Feature Selection Method for Music Mood Score Detection

Masato Miyoshi[1] Satoru Tsuge[2] Tadahiro Oyama[3] Momoyo Ito[1] Minoru Fukumi[1]

[1]The University of Tokushima, 2-1, Minami-josanjima, Tokushima, Tokushima, 770-8506, Japan
{m.miyoshi,momoito,fukumi}@is.tokushima-u.ac.jp

[2]Daido University, 10-3, Takiharu-cho, Minami-ku, Nagoya, Aichi, 457-8530, Japan
tsuge@daido-it.ac.jp

[3]Kobe City College of Technology, 8-3, Gakuen-higashimachi, Nishi-ku, Kobe, Hyogo, 651-2194, Japan
oyama@kobe-kosen.ac.jp

Abstract

In general, music retrieval and classification methods using music moods use a lot of acoustic features similar to music genre classification. These features are used as the spectral features, the rhythm features, the harmony features, and so on. However, all of these features may not be efficient for music retrieval and classification using music moods. Hence, in this paper, we propose a feature selection method for detecting music mood scores. In the proposed method, features which have strong correlation with mood scores are selected from a lot of features. Then, these are input into Multi-Layer Neural Networks (MLNNs) and mood scores are detected every mood labels. For evaluating the proposed method, we conducted the music mood score detection experiments. Experimental results show that the proposed method improves the detection performance compared to not use the feature selection.

Key words: Music retrieval and classification, Music mood, Mood score detection, Feature selection, Feature extraction

1 Introduction

In recent years, it has become possible to access a lot of music data because we can get music from distributions on the Internet and save these on large capacity portable digital audio players and personal computers. As a result, it is difficult to retrieve and classify these music data by hand. Hence, efficient music retrieval and classification methods have been proposed. These methods retrieve and classify the music by using a genre, an artist name, a mood, and so on. We consider that the music retrieval method by the moods is efficient for users because they can intuitively retrieve music. In addition, the music retrieval method using the moods can retrieve the music without music information such as a genre, an artist name, and so on. In order to retrieve the music using music mood, it is necessary to set a suitable mood to each song and/or set a score of each mood to each song.

In general, in the retrieval/classification method using moods, we use a lot of acoustic features[1, 2, 3, 4] used in music genre classification[5, 6, 7]. As acoustic features, there are spectral features (e.g. Centroid, Rolloff, Mel-Frequency Cepstral Coefficients (MFCCs)), rhythm features (e.g. Power spectrum peaks, Beat histogram), and harmony features (e.g. Major components, Chromagram). However, all of these features may not be efficient for the music retrieval/classification using music moods. Some features might cause a degradation of the detection performance. Furthermore, we consider that there are suitable feature parameters for each mood. Therefore, in this paper, we propose a feature selection method for detecting the music mood scores. If the proposed method selects the efficient features, the mood score detection performance is improved.

This paper is constructed as follows. First, in section 2, we describe the details of the proposed method. In section 3, we conduct experiments to evaluate the proposed method. Finally, in section 4, we conclude this paper and describe future works.

2 Mood score detection using feature selection method

In this section, we describe the mood score detection using the proposed method. A flowchart of the mood score detection is shown in Figure 1. First, in the feature extraction, efficient features are extracted for mood score detection. Next, the proposed

978-1-4577-0005-7/11/$26.00 ©2011 IEEE

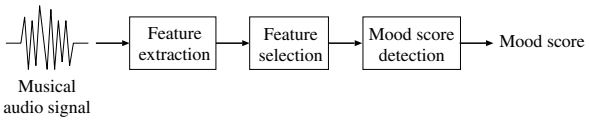

Figure 1: Flowchart of the mood score detection

method selects the extracted features using correlation information between the feature and the score of each mood. In this process, we can get new features for detecting the mood score. Finally, mood scores are detected by Multi-Layer Neural Networks (MLNNs). We describe the details of these processes in the following sections.

