THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.2E 2007. 6 pp. 56~62

Music Emotion Classification Based On Three-Level Structure

Hyoung-Gook Kim*, Jinguk Jeong**
*Intelligent Multimedia Signal Processing, Kwangwoon University
**Samsung Advanced Institute of Technology

(Received May 9 2007; Accepted Jun 18 2007)

Abstract

This paper presents the automatic music emotion classification on acoustic data. A three-level structure is developed. The low-level extracts the timbre and rhythm features. The middle-level estimates the indication functions that represent the emotion probability of a single analysis unit. The high-level predicts the emotion result based on the indication function values. Experiments are carried out on 695 homogeneous music pieces labeled with four emotions, including pleasant, calm, sad, and excited. Three machine learning methods, GMM, MLP, and SVM, are compared on the high-level. The best result of 90.16% is obtained by MLP method.

Keywords: *automatic music emotion classification; timbre and rhythm features; machine learning methods*

I. Introduction

Music is perceived historically and pervasively as an important carrier of human emotions. Although there is still no consensus on how music expresses emotions or produces emotional effects in listeners; how to define emotions; how to evaluate the emotional states of listeners; and other difficult issues in musicology and psychology, there is solid empirical evidence from psychological research that listeners from the same culture often agree rather strongly about what type of emotion is expressed in a particular piece. Then our question is that whether a computer can do so by means of machine learning.

In this paper we define the automatic music emotion classification as to classify a piece of music into one of the predefined emotions automatically. The attempt is to evaluate the possibility of "perceiving" emotional state of music only from acoustic signals automatically by a computer and to evaluate the effectiveness of the relevant algorithms to this task rather than to give other systematic conclusion in the music psychological research.

Firstly, the basic issue is the emotion types to classify with music. Typically Russell decomposes emotions along a "valence" dimension from negative to positive and an "arousal" dimension from inactive to active [1]. Figure 1 gives an illustration of the two dimensions and the basic emotions.

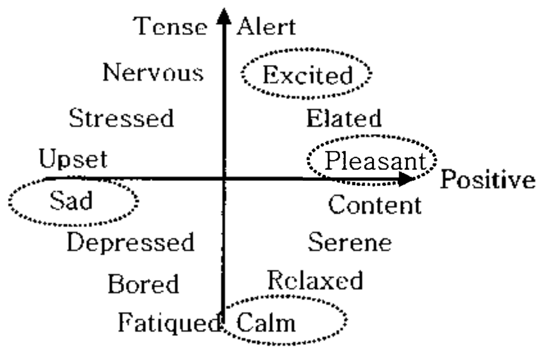


Fig. 1. Dimensional map of basic emotions.

Corresponding author: Hyoung-Gook Kim (hkim@kw.ac.kr)
Dept., of Radio Science and Engineering, Kwangwoon University, 447-1, Wolgye-Dong, Nowon-Gu, Seoul, 139-701

Among the basic emotions four emotions are selected. They are calm, sad, pleasant, and excited. Our criterion is to only select the relatively consistent and widely accepted emotions in music. So we select the four emotions distributing on the corners of dimensional map separately. Furthermore, we substitute excited to tense since the former is easier to be labeled consistently by different listeners.

Secondly, how to get the ground truth data of the emotions expressed in music? Here the verbal self-report criterion is adopted, that is, listeners are asked to describe that the music piece is supposed to indicate one of the emotions or none of them in response to different genres of music. 5 females and 5 males in the ages of 20~35 attend it in an ordinary office cubic via earphone listening to music played by a computer. Finally 695 homogeneous pieces are labeled from hundreds of western classical, dance, march, jazz, electronic, popular, and rock, with the average length of 3 minutes. Among them, 286 pieces (68/calm, 59/excited, 85/pleasant, 74/sad) are labeled consistently and are used as training data; 409 pieces (100/calm, 100/excited, 107/pleasant, 102/sad) are labeled as testing data.

After the ground truth data are collected, we attempt to computationally model the underlying mechanism of emotion classification by three computational levels. Figure 2 illustrates the system framework.

