Classification of Music Emotion with Gaussian Mixture Model (GMM)

Myat Ko Ko
University of Computer Studies, Yangon
pekyaw@gmail.com

Abstract

Music is a link between cognition and emotion, and people are not able to share same feeling for a song. There has a need to process vast qualities of musical data. One of the operations is music emotion classification which is very popular today and an automatic extraction is needed, relating to various aspects of music. Music emotion recognition through a learning model is considered in this paper. In order to capture the salient nature of music signals features such as cepstral is applied. Classification of music signals is considered by Gaussian Mixture Model (GMM). In this approach, Thayer's model is adopted for the description of emotions. This music mood detection approach is validated through an experimental study on a dataset containing 60 famous popular songs from English albums.

1. Introduction

Music is present in every culture, and it plays a prominent role in people's everyday lives. According to a recent report, the prevalence of most leisure activities, such as watching TV or movies or reading books, has been overtaken by music listening. As the music databases grow, more efficient organization and search methods are needed. Emotion classification of music has become a matter of interest, mainly because of the close relationship between music and emotions. There have been papers which present the relationships between emotions and audio features automatically extracted from the signal (raw digital data).

Music Emotion Classification (MEC) is still a challenging task. One of the major difficulties lies in the fact that emotions are hard to be described in a universal way. The adjectives used to describe emotions are ambiguous, and the use of adjectives for the same emotion can vary from person to person. Listening mood, environment, personality, age, cultural background etc, can influence the emotion perception. Because of these factors, classification methods that simply assign one emotion class to each song in a deterministic manner do not perform well in practice. Studies on emotions have been carried out for many centuries, including research on various levels and aspects of music perception. Recent studies focus both on cognitive and neurophysiological aspects of emotions such as auditory disgust and pleasure responses, intrinsic and expressive emotions, expectation, and personality correlates. Emotions are an inherent part of music and their role cannot be overvalued. Research on this topic is interdisciplinary, relating to music emotion and psychology, philosophy, musicology, and also biology, anthropology, and sociology. Emotions can be communicated via musical structures, and they also affect performance. The Thayer's emotion plane [1] is commonly adopted to avoid the ambiguity of adjectives. It defines the emotion classes dimensionally in terms of arousal (how exciting/calming) and valence (how positive /negative).

The outline of this paper is as follows. Related work on music emotion classification is presented in Section 2. A brief description on mood model used in this paper is described in Section 3. In Section 4, derivation of cepstral feature and some background on Gaussian Mixture Model is outlined. Section 5 and Section 6 deal with the proposed framework and experiments carried out. Discussion is given in Section 6 and conclusions are drawn in Section 7.

2. Related Work

In [2], there is an analysis of the associations between emotion categories and audio features automatically extracted from raw audio data. This work is based on 110 excerpts from film soundtracks evaluated by 116 listeners. This data is annotated with 5 basic emotions (fear, anger, happiness, sadness, tenderness) on a 7 points scale. Exploiting state-of-the-art Music Information Retrieval (MIR) techniques, the system extract audio features of different kind: timbral, rhythmic and tonal. The classifier is Support Vector Machines. In [3], author model emotions as continuous variables composed of arousal and valence values (AV values). The *R2* statistics of the AV values reach 60% and 19% by SVR (Support Vector Regression). The classification accuracy reaches 84% and 68%, competitive to existing categorical approaches.

A user study on the usefulness of the PANAS-X emotion descriptors as mood labels for music is presented in [4]. It describes the attempt to organise and categorise music according to emotions with the

help of different machine learning methods, namely Self-organising Maps and Naive Bayes, Random Forest and Support Vector Machine classifiers. Also, in this paper, it is described that with something so subjective as emotion, it might be more promising to build individual emotion classifiers for each listener than to try and derive a general notion of what song belongs to which emotional class.

In [5] two fuzzy classifiers are adopted to provide the measurement of the emotion strength. The measurement is also found useful for tracking the variation of music emotions in a song. Results are shown to illustrate the effectiveness of the approach. Fuzzy k-NN classifier (FKNN) and Fuzzy Nearest-Mean classifier (KNM) are used. It is concluded that their presented approach performs better than conventional deterministic approaches because it is able to incorporate the subjective nature of emotion perception in the classification. A novel approach to extract musical segments of significant emotional expressions is proposed in [6]. The first step of this approach is to extract music segments that most listeners will report similar emotion. Such segments are called segments of significant emotional expression (SSEE).