2.1 Feature extraction

In the feature extraction, features which are related to intensity, timbre, rhythm, and harmony are extracted from musical audio signals. We describe each feature as follows:

2.1.1 Intensity features

The following 4 features are used as intensity features:

- **Frame Energy**[3]
  Frame Energy indicates logarithm power of audio signals. We consider that the music with high Frame Energy is brightness and/or cheerful music.

- **Log Spectrum Sum and ΔLog Spectrum Sum**[4]
  Log Spectrum Sum is the summation of logarithm power spectrum and ΔLog Spectrum Sum is the linear regression coefficients of the Log Spectrum Sum.

- **Low Energy Frame**[4]
  Low Energy Frame is the number of frames that Frame Energy is less than the threshold value. We consider that the music with many low energy frames expresses quietness.

2.1.2 Timbre features

The following 6 features are used as timbre features:

- **Centroid**[5]
  This indicates the centroid frequency of the spectrum. We use this feature as a brightness feature of music.

- **Bandwidth**[4]
  If the number of instruments in music is high, we consider that this music is cheerful. Therefore, we use the variance of the spectrum as a cheerful feature of music.

- **Rolloff**[5]
  We consider that the music in which the low-frequency band is emphasized expresses darkness. On the contrary, the music in which the high-frequency band is emphasized expresses brightness. For representing the frequency band emphasis, we use the Rolloff feature.

- **Variation of timbre**
  We use Flux[5] and Cosine Similarity[4] as the features for the variation of timbre in music.

- **Zero-Crossing**[4]
  For the pitch feature, we use the Zero-Crossing which is related to fundamental frequency in music.

2.1.3 Rhythm features

The percussive sounds, such as bass drum, high-hat, cymbal, and so on, are impulse signals in time domain. These frequency elements exist up to the high-frequency band. Hence, we use Power Spectrum Peaks[4] which indicate the number of power spectrum peaks in audio signals.

Power Spectrum Peaks are calculated as follows. First, we divide frequency into three bands, which are low-frequency band, middle-frequency band, and high-frequency band. The power spectrum is calculated every frequency bands. Next, the summation of the power spectrum is calculated in each band and this processing is repeated only constant time length. The signal which represents time variation of the power spectrum summation is applied moving average filter. We calculate differential of the filtered signal. Finally, the peaks of differential signal are detected in each frequency band and the summation of the peaks of each frequency band is calculated. That is used as a rhythm feature.

2.1.4 Harmony features

We calculate chroma vectors[8] to extract the harmony features of music. The following 3 harmony features are calculated using chroma vectors.

- **Chroma Vector Flux**[4]
  To extract variation of pitch, Chroma Vector Flux is calculated. Chroma Vector Flux indicates euclidean distance between chroma vectors.

- **Major and Minor Code Components**[4]
  The code of music is one of the mood features of music. For example, a major code music is felt as brightness, minor code music is felt as darkness. Hence, we calculate major and minor code components and use them as harmony features.

2.2 Feature selection method

We consider that the suitable features are different in each mood. If we use all features described in previous section, some features might cause the degradation in the mood score detection performance. Therefore, we need to select the suitable features for each mood. In this paper, we propose the feature selection method using correlation coefficients between features and music mood scores. In addition, the proposed method is compared with Correlation-based Feature Selection (CFS) [9] in section 3. CFS is a feature selection method using correlation between features and mood scores. We describe two feature selection methods in the following sections.

2.2.1 Feature selection using correlation coefficients

We consider that the features with high correlation against the mood scores are suitable for detecting the mood score. Hence, we propose a feature selection method using correlation coefficients. We calculate a correlation coefficient, r_{f_i,g_j}, between a feature, f_i, which indicates a feature of ith dimension and a score of mood label, g_j, in the following equation.

$$r_{f_i,g_j} = \frac{\sum_{k=0}^{N-1} \left(f_{i,k} - \overline{f}_i\right)\left(g_{j,k} - \overline{g}_j\right)}{N \cdot \sigma_{f_i} \cdot \sigma_{g_j}}, \qquad (1)$$

where N is the number of patterns. $f_{i,k}$ and $g_{j,k}$ are a feature, f_i, of the kth music pattern and a score of a mood label, g_j, of the kth music pattern, respectively. $\overline{f}_i$ and $\overline{g}_j$ are an average of feature, f_i, and an average of a score of a mood label, g_j, respectively. σ_{f_i} and σ_{g_j} is a standard deviation of a feature, f_i, and a standard deviation of a mood label g_j, respectively. Then, features which are more than threshold value r_{th} are selected.