The low-level extracts the relevant features from the music signals. It is suggested in [2] that the structural features, performance features, states or traits of listeners, and relative context are four types of important input variables. Among them the structural features are ranked the most important factors and are subdivided into segmental and suprasegmental types. Segmental features consist of the acoustic characteristics, which are described by the duration, energy, pitch, and timbre or harmonic structure of the tones. The segmental effects on emotion are relatively stable and universal. Suprasegmental features in music are melody, tempo, rhythm, harmony, and other aspects of musical structure and form. So basically the above features should be extracted from the complex waves and evaluated in an emotion classification system. Here three sets of features, representing energy, timbre, and rhythm, are adopted as the low-level features.

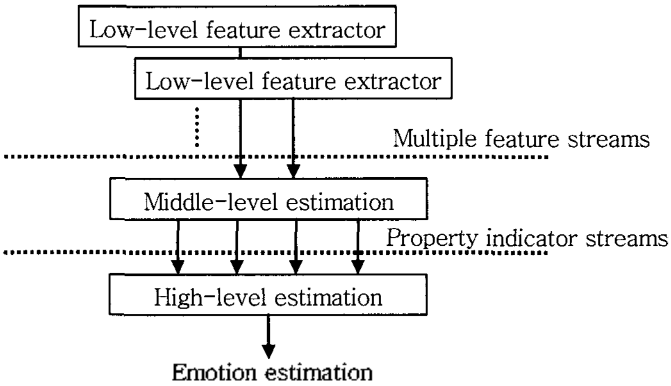


Fig. 2. System framework.

The middle-level aims at estimating the indication functions of those properties that are defined to bridge the semantic gap between the low-level features and the high-level concepts. For emotion classification, the instrument type, the timbre property from dark to bright and the tempo property from slow to fast, not exclusively here, can all be considered as the properties linked between the low-level and the high-level. The a posterior probability of each property can be estimated from the low-level features by means of Bayesian theory. In this paper, the middle-level property is simply defined as the emotion indicator of each analysis window. Since the analysis window is always much shorter than the duration of the given piece of music, the middle-level here can be seen as a linker between the low and high levels in time.

The high-level aims at estimating the final result based on the middle-level indication functions. Various machine learning approaches can be adopted here. For example, if the middle-level data only compose a small sample set, support vector machine (SVM) is a good classifier for it. Bayesian network can be used, depending on the dependence assumptions on the middle-level properties. If the properties are modeled by hidden states with causality in time, hidden Markov model (HMM) solves the estimation problem well. In this paper, Gaussian mixture model (GMM), neural network, and SVM are evaluated on this level.

The previous work can be found in [3] and [4]. [3] uses intensity, timbre, and rhythm features and GMM classifiers to recognize the 4 emotional states of exuberance, anxious, contentment, and depression. The detailed timbre features include centroid, bandwidth, roll off, flux, and sub-band spectral contrast features. And the rhythm features are described by strength, stability, and ratio of onsets. [4] tries to recognize the 4 emotions of happiness, sadness, anger, and fear by only using 3 features, that is, mean and variance of the silence ratio and the beat rate estimated by a beat-tracking algorithm.

Unlike the previous work, this paper proposes a three-level computational structure, employs a perceptual rhythm feature, which can be fused with other streams on each level, and evaluates them on a larger corpus.

Section II describes the rhythm feature extraction process in detail. Section III gives the middle-level estimation process. Section IV explains the three machine learning methods on the high-level. Experimental results of the middle-level and the high-level estimations are presented in Section V. And Section VI draws the conclusion.

II. Feature Extraction

Energy, timbre and rhythm features are extracted from the music signals. Energy and timbre are analyzed in a short-term window, 20~30 milliseconds typically. The timbre features presented in [3] are adopted here.

In this paper we mainly describe the rhythm feature extraction process. This rhythm feature aims at encoding the rhythm information into a spectrum whose coefficients represent the strength at the dynamic periodicities along the time axis. It is based on the modulation spectrum estimation. For computation efficiency, the modulation spectrum is obtained by applying Fourier transform on sub-band filtered signals using a long analysis window.