The research mainly adopts the essence of self-report method and dimensional approach. There is a brief introduction to the dimensional approach and the tool for subjects to report their appraised emotion. A dimensional structure can be derived from any type of response data. The most common sources to produce a dimensional structure are similarity judgments of emotion words or facial expressions. They are analyzed using factor analysis or other multidimensional scaling methods. Many scientists like to describe emotions in terms of continuous dimensions as a convenient method, rather than discrete classifications. The two most generally used dimensions are valence (positive/negative) and arousal or activation (calm/excited) [7].

In [8], the authors propose two novel methods, bag-of-users model and residual modeling, to accommodate the individual differences for emotion-based music retrieval. The proposed methods are intuitive and generally applicable to other information retrieval tasks that involve subjective perception. Evaluation result shows the effectiveness of the proposed methods. The music database consists of 60 popular songs from English albums. 99 subjects are recruited from the campus, making each song annotated by 40 subjects. Each song is represented by 80- dimension Mel-frequency cepstral coefficients. Support vector regression (SVR) is adopted to train the regression models. The bag-of-users model provides a better way to aggregate the individual perceptions of the subjects, while the residual modeling makes a personalized system focus on music content and user perception in different stages. The novel perspectives introduced in this paper can be applied to other applications that involve subjective human perception.

3. Thayer's Model

Emotions can be represented in multidimensional space, for instance in 2 dimensions, on activation vs. quality plane. For the description of music emotions, Thayer's model will be adopted. As shown in Figure 1, the 2D emotion space (2DES) is divided into 4 quadrants, and different emotions are placed on the plane in such a way that each emotion (a point in 2DES) can be represented by a 2x1 vector. This results in a valence-arousal plane. The right (left) side of the plane refers to the positive (negative) emotion, whereas the upper (lower) side refers to the energetic (silent) emotion. To be consistent with the 2DES model, 4 emotion classes are defined, each corresponding to a quadrant.

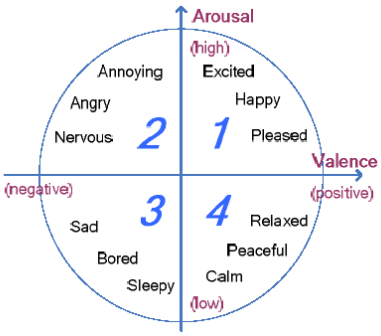

Figure 1. Diagram of Thayer's model.

4. Background

Features are generally extracted to represent the salient nature of the signals. In this work, mel frequency cepstral coefficients which has been most widely used short-time features speech processing applications like speech recognition and speaker verification is adopted. The features extracted for any classification problem are of little use without a good classifier. Classifiers are also known as the back-end of the classification problems. Gaussian Mixture Model has been used in many statistical and signal processing applications. It is the most commonly used classifier for speech/music discrimination. Mel frequency cepstral coefficients and GMM, the selected classifier for the proposed work are described below.

4.1 Mel Frequency Cepstral Coefficients (MFCC)

It represents the shape of the spectrum with very few coefficients. It is the coefficients of the Mel cepstrum. The cepstrum, is the Fourier Transform (or Discrete Cosine Transform DCT) of the Mel cepstrum. The Mel cepstrum is the cepstrum computed on the Mel bands instead of the Fourier spectrum. The use of Mel scale allows better to take into the mid-frequencies part of the signal. Figure 2 shows the general steps involved in MFCC calculation.

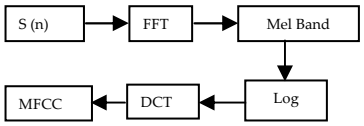

Figure 2. Step by step MFCC feature extraction.

4.2 Gaussian Mixture Model (GMM)

In the Gaussian mixture model classifier, each class probability density function is assumed to consist of a mixture of a specific number of multidimensional Gaussian distributions, whose parameters are estimated using the training set. Gaussian classifier is a typical parametric statistical classifier, assuming a particular form for the class probability density functions. GMM is the most widely used clustering method and is the one based on learning a *mixture of Gaussians.* Mixture model with high likelihood tends to have the traits: component distributions have high "peaks" (data in one cluster are tight); the mixture model "covers" the data well (dominant patterns in the data are captured by component distributions).