$$\begin{cases} \left|r_{f_i,g_j}\right| > r_{th} \longrightarrow (\text{Select feature } f_i) \\ \left|r_{f_i,g_j}\right| \leq r_{th} \longrightarrow (\text{Not select feature } f_i)\,, \end{cases} \qquad (2)$$

These processing is able to give us features which have strong correlation with mood scores.

2.2.2 Correlation-based feature selection (CFS)

CFS is proposed in [9]. CFS considers not only correlation between the feature and the mood score but also correlation between features This method uses a evaluation value, $Merit_S$, calculated by the following equation for selecting the features.

$$Merit_S = \frac{D \cdot \overline{r}_{f,g_j}}{\sqrt{D + D\,(D-1)\,\overline{r}_{f,f}}}, \qquad (3)$$

where D is the number of features in a feature set, S, and $\overline{r}_{f,g_j}$ is the average correlation coefficient between the features in a feature set, S, and a mood label, g_j. $\overline{r}_{ff}$ is the average correlation coefficient between features. In this paper, $\overline{r}_{f,g_j}$ and $\overline{r}_{ff}$ are calculated in the following equations.

$$\overline{r}_{f,g_j} = \frac{1}{D}\sum_i \left|r_{f_i,g_j}\right|, \qquad (4)$$

$$\overline{r}_{f,f} = \frac{2}{D\,(D-1)}\sum_i \sum_{j,j>i} \left|r_{f_i,f_j}\right|, \qquad (5)$$

f_i and f_j are features which are included in feature set, S. r_{f_i,f_j} is a correlation coefficient between feature, f_i, and feature, f_j.

A feature set is searched that $Merit_S$ is high by the best first search. If 5 consecutive non-improvement occurs in a feature set, S, a search is stopped. After searching a feature set, locally predictive features are added to the selected feature set. If a correlation coefficient between an unselected feature and the mood score is higher than the highest correlation coefficient between an unselected feature and any one of the selected feature, its feature is added to the selected feature set.

2.3 Mood score detection

MLNNs are used for mood score detection. MLNNs are trained using the Back-Propagation algorithm (BP). These MLNNs are trained to use input acoustic features to detect each mood score for every music patterns. Each output layer unit corresponds to 7 levels of mood scores. This is shown in Figure 2. Detected mood score is given as a mood score which corresponds to the output layer unit with the highest output value. In addition, input value are normalized in the 0 to 1 range on each feature axis. MLNNs are constructed for each mood score detection.

3 Experiments

In this section, in order to evaluate the proposed method, we conduct the mood score detection experiments. In section 3.1, the feature extraction and the feature selection conditions are described. Next,

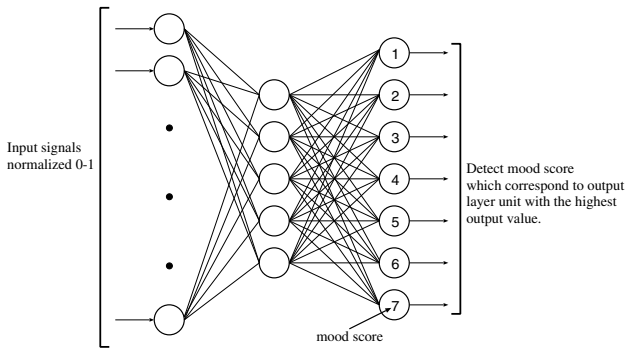

Figure 2: Mood score detection by MLNNs

we describe other experimental conditions in section 3.2 and report the experimental results in section 3.3.

