

Detecting and Classifying Emotion in Popular Music

Chia-Chu Liu, Yi-Hsuan Yang, Ping-Hao Wu, and Homer H. Chen

Graduate Institute of Communication Engineering, National Taiwan University
b90901034@ntu.edu.tw, b91901109@ntu.edu.tw, b89901043@ntu.edu.tw, homer@cc.ee.ntu.edu.tw

Abstract

Music expresses emotion. However, analyzing the emotion in music by computer is a difficult task. Some work can be found in the literature, but the results are not satisfactory. In this paper, an emotion detection and classification system for pop music is presented. The system extracts feature values from the training music files by PsySound2 and generates a music model from the resulting feature dataset by a classification algorithm. The model is then used to detect the emotion perceived in music clips. To further improve the classification accuracy, we evaluate the significance of each music feature and remove the insignificant features. The system uses a database of 195 music clips to enhance reliability and robustness.

Keywords: Music emotion, Emotion Classifier

1. Introduction

Music plays an important role in our daily life. It is often played in ceremonies and religious occasions. The influence of music becomes more widespread as we enter the digital world. Due to the growing capacity of mass storage, larger music databases become ubiquitous in personal computers, MP3 players, and other modern devices. As the music databases grow, more efficient organization and search methods are needed.

Traditional music classification uses tags such as singers, authors, or the names of the songs. As opposed to the abundant traditional music classification methods, only a few content-based classification approaches have been developed, one of which discriminates music by genre [1], [2]. Music genre can be partitioned into several types, such as classical, rock, jazz, popular, country, folk, and so on. Genre classification can achieve an overall accuracy higher than 90%. Another content-based approach classifies music by the perceived emotion. This is much harder than the genre approach because of the following reasons: First, the perceived emotion of music is very subjective. Listening mood, environment, personality, cultural-background and so on can have influence on perceived emotion. Second, the

adjectives describing emotion may be ambiguous. For instance, a “happy” song or a “jovial” song may refer to the same thing. Third, it is inexplicable how music arouses emotion. What intrinsic quality of music, if any, creates a specific emotional response in the listener is still far from well-understood.

Music emotion detection and classification has been studied before, most of them adopted pattern recognition approach. Wang et al. [3] extracted features from MIDI files and used a support vector machine (SVM) to classify music into 6 classes: joyous, robust, restless, lyrical, sober, and gloomy. High classification accuracy was reported; however, one can not easily transcribe real world music into symbolic form, as done in MIDI files. Li et al. [4] divided emotion into 13 categories and combined them into 6 classes. Then, they adopted MARSYAS [5] in their system to extract music features from acoustic data and used SVM to train and recognize music emotion. Liu et al. [6] presented a hierarchical mood recognition system, which uses a Gaussian mixture model (GMM) to represent the feature dataset and a Bayesian classifier to classify music clips. However, these methods are not especially designed for pop music. To our best knowledge, the work described here represents the first attempt to address the problem.

The paper is organized as follows. In Section 2, we introduce the taxonomy of emotion used in our work. Section 3 gives an overview of the proposed system, and Section 4 describes a new algorithm that improves the classification accuracy. Experimental results are in Section 5, and conclusions in Section 6.

2. Taxonomy

Humans feel music by the sense of hearing. Different melody contours, pitch scales, tempi, and loudnesses result in different emotional responses. For these reasons, efforts have been made to establish a universal standard to describe emotion. There are numerous works in psychology that try to classify emotions. The earliest approach is by Hevner [7] who came up with an adjective checklist that contains 67 adjectives. Many new approaches have been developed lately. We adopt Thayer’s model [8] shown in Fig. 1.

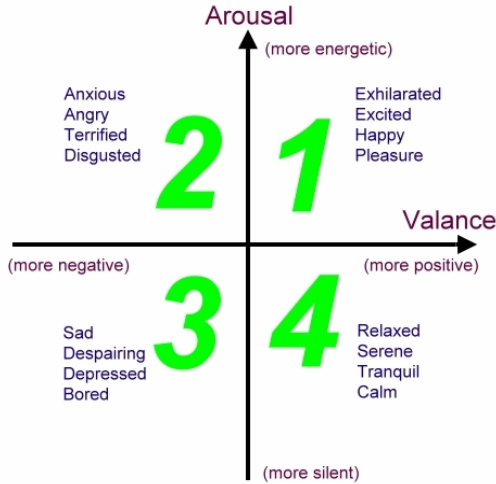


Fig. 1: Thayer's model of mood.