For a standard MP3 file with the properties of 44100Hz, stereo and 128kbps, the proposed feature is extracted directly on MP3 files as follows:

<Step 1> Partially decode the MP3 file to narrow-band-pass filtered samples $\{s^n(t)\}$, $n \in [1,576]$. n is the index of sub-bands. 576 sub-bands can be obtained on the intermediate level of decoding process. The bandwidth of each sub-band is about 38Hz.

<Step 2> Full wave rectification followed by a low-pass filtering is a typical envelope detection method. The smoothed envelope is computed by using recursively smoothed periodograms. Here, a one-pole filter with α=0.96875 is adopted to obtain a smoothed envelope. The advantage of recursive smoothing is its computational simplicity and the fact that no measurement delay is introduced.

$$\bar{s}^n(t) = \left|s^n(t)\right| \tag{1}$$

$$x^n(t) = (1-\alpha)\bar{s}^n(t) + \alpha x^n(t-1), \quad 0.95 \le \alpha < 1 \tag{2}$$

<Step 3> The deviation between two time-adjacent filtered samples is employed for emphasizing signal variation.

$$\hat{x}^n(t) = x^n(t) - x^n(t-1) \tag{3}$$

<Step 4> Long-term Fast Fourier Transform (FFT) is applied on hamming windowed deviation signals. The analysis window is 13.4 seconds (512 samples) and the window shift is 1 second (38 samples). For the i^{th} frame, the k^{th} sample is

$$y_i^n(k) = \hat{x}^n(38 * i + k), \quad 0 \le k < 512 \tag{4}$$

The result of FFT is described as in Equation (5), where $w(k)$ is the weight of hamming window.

$$Y_i^n(m) = \sum_{k=0}^{512} w(k) \cdot y_i^n(k) \cdot e^{-j2\pi km/512}, \quad 0 \le m < 512 \tag{5}$$

<Step 5> The power of modulation spectrum is smoothed by a log-scale triangular filter-band $\{H(p,m) | 1 \le p < 12,\ 0 \le m < 256\}$, where the p^{th} filter is given by

$$H(p,m) = \begin{cases} 0 & m < f(p-1) \\ \dfrac{2(m - f(p-1))}{(f(p+1) - f(p-1))(f(p) - f(p-1))} & f(p-1) \le m \le f(p) \\ \dfrac{2(f(p+1) - m)}{(f(p+1) - f(p-1))(f(p+1) - f(p))} & f(p) \le m \le f(p+1) \\ 0 & m > f(p+1) \end{cases} \tag{6}$$

and $\{f(p) | 0 \le p < 13\}$ is the center frequency of each filter, which increases logarithmically. Then, the 12-order dynamic feature is calculated as the log-energies at the output of filter-bank:

$$c_i^n(p) = \ln\left(\sum_{m=0}^{256} \left|Y_i^n(m)\right|^2 H(p,m)\right), \quad 1 \le p \le 12 \tag{7}$$

For each frame, dynamic rhythm features on the lowest 5 sub-bands are extracted to constitute 60-dimensional feature vector.

After the energy, timbre, and rhythm features are extracted, the energy value and the timbre vector are connected into one feature vector. Then a principle component analysis (PCA) is used to extract the uncorrelated feature coefficients, from the timbre feature vector and the rhythm feature vector respectively.

III. Middle Level Estimation

The middle-level aims at estimating the indication functions that represent the probabilities of emotion types with each observation of low-level features. It can be solved by Bayesian theory. With the assumption of equal prior probabilities of emotion types, the a posterior probabilities are reduced to the observation likelihood, which can be modeled by GMM. So the GMM of each emotion type on the low-level features are trained by EM algorithm.

$$\Theta(x \mid i) = \sum_j \lambda_j N(m_j, \Sigma_j), j = 1,2,\cdots,M \quad , \quad \sum_j \lambda_j = 1 \tag{8}$$

Here, x is the one frame of PCA transformed coefficients. j is the index of mixture. i indicates emotion.

The middle-level firstly estimates the likelihood of each emotion type with each observation frame,

$$L(x_s \mid i) = \log(\Theta(x_s \mid i)) \tag{9}$$

Here, s stands for different feature stream. The timbre stream and the rhythm stream are involved.