3.1 Feature extraction and feature selection conditions

In this experiment, we used stereo audio signals which are sampled at 44.1kHz, quantized in 16 bits. Audio signals are framed by the slide window method and features are extracted every frames. In this experiment, the intensity features, the timbre features, and the rhythm features are extracted under the conditions that a frame length and a frame shift length are 23.2 ms and 11.6 ms, respectively. The harmony features are extracted under the conditions that a frame length and a frame shift length are 185.8 ms and 80.0 ms, respectively. The Blackmann window is used as the windowing function.

Means and standard deviations of the intensity features (except Low Energy Frames), the timbre features are calculated every 5 seconds. Power Spectrum Peaks are also calculated every 5 seconds. Means of Chroma Vector Flux are calculated every 1 seconds. The total number of dimensions of a feature vector, which is used for mood score detection, is 75. In addition, the threshold value, r_{th}, in equation (2), is set from 0.05 to 0.5 with 0.05 increments.

3.2 Experimental conditions

In this experiment, we use the RWC Music Database[10]. From this database, we use 406 music patterns which are selected from 226 music data. Some music patterns have been selected from a music datum and time length of each music pattern is 15 seconds. Mood scores of 406 music patterns are measured against each mood label, which are Brightness, Cheerful, Up-tempo, and Airiness, by 6 test subjects. The mood score of each music pattern is represented by 7 levels. For training and evaluation data, medians of mood scores which are measured by 6 test subjects are used as mood scores of each music pattern in each mood label. This experiment is conducted by 10-fold Cross-validation method. We divide music patterns into 10 segments in each mood label. 9 segments are used as training data and 1 segment is used for evaluation data. This experiment is run over 10 times in each mood label.

The number of input layer units, hidden layer units and output layer units are 75, 5, and 7, respectively. The number of iterations for training is 50,000. The training coefficient and the momentum term are set to 0.01 and 0.7, respectively.

3.3 Experimental results

The experimental results are shown in Figure 3 and 4. In these figures, “Baseline” indicates accuracies of non-feature selection method, “Proposed” indicates

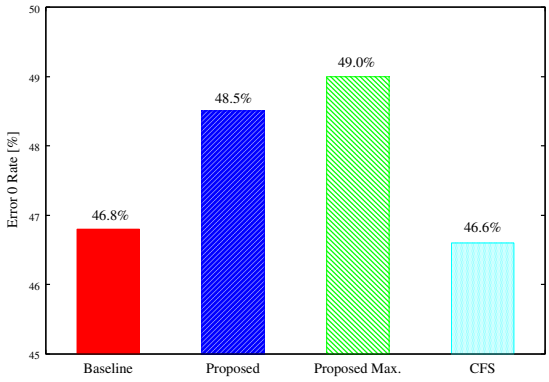

Figure 3: Experimental results (“Error 0 Rate”)

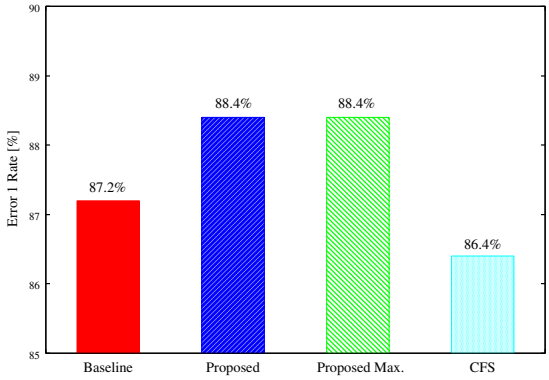

Figure 4: Experimental results (“Error 1 Rate”)

accuracies which are the highest in the average accuracies of each threshold value, “Proposed Max.” indicates accuracies which the highest accuracies are averaged for each mood label. In this experiments, two evaluation measurements, “Error 0 Rate” and “Error 1 Rate”, are used. “Error 0 Rate” indicates the rate of the music patterns for which the difference between detected and ground-truth score is 0, “Error 1 Rate” indicates the rate of the music patterns for which the difference between detected and ground-truth score is less than 1. From the experimental results, compared “Proposed” with “Baseline”, we can see that “Error 0 Rate” is improved 1.7% (Error Reduction Rate (ERR) is 3.2%), “Error 1 Rate” is improved 1.2% (ERR is 9.1%). Furthermore, compared “Proposed Max.” with “Baseline”, we can see that “Error 0 Rate” and “Error 1 Rate” are improved 2.1% (ERR is 4.1%) and 1.2% (ERR is 9.6%), respectively. From these results, we conclude that the proposed method is efficient for mood score detection.

