Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, 10-13 July, 2016

MUSIC EMOTION RECOGNITION BASED ON TWO-LEVEL SUPPORT VECTOR CLASSIFICATION

CHINGSHUN LIN, MINGYU LIU, WEIWEI HSIUNG, and JHIHSIANG JHANG

Department of Electronic and Computer Engineering
National Taiwan University of Science and Technology
43, Section 4, Keelung Rd., Taipei, Taiwan
E-MAIL: mentortw@gmail.com and B9902010; M10402101; M10402150@mail.ntust.edu.tw

Abstract:

Music emotion recognition (MER) detects the inherently emotional expression of people for a music clip. MER is helpful in music understanding, music retrieval, and other music-related applications. As volume of online musical contents expands rapidly in recent years, demands for retrieval by emotion have been emerging lately. Determining the emotional content of music computationally is an interdisciplinary research involving not only signal processing and machine learning, but the understanding of auditory perception, psychology, cognitive science, and musicology. One of the challenges in evaluating automatic music emotion detection is that there is currently no well-developed emotion model for music emotion description. Moreover, owing to the low transparency of the acoustic feature based music emotion recognizer, it is difficult to interpret data generated by this mechanism. In this study, a two-level classification system based on both music genre and music feature pre-described by the domain knowledge is proposed. This framework has the advantage of utilizing the most suitable acoustic information. Experiments will be conducted via the measure of the correlation between diverse emotional expressions and various musical cues. To verify the performance of overall system, the proposal model will also be evaluated based on the consistency between music features and ground-truth emotions.

Keywords:

Affective computing; Music emotion recognition; Music information retrieval; Feature extraction; Support vector machine.

1 Introduction

In the music industry, the demands for music retrieval and recommendation attract lots of researchers to explore the acoustic characteristics for music genre classification [1]. Moreover, the music emotion recognition system has also been developed under the necessity of music understanding. Unlike image or video processing, there are fewer significant music features with strong link to the emotion perception. Moreover, users' emotional responses to music vary from person to person as nature of music is complex. As a result, music emotion recognition (MER) system has been developed for years with limited progress. Since emotion perception is intrinsically subjective, one could perceive different emotions even when listening to the same song. The subjectivity, introduced by subjects' background such as sex, generation, culture and personality, makes the evaluation of an MER system difficult since a consensus of classification is hard to reach. In addition, it is generally not easy to describe and label emotion in a universal way because the adjectives used to describe emotions could be ambiguous. A variety of reasons that might evoke a certain emotion, no matter from the same or different music, are still far from fully understood.

To explore the relationship between music and emotion, a hierarchical framework using acoustic characteristics such as timbre, intensity, and rhythm as features was proposed for music emotion detection [2]. A similar work has been done using the fuzzy classifier based on some specific rules and assumptions [3]. However, neither one's individual response to specific music, nor the complicated nature is taken into consideration. On the other hand, the ambiguity of emotional description has been reduced by introducing the emotion classes in terms of valence (positive or negative stress) and arousal (high or low energy) [4]. However, even with the emotion plane, a specific emotion could distribute over a certain region on the valance-arousal (VA) plane, in which emotion states may vary according to the ever changing melody. For instance, the first quadrant of the emotion plan contains emotions such as excited, happy, and pleased, which are different adjectives in nature. This obscurity might mislead the subjects into an inconsistent

978-1-5090-0390-7/16/$31.00 ©2016 IEEE

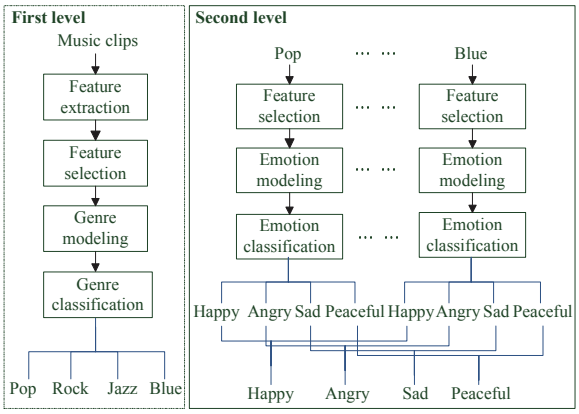

FIGURE 1. Block diagram outlining music emotion recognition system based on two-level support vector classification.

