A Study on Query-by-Any-Word Based Music Retrieval System

Shinji Sako, Ai Zukawa, and Tadashi Kitamura

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
{s.sako,kitamura}@nitech.ac.jp, aizkw@mmsp.nitech.ac.jp

Abstract. Recently, commercial interest in the field of music information retrieval (MIR) has been growing rapidly. This paper describes a MIR system that accept any Japanese word as query. Previous studies focused on emotion based MIR system generally uses limited words such as major adjectives or kansei words. However, emotion of the music is represented by various words in practice. Music review is a one of good example. Word can also express complicated emotions with which various emotions are mixed. Starting from this point of view, we propose a method for MIR system that is able to find the appropriate music directory from any word as query. There are three main issues in this study. First one is how to mapping music and emotion. We introduce two-dimensional space which can represent emotion and music in a unified space. This space is obtained automatically from the emotion evaluation data of words and music. Second issue is extraction method for musical feature in order to map to the emotion space from given music. In our approach, optimal feature parameters are automatically selected with respect to each axis of the emotion space. Third issue is how to cope with any query word. Our method can find a music pieace corresponding to emotion of any word by measurement of relationship between each basic word for the given query word. A feature point of this approach is the use of co-occurrence probability of words obtained from a large scale of web text corpus. We performed a subjective evaluation experiments using 100 classical musical pieces and 50 Japanese words that are often used in music reviews. The experimental results show that our proposed system can find the correct music piece which matches mostly given query word.

Keywords: emotion based music retrieval system, co-occurrence probability of word.

1 Introduction

The general approach to music retrieval by emotion is used adjective words (e.g. happy, sad and etc.). Sato, et. al. proposed the retrieval method for music works using emotion word [1]. Using the affective value scale of music, they established the relation between music structure and six factors of emotion (positive, negative, affection, strength, frivolousness, and solemnity). Levy and Sandler[2] also proposed a semantic space using high-volume social tags, because tags for music capture sensible attributes grounded in individual tracks. Moreover, Kumamoto and Ota[3] suggested an impression-based music retrieval system with 10 adjective-pairs. It asks users to select one or more pairs

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 228–233, 2014.

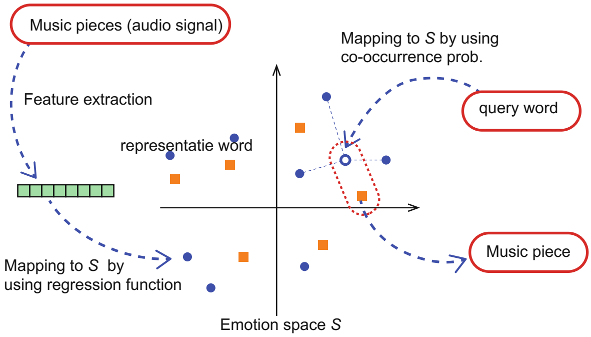
Music pieces (audio signal)
Feature extraction
representatie word
Mapping to S by using co-occurrence prob.
query word
Mapping to S by using regression function
Music piece
Emotion space S

Fig. 1. System outline

and estimate each pairs on a seven-step scale, and outputs musical piece which is suitable for the given word. However, there exists many ways to to represent emotion of music in practive. For example, text of music review consists of many kind of word to represent impression or emotion of music pieces. From this point of view, we proposed the MIR system which can find music piece from any query word. In order to express any words, relation of words is utilized. Emotion space is configured as an indicator of emotion music and words, and some words whose impression is known were mapped in the space. Any query words can be mapped from these words by using similarity of words. Emotion of music is also mapped to the identicle emotion space from music features that are extracted audio signal. We can calculate the distances between music and word in the emotion space.

In section 2, we propose the approach which deals with any words and describe emotion space. Next, section 3 presents how to map query words, and section 4 presents how to map music in the space. Section 5 shows the subjective evaluation experiment. Finally, conclusions and future work are drawn in section 6.

2 Emotion Space for MIR

The outline of our system is shown in Fig. 1. Any music pieces are mapped in the emotion space, and the words whose emotion is already known *representative words*, are also mapped by the similarity. When a query word is given to the system, the position of query word on emotion space is determined using representative words. Then, the Euclidian distances between query word and music pieces in the emotion space is calculated, and the music piece which is the closest to query word is selected as the search result.