Figure 5 and 6 show the accuracies of each mood label as a function of threshold value, r_{th}. In these figures, points indicate the highest accuracies for each mood label. These results show that threshold values in which the highest accuracies of each mood label are different. For this reason, it is important to determine the threshold values of each

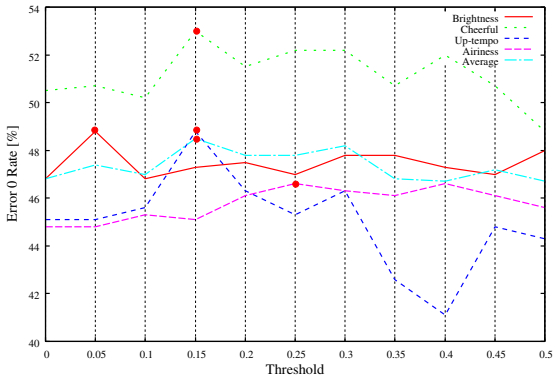

Figure 5: "Error 0 Rates" every threshold values in each mood label

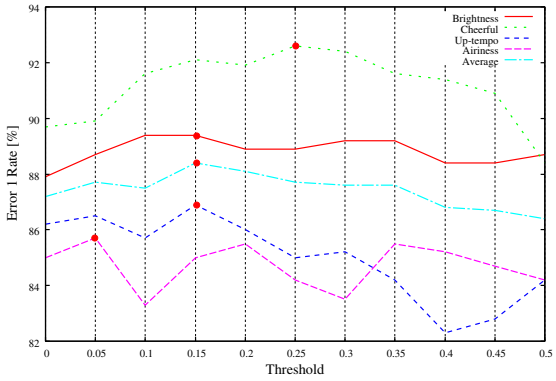

Figure 6: "Error 1 Rates" every threshold values in each mood label

Table 1: Features which have strong correlation with each mood label (Cond.:$|r_{f_i,g_j}| > 0.6$)

Mood	Feature
Brightness	Log Spectrum Sum(Mean)
Cheerful	Log Spectrum Sum(Mean) Cosine Similarity(Mean) Major Code Component
Up-tempo	Chroma Vector Flux(Mean)
Airiness	Power Spectrum Peaks

mood because the suitable combination of features to detect the mood score is different every moods. Table 1 shows that features which have strong correlation (correlation coefficient $|r_{f_i,g_j}|$ is more than 0.6) with each mood label. In this table, "Mean" indicates the mean of feature parameters. From this table, Log Spectrum Sum, which represents intensity of the frequency domain, has strong correlation with "Brightness" and "Cheerful". Chroma Vector Flux and Power Spectrum Peaks, which represent variation of pitch and power spectrum, have strong correlation with "Up-tempo" and "Airiness". We conclude that features which represent intensity and variation of pitch and power spectrum are efficient

Table 2: The average number of selected features

Mood	Brightness	Cheerful	Up-tempo	Airiness
CFS	10.4	11.2	6.1	8.3
Propose	73.7	68.0	71.5	68.9

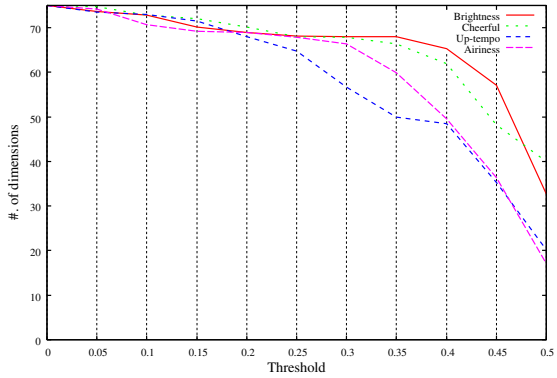

Figure 7: The average number of features selected by the proposed method

for these mood labels.