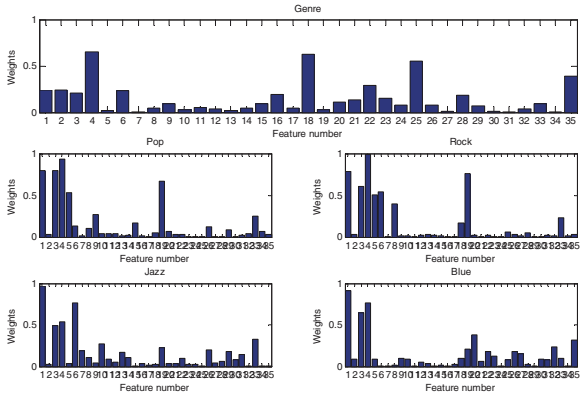

FIGURE 2. Feature weight extraction based on RReliefF for music genre classification.

emotional state and a diverse experimental result. Moreover, people tend to give more subject evaluation for valence than for arousal, and therefore a larger variance in statistics. Another problem of using the VA plane to denote the music emotion is that the participant might be confused with the degree of arousal and valence, especially for the latter.

Meanwhile, different classifiers based on the machine learning algorithms such as K-nearest neighbors [5], support vector regression [6], Gaussian mixture models [7], and support vector machines [8] have also been proposed. Unlike dealing with the music factors and the emotion states on the VA plane, we use the acoustic features directly extracted from the raw music clips and apply them to two SVM classifiers. The main idea of this work is to add one more classification level to the overall system for a higher classification rate. First of all, we evaluate and select the features directly related to the perceptual characteristics of music by RReliefF [9]. Second, a two-level support vector machine, a learning algorithm capable of learning from the labeled training data, is used to classify patterns from the aspects of both music genre and music emotion. The music genre composed and categorized with lots of emotion issues, on the other hand, serves as an important and straightforward factor when recognizing the music emotion. The music factors extracted from the music genre in advance may provide a clear picture for the understanding of the music emotion. With this conception in mind, designing the music genre classification as a preprocessor with a following music emotion recognizer motivated us to design a SVM-based two-level classifier [10]. Therefore, the aim of this study is to develop a more intuitive method that efficiently links the music emotion and acoustic feature together, and thus to increase the overall accuracy for the identification. Once the SVM-based music genre classification is complete, we send the primary result to the music emotion classifier with the same mechanism for evaluating the usefulness of the proposed system. Fig. 1 illustrates the block diagram for such a music emotion recognition system, which would be explained in more details below.

2 Feature Extractions and RReliefF

2.1 Music information retrieval

The arousal may be determined by tempo, pitch, loudness, and timbre, whereas the valence may depend on mode and harmony [11]. These low-level music features are able to provide more intuitive stimuli for human perception. However, to properly recognize the music emotion, more sophisticated features such as pulse clarity, roughness, and inharmonicity should be considered as well. In this study, we parameterize the music information retrieval (MIR) toolbox as the feature extraction. This extraction is mainly based on four categories of musical elements, i.e., rhythm, timbre, tonality, and dynamics, which have been constructed as a 35-dimensional feature vector in this application [12]. These features are described below:

2.1.1 Rhythm

The rhythm, defined as the time-variant beat or tempo (#2), is used to represent the timing of sound events. Beat usually represents a periodic length of 1/4 note, whereas tempo is usually defined as the beats per minute for representing the global rhythmic feature. In general, fast music with high tempo usually makes listeners tense or exciting. On the contrary, slow music with low tempo is boring or relaxing. Similarly, regularity of beats appeases listeners, but irregular beats make people

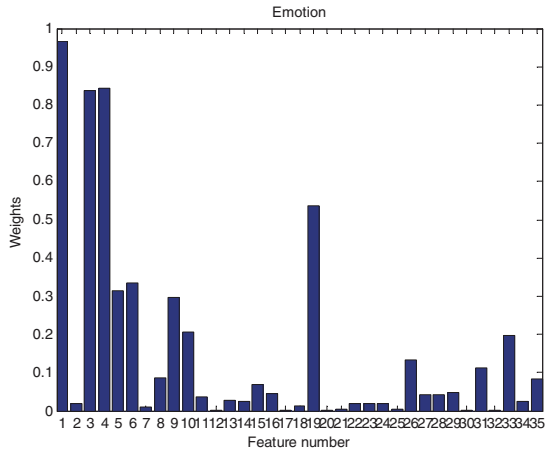

FIGURE 3. Feature weight extraction based on RReliefF for music emotion recognition.