The Gaussian mixture architecture estimates probability density functions (PDF) for each class, and then performs classification based on Bayes' rule:

$$P(C_i \mid X) = P(X \mid C_i).\frac{P(C_i)}{P(X)} \qquad (1)$$

where $P(C_i|X)$ is the PDF class j, evaluated at X, $P(C_i)$ is the prior probability for class j, and $P(X)$ is the overall PDF, evaluated at X. Unlike the unimodal Gaussian architecture, which assumes $P(X|C_j)$ to be in the form of a Gaussian, the Gaussian mixture model estimates $P(X \mid C_j)$ as a weighted average of multiple Gaussians.

$$L_j = \prod_{i=0}^{N_{train}} P(X_i \mid C_j) \qquad (2)$$

where w_k is the weight of the k-th Gaussian G_k and the weights sum to one. One such PDF model is produced for each class. Each Gaussian component is defined as:

$$G_k = \frac{1}{(2\pi)^{n/2} \mid V_k \mid^{1/2}} \cdot e^{[-1/2(X-M_k)^T V_k^{-1}(X-M_k)]} \qquad (3)$$

where M_k is the mean of the Gaussian and V_k is the covariance matrix of the Gaussian.

Free parameters of the Gaussian mixture model consist of the means and covariance matrices of the Gaussian components and the weights indicating the contribution of each Gaussian to the approximation of $P(X \mid C_j)$. These parameters are tuned using a complex iterative procedure called the estimate-maximize (EM) algorithm, that aims at maximizing the likelihood of the training set generated by the estimated PDF. The likelihood function L for each class j can be defined as:

$$\ln(L_j) = \sum_{i=0}^{N_{train}} \ln(P(X_i \mid C_j)) \qquad (4)$$

5. Proposed Method

A Matlab based music emotion classification method has been developed. There are four main parts in proposed method. They are preprocessing, feature extraction, modeling, and classification.

5.1 Preprocessing

In preprocessing, songs having 44 kHz mp3 stereo format are converted to a uniform format (22 kHz sampling rate, 16 bits resolution, and mono channel .WAV). In Figure 3, preprocessing step is illustrated with a block diagram. Chorus part of each music file is manually trimmed. A music database is used, which contains 60 popular songs from Western albums and each song is annotated by 40 participants.

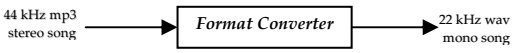

Figure 3. Block diagram of preprocessing.

5.2 Feature Extraction

After preprocessing, it is followed by feature extraction. Mel Frequency Cepstral Coefficients

(MFCC) features are applied in this approach. A total of 13 MFCC coefficients are extracted from music frames of size 20 ms (440 samples). Here music frames from chorus part are overlapped by 50%. To have compact feature dimension, each song is represented with means of 13 MFCC coefficients calculated over chorus frames.

5.3 Modeling

This step generates the classification according to the features of the training samples. Once cepstral features have been extracted from chorus part of each song, they are taken as input parameter of GMMs. To learn the parameter of music emotion, 10 songs have been collected for each quadrant to use as training samples. Figure 4 depicts how music emotion models are generated. Popular Expectation-Maximization (EM) algorithm is used to find the mixture of Gaussians.

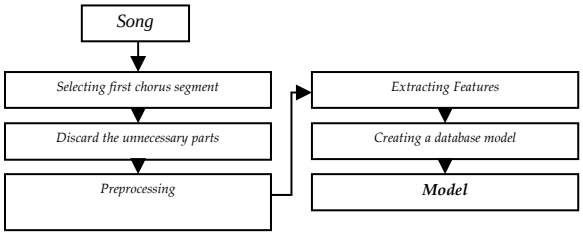

Figure 4. The block diagram of modeling.

5.4 Classification

The final output of the proposed system is music emotion in four phases. The four portions of test data classification are in 4 quadrants as Thayer's model is used. An hierarchical classification architecture is used. First, a GMM model is applied to differentiate whether the particular song is high energetic or low energetic song. Low energetic songs are further classified into Quadrant I and Quadrant II according to the model obtained above. Similarly, high energetic songs are further classified into Quadrant III and Quadrant IV as shown in Figure 5.

