conducted manually. Table 1 and 2 shows the number of songs in each emotion category. This study adopts Thayer's emotion model, as in [14], and four emotion categories were used: happy, sad, angry, and relax. In general, Thayer's emotion model divides emotions into four quadrants of valence-arousal, as shown in Figure 2.

Table 1. Number of Songs in 4-class

Emotion Class	# of songs
Happy	31
Sad	34
Angry	22
Relax	28

Table 2. Number of Songs in Arousal-Valence

Emotion Class	# of songs
Arousal (high)	53
Arousal (low)	62
Valence (positive)	59
Valence (negative)	56

As stated before, this study adopted a binary classification for each song in the dataset. Therefore, for every emotion category, there are two types of data, positive and negative. Negative data of a category were selected based on the opposite quadrant in the emotion model. For example, a song in "sad" quadrant will be negative data for “happy” quadrant.

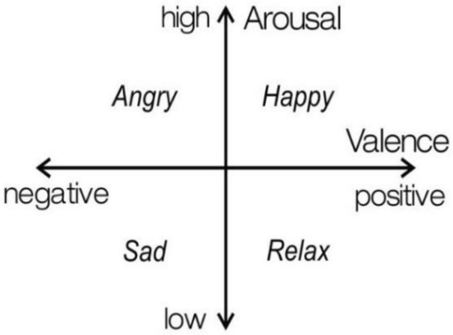

Figure 2. Thayer’s Model of Emotion

2.2. Data Pre-processing

To prepare the audio data for the feature extraction process and to make sure every song in the same format, a pre-processing step is conducted. All songs in the dataset are converted to a uniform format with 44.1 kHz sampling frequency and 128 kbps bit rate.

2.3. Feature Extraction

There are several audio features that commonly used in song emotion recognition, such as dynamics, timbre, harmony, and rhythm. Although there are lots of features, timbre and rhythm were found the most useful and popular for music emotion recognition and classification [9]. In this work, we use jAudio [15] with default parameter values to extract rhythm and timbre features from audio data. Timbre is related to spectral information (for example, MFCC), while the rhythm is related to tempo and beat (for instance Strongest Beat). All the extracted features are listed in Table 3.

Table 3. The Features List

Feature Types	Features	Dimensions
Rhythm	Beat Sum, Strongest Beat, Strength of the Strongest Beat	6
Timbre	Spectral centroid, Spectral Rolloff Point, Spectral Flux, Spectral Variability, Zero Crossing, Strongest Freq via Spectral Centroid, MFCC	38

2.4. Model Training

Most of the works on music emotion recognition and classification have used supervised learning models include k-NN, Naïve Bayes, and SVM. Among them, the SVM method shows good results in previous studies [9].

Basically, SVM looks for the best hyperplane, a line that can separate data into two classes linearly [16]. When the data is not separable linearly, SVM mapping to higher dimension space where data is separated linearly. This method is known as kernel trick [17]. In this study, the implementation was performed using LibSVM [18]. LibSVM is a popular SVM implementation and is often used in music emotion recognition studies. This paper employs RBF kernel and perform grid search tool.

2.5. Classification and Evaluation

This paper evaluated two types of features, namely rhythm, and timbre on the task of Indonesian music emotion recognition. In order to do this, two kinds of the experiment were conducted: 1) valence-arousal classification and 2) 4 class emotion classification. In both of those experiments, three types of features were used and compared: 1) timbre only, 2) rhythm only, and 3) timbre and rhythm. All experiments were performed using a binary classification approach in which every song was determined to be classified in a certain emotional type or not (example: song “A” was “happy” or “not-happy”, song “A” was “angry” or “not-angry”).

All the experiment was conducted by performing ten times 10-fold cross-validation [19]. This method is used when the amount of data used is small. In every experiment, the performance of the classifier model was measured based on the accuracy. The accuracy of each model was the average of 10 times of 10-fold cross-validation. We calculate the average of accuracy using the following formula:

$$accuracy = \frac{\#\ correct\_data\_predict}{\#\ all\_test\_data}\ x\ 100\ \% \qquad (1)$$

3. Result and Discussion

The main objective of this study is to find which features is better for Indonesian music emotion recognition task. To achieve this, we use 115 Indonesian pop songs. The data will be used in music emotion classification. We extract two types of features, which are timbre and rhythm. These two features are then used in experiments. We conducted two types of experiments, the first attempt to classify arousal-valence, and the second experiment to classify four types of emotions. In each experiment, the classification was performed using the timbre, rhythm, and a combination of both. In total, we conducted six experiments, and in each experiment, we measured the accuracy of the classification model.

The first experiment was conducted for Arousal-valence emotion classification. Table 4 lists the results of the classification of valence-arousal. The highest average accuracy for the classification of arousal obtained by using a timbre feature that is 71.3%, while for valence was obtained by using rhythm feature that gives 64.35 %. As seen in Table 4 shows that the accuracy of arousal classification is better than valence.