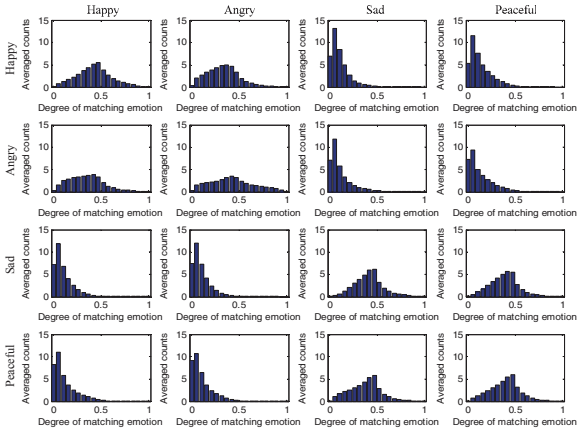

FIGURE 4. Histogram of the degree of matching emotion obtained via cross-validation.

feel annoying or anxious. We take event density (#1) which is the average frequency of onsets to estimate rhythm. The onset is defined as a sudden burst of energy or a change in the short-time spectrum of a music clip. On the other hand, pulse clarity (#3), extracted from the onset curve, is defined as how easy the listener can perceive the underlying rhythmic or metrical pulsation in music.

2.1.2 Timbre

The spectral rolloff (#18), defined as the number of frequency which contains energy below a certain percentage of the total energy, shows the amount of high frequency in the signal. The percentage assigned here, say 90%, is to properly include all the eligible features. On the other hand, the brightness (#19) is obtained by measuring the long-term average spectrum, which represents the percentage of energy above the cut-off frequency 2 kHz. In addition, the zero crossing rate (#4) and the first 13 order MFCCs (#5-#17) have been used to form the feature vector, whereas roughness (#20) is related to the energy of beating appearance based on the lower spectral of neuronal patterns. The lower spectral of neuronal patterns can be obtained by the model of Van Immerseel and Martens [13].

2.1.3 Tonality

The chromagram (#21-#32) is calculated in the log-scaled spectrogram for representing the frequencies in musical scales. There are 12 chromas corresponding to the 12-tone equal temperaments in each octave and wrapped along the 12 pitch classes (#33) to describe the energy distribution [14]. On the other hand, the mode is a type of scale in music which implies major and minor to estimates the modality based on chroma components. Moreover, inharmonicity (#34) estimates the amount of overtones which deviate from the multiples of the fundamental frequency [15].

2.1.4 Dynamics

In the dynamic analysis, we detect the vitality of the music clip by computing its root mean square (RMS). The RMS of a discrete-time signal estimates energy feature (#35) according to the loudness of a syllable. It is worth noticing that the music energy is consistent with the loudness people perceive in listening.

2.2 RReliefF

In this study, RReliefF, a feature weighting approach considering interrelationship among features, is utilized for feature selection in both classification levels for its effectiveness [9]. It evaluates the features one by one and weights each feature according to its importance. For this evaluation, RReliefF selects instances randomly and then find its K-nearest instances. The distances between all instances and the instance that are processing are then updated. Since the extracted features do not always reflect the emotion evoked by the music, weighting the features according to the relevance would not only raise the representation, but reduce the feature dimensions. As a result, Fig. 2 shows the feature weights extracted by RReliefF for the music genre classification, whereas Fig. 3 presents the weights for the music emotion recognition.

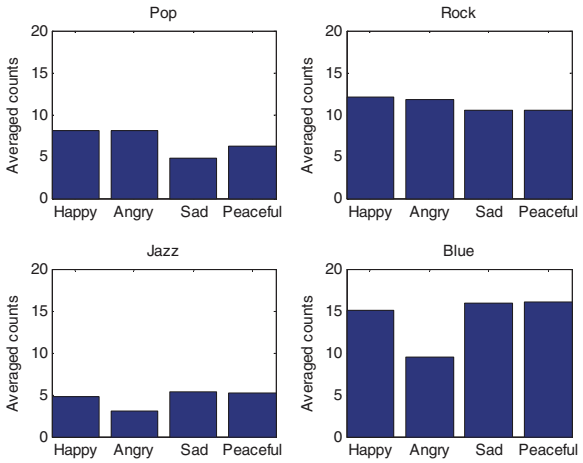

FIGURE 5. Results of music emotion recognition for various music genres.

3 Experimental results

In this experiment, the music database is made up of 301 songs selected from a number of albums. The detailed quantity for each music genre and emotion are tabulated in Table 1. All the music is separated into 30-second clips, and each music clip is taken as a training or testing pattern, which originally belongs to one of the four genres with various moods form the databases. The music corpus we use in this application is mainly consisted of western tonal music, since the music perceptual features we use are based on psychological theories in western cultures. Western music genres, including pop, rock, jazz, and blue are used because they are easier to arrive an agreement on the subject evaluation. On the other hand, since people are more sensible to the valence instead of arousal features, we define one representable term for each quadrant, i.e., happy, angry, sad, and peaceful in Thayer's VA plane for convenience [4]. More specifically, the ground truth is set via a subjective test provided by the well labeled databases which contain the averaged subjects' opinions in terms of emotion indicated in the emotion plane for each music clip. To make a fair comparison, the music clips are all re-sampled to the uniform format of 44.1 kHz, 16-bit depth resolution, and mono channel PCM wave files.