In Fig. 1, we divide the 2-dimensional emotion space (2DES) into 4 quadrants and place different emotions on the plane so that each emotion (a point in 2DES) can be represented by a 2x1 vector. In this paper, we only concentrate on the quadrant the point of interest falls in. The right side of the plane (quadrants 1 and 4) refers to the positive emotion, while the left side (quadrants 2 and 3) refers to the negative emotion. The points representing energetic emotions are located on the upper half plane (quadrants 1 and 2), while the points representing silent emotions fall on the lower half plane (quadrants 3 and 4). To be consistent with the 2DES model, we define 4 emotion groups, each corresponding to a quadrant.

3. System Overview

The proposed system can be divided into two stages: the model generator (MG) and the emotion classifier (EC). The MG generates a model according to the features of the training music clips, while the EC applies the resulting model to classify the test music clips. The details of the system are described in the following sub-sections.

3.1. Preprocessing

The preprocessing procedure is shown in Fig. 2. Music clips are downsampled to 22,050 Hz, 16 bit, mono channel PCM signals from files with CD-quality format. Preprocessing is performed in both the MG and the EC stages.

3.2. Model generator

Fig. 3 shows the architecture of the MG. First, we collect a large number of popular songs and choose a 25-second segment from each song as the initial music

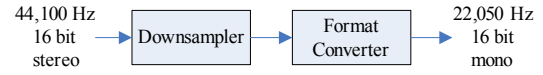


Fig. 2: Flow chart of pre-processing.

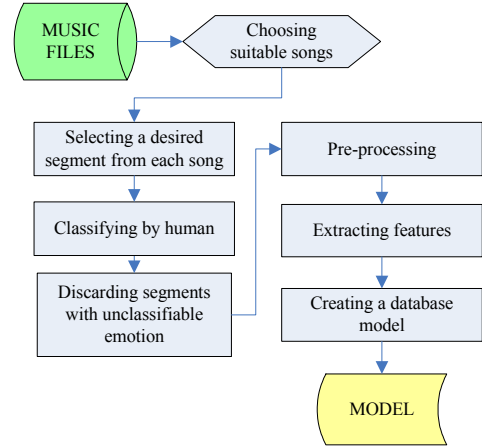


Fig. 3: Flow chart of the model generator.

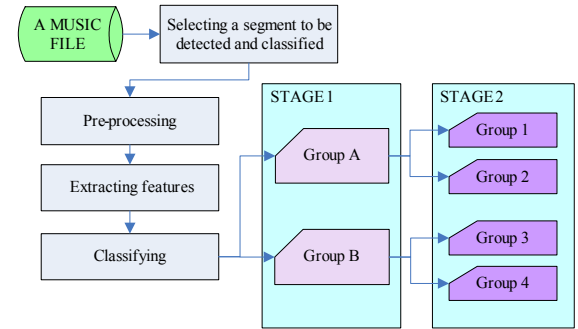


Fig. 4: Flow chart of the emotion classifier.

dataset. Next, volunteers are asked to classify the songs subjectively. If less than half of the subjects have the same emotional response (group 1, 2, 3, or 4) to a song segment, this segment is discarded.

The remaining segments are preprocessed, as described in Section 3.1, and saved in WAV format. Then, we extract music features using PsySound2 [9]. Among the numerous features, 15 are chosen based on the recommendation of [10]. We put the segments into the groups voted by the subjects and calculate the mean of each feature of the four groups by using the following equation

$$\mu(g, f) = \frac{1}{N_g} \sum_{n=1}^{N_g} F_{g,f,n} \quad (1)$$

where $\mu(g, f)$ is the mean of feature f ($f=1,2, \dots, 15$) in group g ($g=1,2,3,4$), $F_{g,f,n}$ is the value of feature f of the n th segment in group g , and N_g refers to the total number of segments in group g . These mean values form the core of the model and are an integral part of the EC.

3.3. Emotion classifier

As shown in Fig. 4, we start by choosing a music segment (denoted as X) with an unknown emotion and preprocessing it according to the procedure described in Section 3.1. Next, 15 features are extracted from X using PsySound2. Then, we choose the Nearest-Mean classifier to be our emotion classifier because of its outstanding experimental results among several classifiers.

In the Nearest-Mean classifier, we compute the sum of the squared error (SSE) between the feature values of X and the mean of each group. The group whose mean has the minimum SSE is the group to which X is assigned. That is, X is classified as:

$$G(x) = \{g \mid \min(\sum_{f=1}^F (X_f - \mu(g, f))^2), g \in \{1, 2, \dots, G\}\}, \quad (2)$$

where $G(X)$ denotes the predicted group of X and X_f is the value of feature f of X. F is the total number of features, and G is the total number of groups.