Secondly, the likelihood is transformed to a probability measure by the rule

$$P_{i,s} = \frac{\exp(L(x_s \mid i))}{\sum_i \exp(L(x_s \mid i))} \tag{10}$$

Thus the probability measure is in the range of (0,1) and facilitates the high-level processing.

IV. High Level Estimation

The high-level aims at estimating the emotion type of the given piece of music based on the middle-level indication functions:

$$P_{i,s}(t), i \in \{calm, sad, excited, pleasant\}, s \in \{timbre, rhythm\} \tag{11}$$

GMM, multi-layer perception (MLP) network, and SVM are compared on this level.

For GMM, the 8 values of P at the same time t are connected into one feature vector. Then the high-level GMM is trained by means of EM algorithm. In order to obtain the sufficient training data, as well as to synchronize the indication functions, the rhythm feature is extracted with the same analysis shift with that of the timbre feature during the training process. In testing the rhythm feature and the timbre feature can be extracted with much longer analysis shift to increase the classification speed without the loss of precision.

For MLP, the input is the same with that of high-level GMM. The MLP is trained by means of back propagation (BP) algorithm.

SVM is originally designed for binary classification. Hereit is extended for K-class classification by constructing $K(K-1)/2$ "one-against-one" binary classifiers and combining them with the rule of majority voting. Each binary classifier is trained on data $\boldsymbol{W}_t$ (t=1~N, the number of training samples) from class i and class j by solving the following binary classification problem:

$$\min J(W,e,b) = \frac{1}{2}W^T W + C\sum_t e_t \tag{12}$$

subject to

$\boldsymbol{w}^T \Phi(\boldsymbol{x}_t) + b \ge 1 - e_t, if \quad \boldsymbol{x}_t \in \{class\ i\}$

$\boldsymbol{w}^T \Phi(\boldsymbol{x}_t) + b \le -1 + e_t, if \quad \boldsymbol{x}_t \in \{class\ j\}$

$e_t \ge 0$

After training, the decision function is given by

$$sign(\boldsymbol{w}^T\Phi(\boldsymbol{X})+b) \tag{13}$$

In combining the decisions of the *K(K−1)/2* binary classifiers, the sample is assigned the class when at least k classifiers are agreed on the identity, where

$$k=\begin{cases}\frac{K+1}{2}, \textit{if K is odd}\\ \frac{K}{2}, \textit{if K is even}\end{cases} \tag{14}$$

V. Experimental Result

The timbre features are extracted every 23 milliseconds; while the rhythm features are extracted every 1 second. Then they are transformed to the low-level feature frames via PCA transformation. The middle-level GMMs are trained on the two sets of features separately. The likelihood is transformed to probability measures, which indicate the emotion property of each frame.

Experiments are carried out to evaluate the performance of the middle-level GMM. The optimal parameters, including the dimension of PCA and the mixture number of GMM, are selected empirically. As a comparison to PCA, linear discriminant analysis (LDA) is also adopted and evaluated. The frame likelihood is summarized to estimate the likelihood of the whole piece of music with respect to each emotion type. Then the emotion of the music piece is decided to be the one with the largest likelihood. The performance is measured by the classification precision of music pieces. Figure 3 illustrates the result.

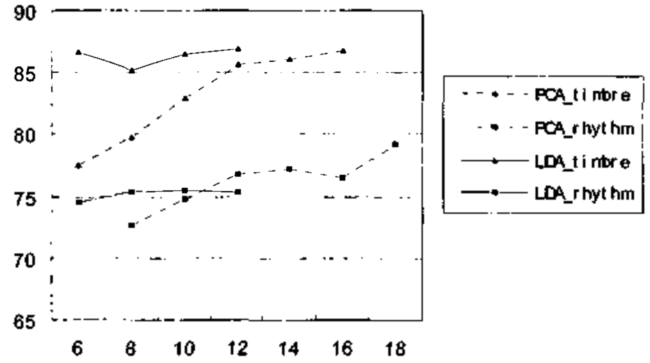


Fig. 3. Music emotion classification precision of middle-level GMM.