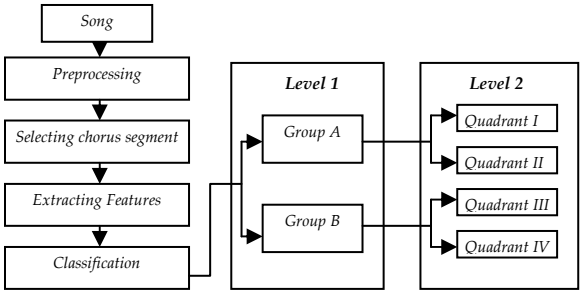

Table 1: Quadrant-wise Classification Results

	Quadrant I	Quadrant II	Quadrant III	Quadrant IV
Quadrant I	100%	0%	0%	0%
Quadrant II	0%	100%	0%	0%
Quadrant III	0%	0%	100%	0%
Quadrant IV	0%	0%	0%	100%

Figure 5. The block diagram of classification.

6. Experiments

To study the performance of the proposed method, experiments are carried out on the popular English song data set which is also used in [9]. It consists of songs from various genres such as rock, pop, classical, country, alternative, punk, jazz, Christian and gospel, metal and acoustic. In this data set, 15 songs are contained in the each quadrant. From each quadrant, chorus music clips extracted from 10 songs and used as training data and the rest 5 songs are applied as test data. There is no overlapping in training and testing songs.

The classification results are computed using the percentage of songs which is correctly grouped into their respective quadrants. Table 1 shows the results achieved by the proposed music emotion classification method. As it can be seen, no test songs are designated into wrong quadrants and thus achieving high classification accuracy.

7. Discussion

The system used a labeled Western song database and it may be difficult to have annotated songs by a group of participants. In the earlier stage of this work, the system has been trained with initial 25 seconds of song leading to very low classification accuracy in all quadrants. The system is more suitable for using trimmed chorus part of the song resulting higher accuracy in each group. Here, only cepstral features (MFCC) is used to learn the model of music emotion, but investigation of including other features should be done. As hierarchical classification style is used, number of Gaussian Mixture Models to be trained is also reduced.

8. Conclusion

In this paper, a method for music emotion classification is developed. The method adopts Thayer's mood taxonomy model containing 3

adjectives in each quadrant. Cepstral coefficients are found to be useful features for music emotion detection task. Classification is designed using a hierarchical structure of Gaussian Mixture Models. From the preliminary experiments, the overall accuracy gained on test songs for all adjective groups are promising. Due to the restriction of available resources, the proposed method is validated with a small data set. To find out the acoustic features which have relationship with music emotion and to perform more experiments with a specific music genre is the one possible direction to be done in future.

9. References

[1] R. E. Thayer, "*The Biopsychology of Mood and Arousal*," Oxford University Press, 1989.

[2] C. Laurier, O. Lartillot, T. Eerola, P. Toiviainen, "Exploring Relationships between Audio Features and Emotion in Music," *Conference of European Society for the Cognitive Sciences of Music*, 2009.

[3] Y. H. Yang, Y. C. Lin, Y. F Su, and H. Homer, "*Music Emotion Classification: A Regression Approach*," *International Conference on Multimedia and Expo 2007*, pp. 208-211, 2007.

[4] D. Baum, "EmoMusic – Classifying Music According to Emotion," *International Workshop on Self-Organizing Maps (WSOM'07)*, 2007.

[5] Y.-H. Yang,C.-C Liu, and H.-H. Chen, "Music emotion classification: A fuzzy approach," *ACM MM*, pp. 81–84, 2006.

[6] T. -Lin Wu and S.-K. Jeng, "Extraction of segment of significant emotional expressions in music," *2006 International Workshop on Computer Music and Audio Technology*, pp. 76–80, 2006.

[7] J. A. Russell, "A circumplex model of affect ." *Journal of Personality and Social Psychology*, 39, 1980, pp. 1161-78.

[8] Y.-H. Yang, C.-C Liu, and H.-H. Chen, "*Personalized Music Emotion Recognition*," *SIGIR'09*, 2004.

[9] Y.-H. Yang, Y.-F. Su , Y.-C. Lin and H.-H. Chen, "Music Emotion Recognition: The Role of Individuality," *Proc. ACM Int. Conf. HCM'07*, pp. 13-22, 2007.