In our hierarchical Nearest-Mean classifier, the means of a feature f in group A and B are obtained as follows:

$$\begin{aligned} \mu(A, f) &= 0.5 \cdot [\mu(1, f) + \mu(2, f)], \text{ and} \\ \mu(B, f) &= 0.5 \cdot [\mu(3, f) + \mu(4, f)], \end{aligned} \quad (3)$$

where the notation is the same as those in Eq. (1). The SSEs between the feature values of X and the means of groups A and B are calculated. Then X is assigned to either group A or group B by:

$$G_1(X) = \{g \mid \min(\sum_{f=1}^{15} (X_f - \mu(g, f))^2), g \in \{A, B\}\}. \quad (4)$$

So far we have determined whether X belongs to the upper or the lower plane of the 2DES. Next, if X belongs to group A, we classify it into either group 1 or group 2 by:

$$G_{2a}(X) = \{g \mid \min(\sum_{f=1}^{15} (X_f - \mu(g, f))^2), g \in \{1, 2\}\}; \quad (5)$$

otherwise, it is assigned to group 3 or 4 by:

$$G_{2b}(X) = \{g \mid \min(\sum_{f=1}^{15} (X_f - \mu(g, f))^2), g \in \{3, 4\}\}. \quad (6)$$

Fig. 5 shows the projection from the original features to the first two principal components. Based on the distribution of the points, we can expect that (a) it is easy to separate group A from group B, therefore we adopt the hierarchical approach, (b) it is easier to separate group 1 from group 2 than it is group 3 from group 4, and (c) the songs in group 2 are correctly classified more easily than those of group 1 due to the denser point distribution, which reduces the distances between the points and their mean and enhances

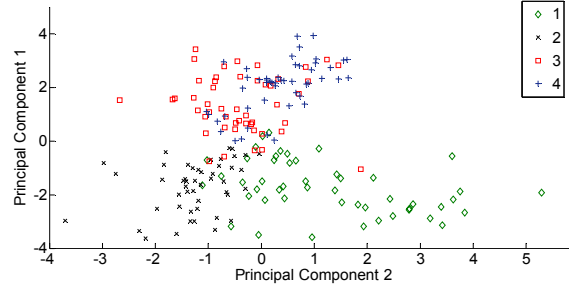


Fig. 5: Transformation into the navigation space by PCA

accuracy. The observations (a) and (b) are the reason why we design the hierarchical architecture in order to minimize the classification error.

4. Accuracy Enhancement

To enhance the classification accuracy, we propose an optimization algorithm based on the observation that not all 15 features are vital for classification. Therefore, we may discard the irrelevant features to improve results. This is done in an iterative fashion. Take Type 1 EC as an example. First, we obtain the accuracy Acc_0 using all 15 features. Then, we remove a feature, t ($t=1, 2, \dots, 15$), from the system and perform all the procedures of MG and EC; that is, 14 features are used to generate model and to classify the segments. We denote $R(X)$ as the voted answer of the group that X belongs to and define $G(X, t)$ and $Acc(c, t)$ as follows:

$$G(X, t) = \{g \mid \min(\sum_{f=1, f \neq t}^{15} (X_f - \mu(g, f))^2), g \in \{1, 2, 3, 4\}\}, \quad (7)$$

$$Acc(c, t) = \frac{\{\# X \mid G(X, t) = R(X) = c\}}{\{\# X \mid R(X) = c\}}. \quad (8)$$

$Acc(c, t)$ denotes the overall accuracy after removing feature t from the system. If we can find a $t=t_0$ such that $Acc(c, t_0)$ is higher than Acc_0 and $Acc(c, t_0)$ is larger than any $Acc(c, t)$, for $t=1, 2, \dots, 15$ and $t \neq t_0$, we claim feature t_0 as a “weak” feature and discard it from the system. Next, we test the removal of another feature of the system again (13 features left now) and examine whether we can improve the accuracy by removing this feature. This procedure is repeated until there is no more improvement.

5. Experiments and Results

We collected 243 popular songs from Western, Chinese, and Japanese albums and chose a 25-second segment from each song. Seven subjects were asked to label the emotion of the segments. After removing the unclassifiable ones, as explained in Section 3.2, we

obtained 195 segments (listed in Table 1) with their group of emotion annotated by the subjects.