It is revealed clearly that with the same dimension of feature space LDA outperforms PCA by a great margin, because LDA aims at extracting the features which separate the classes to a maximal extent while PCA only focuses on the features which approximate the original feature space with the lowest mean-square-error and does not consider the discriminability of these features. But unlike PCA whose classification precision can be improved by means of augmenting more features, such a potential of LDA is quite limited, because the nature of LDA only permits *K*−1 orthogonal feature axes be spanned (*K* is the number of classes). Here we have increased the possible dimensionality of LDA by unsupervised-clustering the classes into more sub-classes. Or else only three features can be extracted and the performance is quite unsatisfied. So PCA is used for the reason of robustness.

The input of high level is the values of eight indication functions, four from the timbre stream and four from the rhythm stream. The values are input into the high-level module every one second. The high-level decision is made based on various learning methods.

In GMM mode, the output is the observation likelihood of indication function values with respect to each emotion type:

$L(P_t \mid i)$, $P_t = \{P_{i,s}(t)\}, i \in \{calm, sad, excited, pleasant\}, s \in \{timbre, rhythm\}$.

Then the final decision is made by accumulating the likelihood of each input P_t, with the assumption of independence of P_t each other:

$$\hat{i} = \arg_i\{\sum_t L(P_t \mid i)\}$$.

In MLP mode, the final decision rule is the same with that of GMM mode. Only the output of MLP substitutes P_t. That is the estimation of the a posterior probability of each emotion type with respect to the indication functions values.

The scheme of SVM is based on the majority voting rule. Since the collection of the binary classifiers has estimated the emotion type with respect to the indication function values every second, the final decision is made by selecting the emotion that counts the most.

Table 1 gives the high-level precision.

The best result of 90.16% on four emotions is obtained with MLP high-level estimation. But there is no

Table 1. Music emotion classifiction precision using different high-level method.

High-level methods	GMM	MLP	SVM
Precision	89.73%	90.16%	89.69%

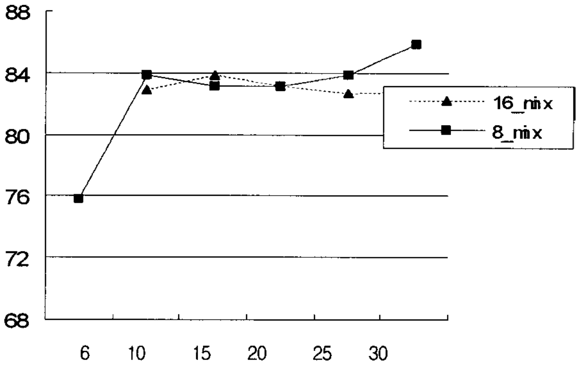


Fig. 4. Music emotion classification precision of feature fusion method.

significant difference found among the three classification methods. In the case of GMM, the process of music emotion classification is more simple and fast compared with SVM and MLP. Therefore, it could be suitable to use GMM in real-world applications, where training is done off-line and only the classification should be fast as possible for on line classification of music emotions.

In order to illustrate the errors more clearly, Table 2 gives the confusion matrix of GMM high-level estimation. The confusion matrices of the three learning methods are quite similar so others are not repeated here.

Firstly, the choice of different learning methods on the high-level has little effect on the classification precision. We think that the relatively simple design of indication functions has limited the potential of high-level learning methods. But even with such a simple fusion method as only subjects to the likelihood of timbre stream and rhythm stream, the precision has been improved to 90% approximately.

Secondly the confusion matrix reveals that the discriminability among the four types of emotions in music has been captured by the current features rather well. It can be seen that almost all of the calm songs are recognized correctly. Such is the same case with pleasant songs. The most two confusion sets are (a) many exciting songs are misclassified as pleasant and (b) many sad songs are misclassified as calm. After listening to the test songs, we find that all of the errors take place between the categories with similar tempo. In the case of (a), the error pieces of exciting songs sound like pleasant rock, noisy than the pleasant pop, but not as heavy and tense as the mental. In the case of (b), it is very difficult to distinguish the emotion of the error pieces precisely, because all of them are vocal with slow tempo and similar accompaniment with some calm songs. The major difference is that the vocal of calm songs is smooth and relaxed, while the vocal of sad songs sounds trembling a little. It is even suspected that several sad pieces are labeled based on the lyrics, because the Chinese listeners cannot tell the difference of them from some Korean popular music in the calm category. Clearly the differences of singing style and lyrics content cannot be represented by current features.