In addition, to reduce the effects of variance caused by a diverse range of values among different features, each feature is normalized by the z-score before sending to SVMs for training and testing. We evaluate the performance of classification by the *10*-fold cross-validation, in which the whole dataset is randomly divided into two parts as both training and testing

TABLE 1. Number of music clips used for music emotion recognition.

	Happy	Angry	Sad	Peaceful	Total
Pop	20	19	18	18	75
Rock	20	19	20	20	79
Jazz	18	15	17	18	68
Blue	22	19	19	19	79
Total	80	72	74	75	**301**

patterns. In addition, all the processes mentioned above are repeated 10^4 times for fair averaged results. Unlike the conventional classification approaches that simply assign one emotion class to each music clip in a dichotomous manner, we introduce the matching degree ranging between 0 and 1 to each clip for a better emotion representation. Fig. 4 shows the histogram of degree of matching emotion obtained via cross-validation in the subjective test. We can observe that more than 30% of a specific emotion distributes around 0.5 for the plots along the main diagonal, showing the classification provided by the SVMs is reliable. In contrast, the highest distributions range from 0.05 to 0.2 for the off-diagonal plots. More information may also be observed from the ambiguity of happy and angry as well as sad and peaceful. This verifies that the arousal, a less subjective feature than valence, is more relative to features extracted by the MIR. Fig. 5 shows the results of the music emotion recognition for various music genres. The most significant results may be found in the pop and blue, in which happy and angry, sad and peaceful are with highest arousal values, respectively.

4 Conclusion

Music is a medium people use to express or perceive emotions, and therefore music emotion recognition (MER) is useful in music retrieval, understanding, recommendation, and any other music-related applications. In this work, we develop an MER system based on a two-level support vector classifier, which is proved to be more reliable than directly recognizing music emotion by using acoustic characteristics as features. Moreover, the relationship between the music genres and music emotions may also be observed from the correlation extracted by the cross-validation.

Acknowledgment

The work reported in this paper was partially supported by grant from the Ministry of Science and Technology of Taiwan under contract MOST 104-2221-E-011-161.

References

[1] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133-141, Mar. 2006.

[2] D. Liu, L. Lu, and H. J. Zhang, "Automatic mood detection from acoustic data," *4th Intl. Conf. on Music Information Retrieval*, pp. 81-87, Oct. 2003.

[3] P. Lucas, E. Astudillo, and E. Peláez, "Human-machine musical composition in real-time based on emotions through a fuzzy logic approach," *IEEE Latin America Congress on Computational Intelligence*, pp. 1-6, Oct. 2015.

[4] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press, New York, 1989.

[5] K. C. Dewi and A. Harjoko, "Kid's song classification based on mood parameters using K-nearest neighbor classification method and self organizing map," *Intl. Conf. on Distributed Framework and Applications*, pp. 1-5, 2010.

[6] B. J. Han, S. M. Rho, R. B. Dannenberg, and E. J. Hwang, "SMERS: Music emotion recognition using support vector regression," *Proc. of Intl. Society for Music Information Retrieval*, pp. 651-656, 2009.

[7] K. Markov and T. Matsui, " Music genre and emotion recognition using Gaussian processes," *IEEE Access*, vol. 2, pp.688-697, Jun. 2014.

[8] C. Y. Chang, C. Y. Lo, C. J. Wang, and P. C. Chung, "A music recommendation system with consideration of personal emotion," *Intl. Computer Symposium*, pp.18-23, Dec. 2010.

[9] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1, pp. 23-69, Oct. 2003.

[10] C. C. Chang and C. J. Lin, LIBSVM: A Library for Support Vector Machines. [Online] Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[11] A. Gabrielsson and E. Lindstrom, "The influence of musical structure on emotional expression," *Music and Emotion: Theory and Research, Oxford University Press*, New York, pp. 223-248, 2001.

[12] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," *8th Int. Conf. on Music Information Retrieval*, pp. 127-130, 2007.

[13] L. Mion and G. D. Poli, "Score-independent audio features for description of music expression," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 458-466, Jan. 2008.

[14] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 96-104, Feb. 2005.

[15] Y. H. Yang and H. H. Chen, *Music Emotion Recognition*, CRC Press, 2011.