The classification results were computed using a cross-validation technique. The whole dataset was randomly divided into 10 parts. 90% of the data were used for generating a model, and the remaining 10% were used for testing. The above process was repeated 50 times and then the accuracies were averaged.

Table 2 shows the results produced by MC + EC. The overall accuracy is 71.95%. The results are consistent with our expectations in Section 3.3: (a) A low percentage of the songs in group A are classified into group B (6.12% from group 1 and 3.33% from group 2 to groups 3 and 4, respectively), and vice versa (9.28% from group 3 and 1.42% from group 4 to groups 1 and 2, respectively), (b) The accuracies of both groups 1 and 2 are higher than the accuracies of groups 3 and 4, and (c) The accuracy of group 2 is the highest of all.

After performing the optimization algorithm described in Section 4, features 4, 5, 3, 13, and 9 are removed one after another from the initial 15 features resulted by the iterative feature pruning. The final results are shown in Table 3, and we can see the great improvement in overall accuracy (from 71.95% to 78.97%, the Acc(avg) curve). The discarded features are associated with sharpness (2 features), timbral width, tonal dissonance, and multiplicity respectively.

6. Conclusions

In this paper, we have proposed an emotion detection and classification system for pop music. The system is designed using a hierarchical framework followed by an accuracy enhancement mechanism. The experimental results show that the system gives satisfactory performance. The system is developed in software and works automatically. Furthermore, the system aims at popular music, so it can be applied to public music database software to provide emotion-based search.

On the other hand, the features that affect the perception of emotion are associated with frequency centroid, spectral dissonance and .pure tonalness. We will find out the deeper relation between these features and music emotion.

7. Acknowledgements

The authors would like to thank Tien-Lin Wu for providing equipment and software for this work and the volunteers for participating in the experiments.

This work was supported in part by grants from Intel and the National Science Council of Taiwan under contracts NSC 94-2219-E-002-016 and NSC 94-2725-E-002-006-PAE.

Table 1. Distribution of segments

	Group 1	Group 2	Group 3	Group 4
# segment	49	48	49	49
Total	195			

Table 2. Results of EC (before optimization)

	Group 1	Group 2	Group 3	Group 4
Group 1	70.89%	22.98%	6.12%*	0%
Group 2	6.25%	90.42%	3.33%*	0%
Group 3	2.04%	7.24%	61.12%	29.60%
Group 4	0%	1.42%*	32.82%	65.76%
Overall	71.95%			

Table 3. Results of EC (after optimization)

	Group 1	Group 2	Group 3	Group 4
Group 1	74.59%	20.29%	5.12%*	0%
Group 2	4.27%	95.73%	0%*	0%
Group 3	2.04%	6.06%*	73.98%	17.92%
Group 4	0%	0%	28.19%	71.91%
Overall	78.97%			

8. References

- [1] C. Xu, N. C. Maddage, and X. Shao, "Automatic Music Classification and Summarization," IEEE Trans. on Speech and Audio Processing, vol. 13, no. 3, pp. 441-450, May 2005.
- [2] C. Xu, N. C. Maddage, and X. Shao, "Automatic Music Classification and Summarization," IEEE Trans. on Speech and Audio Processing, vol. 13, no. 3, pp. 441-450, May 2005.
- [3] K. Umopathy, S. Krishnan, and S. Jimaa, "Multigroup Classification of Audio Signals Using Time-Frequency Parameters," IEEE Trans. on Multimedia, vol.7, no.2, pp. 308-315, Apr. 2005.
- [4] M. Wang, N. Zhang, and H. Zhu, "User-adaptive Music Emotion Recognition," IEEE, International Conference on Signal Processing, pp. 1352-1355, 2004.
- [5] T. Li and M. Ogihara, "Detecting Emotion in Music," Johns Hopkins University, 2003.
- [6] Marsyas, <http://opihi.cs.uvic.ca/marsyas>.
- [7] D. Liu, L. Lu, and H.-J. Zhang, "Automatic Mood Detection from Acoustic Music Data," Johns Hopkins University, 2003.
- [8] K. Hevner, "Expression in Music: A Discussion of Experimental Studies and Theories," Psychological Review, vol. 42, pp 186-204, 1935.
- [9] R. E. Thayer, The Biopsychology of Mood and Arousal, Oxford University Press, 1989.
- [10] PsySound, <http://members.tripod.com/~densil/>.
- [11] E. Schubert, "Measurement and Time Series Analysis of Emotion in Music," Ph. D. thesis, University of New South Wales, 1999.