Table 2. Confusion matrix of GMM high-level estimation.

Emotion types	Calm	Sad	Pleasant	Exciting
Calm	98	1	1	0
Sad	23	78	1	0
Pleasant	1	2	102	1
Exciting	0	0	12	88

As a comparison to the high-level methods, the fusion of different information streams is also evaluated on the feature end, since the middle-level is carried out with the same time granularity of one second. The timbre features and the rhythm features of every one second are connected into one feature vector, on which the PCA transformation is done. And GMM is trained to give the final decision directly. So this method is the flat structure that fuses information in the feature space; while the three-level one is the stack structure that fuses information in the semantic space via modeling the middle-level concepts by machine learning methods. The result is given in Figure 4. The precision is plotted with different PCA transformation dimension and mixture number.

The best result of 85.82% is obtained with 30-dimensional features and 8 mixtures of each GMM. It is much worse than the results of stack structure fusion methods. In fact the result is compatible with the precision of timbre feature. So the fusion in feature space does not use the rhythm feature information appropriately.

VI. Conclusion

This paper introduces the work on automatic music emotion classification by means of machine learning methods. A three-level structure is designed to bridge the gap between the acoustic features and the affective concepts. The low level includes two acoustic feature extraction modules, one for timbre feature and the other for rhythm feature. The middle level is responsible for estimating the indication function values that reveal what middle-level properties exist or to what extent they exist. The high level learns the relationship between the middle-level properties and the emotions indicated.

Three machine learning methods, GMM, MLP, and SVM are compared on the high-level estimation stage. They fuse the likelihood of timbre and rhythm feature streams. All of them perform superior to the feature fusion method. But there is no significant difference found among the three. The best result of 90.16% on four emotions is obtained with MLP high-level estimation. In the case of GMM, the process of music emotion classification is more simple and fast compared with SVM and MLP.

Acknowledgment

The present research has been conducted by the Research Grant of Kwangwoon University in 2007.

References

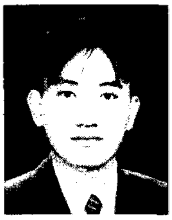
1. B.L. Feldman, and J.A. Russell, "Independence and bipolarity in the structure of affect," Journal of Personality and Social Psychology, 74, 967-984, 1998.
2. P.N. Juslin, and J.A. Sloboda, "Music and emotion: theory and Research," Oxford Univ. Press, 2001.
3. D. Liu, L. Lu, and H.J. Zhang, "Automatic mood detection from acoustic music data," ISMIR 2003.
4. Y.Z. Feng, Y.T. Zhuang, and Y.H. Pan, "Music information retrieval by detecting mood via computational media aesthetics," IEEE/WIC International Conf. on Web Intelligence 2003.

[Profile]

• **Hyoung-Gook Kim**

received the diploma degree in electronicengineering and the Ph. D. degree in computer science from the Technical University of Berlin, Berlin, Germany. From 1998 to 1999, he worked on mobile service robots at Daimler Benz AG and speech recognition at Siemens AG, Germany. From 2002 to 2005, he served as Assistant Professor of the Communication Systems Dept., Technical University of Berlin, Germany. From 2005 to 2007 he was a Project Leader in Samsung Advanced Institute of Technology, Korea. Since 2007 he has been a Professor in the Radio Science and Engineering Dept., Kwangwoon University, Korea. His research interests include audio signal processing, music information retrieval, audiovisual content indexing and retrieval, automatic segmentation, speech enhancement, and robust speech recognition.

• **Jinguk Jeong**



received his B.S., M.S. and Ph.D. in computer science from Sogang University, Korea, in 1998, 2000, and 2004, respectively. He isa researcher in Samsung. His research interest includes multimedia computing system, content-based multimedia indexing and retrieval algorithm, and MPEG compression standard.