Next, the proposed method is compared to CFS. From Figure 3 and 4, accuracies of the proposed method are higher than these of CFS. In addition, accuracies of CFS are lower than these of "Baseline".

For analyzing this result, we investigate the average number of features selected by CFS of each mood. This result is shown in Table 2. From this table, we can see that the average number of features selected by CFS is about 10, which are much less than the number of original features. As a reason, we consider that CFS does not select the features which have strong correlation between features. Hence, the number of features selected by CFS is small. Meanwhile, the average number of features selected by the proposed method is greater than that by CFS. We show the average number of features selected by the proposed method in Figure 7. From this figure, we can see that the proposed method selects the features more than CFS. The number of features selected by the proposed method is significantly decrease under the conditions that the threshold is more than 0.35. Therefore, we conclude that a lot of features are necessary for the mood score detection.

4 Conclusion

In this paper, we proposed a feature selection method for detecting the music mood scores. First, features which are related to intensity, timbre, rhythm, and harmony are extracted from audio signals. Next, the proposed feature selection method selects the features using correlation information between the feature and the score of each mood label. Finally, the

mood scores are detected by MLNNs.

For evaluating the proposed method, we conducted the mood score detection experiments using 406 music patterns. From experimental results, "Error 0 Rate" and "Error 1 Rate" were improved 2.1% (ERR was 4.1%) and 1.2% (ERR was 9.6%) compared to non-feature selection, respectively. In the features selected by the proposed method, we showed that features which represent intensity and variation of pitch and power had strong correlation with mood labels. Furthermore, compared to CFS, the proposed method was efficient rather than CFS because the proposed method selected a lot of features from the original feature set.

From the experimental results, the optimal threshold values were different every mood labels. In the future, we shall study to choose automatically optimal threshold values in the proposed method. Furthermore, if acoustic features which have strong correlation with mood labels are developed, we consider that detection performance is improved. Hence, we plan to study more efficient acoustic features which have strong correlation with mood labels.

Acknowledgment

This research has been partially supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Young Scientists(B), 19700172, Scientific Research(B), 21300036.

References

[1] Dan Liu, Lie Lu, and Hong-Jiang Zhang. "Automatic Mood Detection from Acoustic Music Data". *Proc. of ISMIR2003*, October 2003.

[2] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. "Multi-Label Classification of Music Into Emotions". *Proc. of ISMIR2008*, pages 325–330, September 2008.

[3] Ryo Hirae and Takashi Nishi. "Mood Classification of Music Audio Signals". *The Journal of the Acoustical Society of Japan*, 64(10):607–615 (in Japanese), 2008.

[4] Masato Miyoshi, Satoru Tsuge, Hillary Kipsang Choge, Tadahiro Oyama, Momoyo Ito, and Minoru Fukumi. "Music Impression Detection Method for User Independent Music Retrieval System". *Proc. of KES2010*, pages 612–621, September 2010.

[5] Geroge Tzanetakis and Perry Cook. "Musical Genre Classification of Audio Signals". *IEEE Transaction on Speech and Audio Processing*, 10(5):293–302, July 2002.

[6] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. "Music Type Classification by Spectral Contrast Feature". *Proc. of ICME2002*, 2002.

[7] Tim Pohle, Elias Pampalk, and Gerhard Widmer. "Evaluation of Frequently Used Audio Features for Classification of Music Into Perceptual Categories". *Proc. of ISMIR2005*, September 2005.

[8] Masataka Goto. "A Real-time Music Scene Description System: A Chorus-Section Detecting Method". *MUS2002*, 2002(100):27–34 (in Japanese), 2002.

[9] Mark A. Hall. "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning". *Proc. 17th Int'l Conf. Machine Learning*, pages 359–366, 2000.

[10] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. "RWC Music Database: Database of Copyright-cleared Musical Pieces and Instrument Sounds for Research Purposes". *The Journal of Information Processing Society*, 45(3):728–738 (in Japanese), 2004.