Table 4. Average Accuracies of Arousal-Valence Classification

Features	Arousal	Valence
Rhythm	67.83 %	64.35 %
Timbre	71.3 %	58.26 %
Rhythm + Timbre	70.43 %	57.29 %

The second experiment was conducted for 4-class emotion classification. The results of the average accuracy rate for 4-class classification as in Table 5. Among the four emotion category, the highest average accuracy for 4-class classification is angry, which yield 76.4 %. The highest accuracy for angry class reaches 80 % when using the rhythm feature. Rhythm is a feature associated with the tempo and beat of the music. Most of the song with the emotion of anger has a faster tempo than the relaxed. Therefore, the classification of these two classes, anger and relax, have better accuracy than the happy and sad class.

Table 5. Average Accuracies of 4-class Classification

Features	Happy	Sad	Angry	Relax
Rhythm	61.79 %	69.23 %	76.4 %	74.6 %
Timbre	63.23 %	66.15 %	73.2 %	70.8 %
Rhythm + Timbre	61.85 %	69.23 %	74.8 %	70.6 %

4. Conclusion

This paper presents two parts of the study on Indonesian music emotion recognition. The first part is arousal-valence classification, and the second part is the 4-class emotion classification. For both studies, two types of feature were used, rhythm and timbre. The results showed that rhythm and timbre are better for arousal classification than valence. On the task of 4-class classification, rhythm features with six dimensions achieved the best performance for the angry category.

This study has some limitations that include the number of datasets used, which only 115 songs. In future studies, an experiment on a larger dataset needs to be done. Future studies also can be done using more features. Besides audio, the music also has other parts like lyrics that can contain semantic information in it. Future studies will focus on the use of lyrics as a feature.

References

[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes and M. Slaney, "Content-based music information retrieval: current directions and future challenges, Proceedings of the IEEE., vol. 96, no. 4, **(2008)**, pp. 668 – 696.

[2] S. Mo and J. Niu, "A Novel Method Based on OMPGW Method for Feature Extraction in Automatic Music Mood Classification", IEEE Transactions on Affective Computing., vol. 10, no. 3, **(2019)**, pp. 313-324.

[3] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendations: state of the art and challenges", Computer Science Review., vol. 6, no. 2-3, **(2012)**, pp. 89-119.

[4] C. L. Krumhansl, "Music: a link between cognition and emotion", American Psychological Society., vol. 11, no.2, **(2002)**, pp. 45-50.

[5] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: a review", ACM Trans. on Intelligent System and Technology., vol. 3, no. 3, **(2012)**, pp. 40:1–30.

[6] J. M. Ren, M. J. Wu and J. S. R. Jang, "Automatic Music Mood Classification Based on Timbre and Modulation Features", IEEE Transactions on Affective Computing.,vol. 6, no. 3, **(2015)**, pp. 236-246

[7] M. Sudarma and I. G. Harsemadi, "Design and Analysis System of KNN and ID3 Algorithm for Music Classification based on Mood Feature Extraction", International Journal of Electrical and Computer Engineering (IJECE)., vol. 7, no. 1, **(2017)**, pp. 486 – 495.

[8] L. Lu, D. Liu and H. J. Zhang, "Automatic mood detection and tracking of music audio signals", IEEE Trans. On Audio, Speech and Language Processing., vol. 14, no.1, **(2006)**, pp. 5-18.

[9] Y. Song, S. Dixon and M. Pearce, "Evaluation of musical features for emotion classification", In Proc. of the 13th Intl. Society for Music Information Retrieval Conf., Porto, Portugal, **(2012)** October.

[10] K. C. Dewi and A. Harjoko, "Kids song classification based on mood parameters using k-nearest neighbor classification method and self organizing map", Proceedings 2010 Intl. Conf. on Distributed Framework for Multimedia Applications, Yogyakarta, Indonesia, **(2010)** July 2-3.

[11] T. P. Tomo, G. Enriquez and S. Hashimoto, "Indonesian puppet theater robot with gamelan music emotion recognition", Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO), Zhuhai, China, **(2015)** December 6-9.

[12] I. Binanto, "A method of mood classification on keroncong music", Proceedings of the 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, **(2018)** April 28-29.

[13] X. Hu and J. S. Downie, "Improving mood classification in music digital libraries by combining lyrics and audio", Proceedings of the 10th ACM Joint Conf. on Digital Libraries, Australia, **(2010)** June.

[14] Y. H. Yang, Y. C. Lin, H. T. Cheng, I. B. Liao, Y. C. Ho and H. H. Homer, "Toward multi-modal music emotion classification", In: Huang YM.R. et al. (eds) Advances in Multimedia Information Processing - PCM 2008. PCM 2008. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol. 5353, **(2008)**, pp. 70–79.

[15] C. McKay, I. Fujinaga and P. Depalle, "jAudio: a feature extraction library", Proceedings of the 6th Intl. Society for Music Information Retrieval Conf., London, UK, **(2005)** September 11-15.

[16] C. Cortes and V. Vapnik, "Support vector networks", Machine Learning, vol. 20, **(1995)**, pp. 273–297.

[17] C. J. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol. 2, **(1998)**, pp. 121–167.

[18] C. Chang and C. Lin, "LIBSVM: a library for support vector machines", ACM Trans. on Intelligent System and Technology., vol. 2, no.3, **(2011)**, pp. 27:1–27

[19] I. H. Witten and E. Frank, Data mining practical machine learning tools and techniques, Morgan Kaufmann Publishers, San Francisco, **(2005)**.