2.1 Development of the Emotion Space

In order to set the emotion space, we conducted factor analysis by using an impression of music pieces database[4]. This database contains around 120 subjects who were asked to listen 100 short music pieces and evaluate them at the 7 point scale for the

Table 1. 14 adjective pairs

light – heavy	bright – dark	sad – cheerful	powerful – quiet
calm – violent	dismal – amusing	free – subdued	fast – slow
elegant – rough	lonely – lively	hurried – careless	profound – nimble
relaxed – tense	gorgeous – simple		

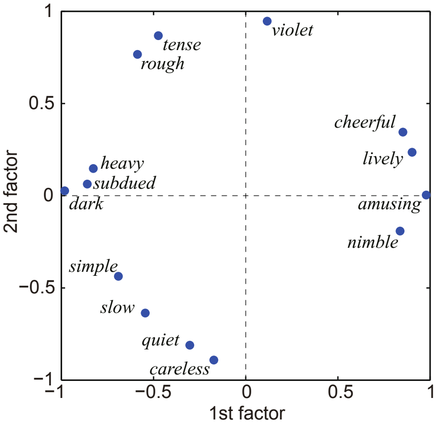

Fig. 2. Emotion space S

adjective-pairs1. Each music clip was segmented in 15 seconds from well-known classical pieces in RWC music database[5], and impression of music was regarded as constant. As a result of further investigation the database,it was found that some subjects with low reliability are included. For example, some subjects did not spent enough time to listen the music. Such kind of outliers are omitted from database in advance. In addition, to get a common emotion space, the data whose values are more than the standard deviation away from the average of all subjects are removed, too.

Factor analysis was conducted on the evaluation scores of 14 adjective pairs. Fig. 1 shows 14 adjective words in the generated emotion space. As the result, 1st factor mainly contains “bright” and “amusing”, it means brightness. On the other hand, because 2nd factor mainly contains “violent” and “hurried”, it means violence.

3 Mapping Method of the Query Word

It is difficult to assign in advance the position of any word on the emotion space. We propose a method for estimating the position on the emotion space automatically by using the relationship between words. In our approach, the relationship between words is represented by co-occurrence probabilities of words pairs. We used the large scale co-occurrence probability database of Japanese that was developed by ALAGIN project[6]. This database contains co-occurrence probability of half a million headword by using 100 million web text corpus. A co-occurrence probability $C(w_1, w_2)$ of word w_1 and

w_2 is defined as Eq. 1. Where, $F(w_1)$ and $F(w_2)$ means individual word frequencies of word w_1 and word w_2, and $F(w_1, w_2)$ means co-occurrence frequency of w_1 and w_2.

$$C(w_1, w_2) = \frac{F(w_1, w_2)}{F(w_1) + F(w_2)} \tag{1}$$

3.1 Definition of Representative Word

To represent the emotion of any word by the relationship of base words, one possible approach is use of well-known adjective words. However, it seems not realistic to express tens of thousands query word by using small number of words. In practice, a lot of base words are needed in order to cope with various words. In this study, we introduced *representative word* as the base word. The representative words are selected which has high degree of co-occurrence probability with adjective word. It is noted that adjective word were converted to noun form The reason of this, it is considered that noun has a high co-occurrence probability than an adjective.

The position of the representative words r in the emotion space is calculated as in Eq. 2. Where a_i is the adjective word, and N means the number of adjective word.

$$\left[\frac{C(r, a_n)}{\sum_{i=1}^{N} C(r, a_i)} \cdot x_{1a_n}, \frac{C(r, a_n)}{\sum_{i=1}^{N} C(r, a_i)} \cdot x_{2a_n}\right] \tag{2}$$

3.2 Calculate the Position of Query Word

In order to determine the position of the query word, we used representative word closely related to the query word. For given query word w, representative words can be found from 300 word high degree of co-occurrence probability The position of w on the emotion space is calculated as Eq. 3. Where $(r_1, r_2, \ldots, r_M)$ are representative words and M means number of representative word.

$$\left[\frac{\sum_{i=1}^{M} (C(w, r_i) \cdot x_{1r_i})}{\sum_{i=1}^{M} C(w, r_i)}, \frac{\sum_{i=1}^{M} (C(w, r_i) \cdot x_{2r_i})}{\sum_{i=1}^{M} C(w, r_i)}\right] \tag{3}$$

4 Mapping Method of Music Piece

4.1 Mapping Function by Using Regression

We considered that to represent the relationship between the emotion and features with the impression evaluation data as same as Sec. 2 by using multiple regression analysis. An average value of *bright – dark* and *calm – violent* are used as dependent variables of 1st factor and 2nd factor, respectively. It is noted that the mean value should be normalized as $[-1, 1]$, because range of emotion space should be $[-1, 1]$.

Table 2. List of feature parameters

Root mean square energy	Low energy	Roll off	Spectral flux
Spectral centroid	Roughness	Zero crossing	MFCC
Harmonic change	Key clarity	Mode	Tempo

Table 3. Result of evaluation experiment

Mean value of score	Number of sample
1 – 2	0
2 – 3	3
3 – 4	23
4 – 5	24

4.2 Feature Parameter and Variable Selection by Stepwise Procedure

Many studies have been done to extract musical features from audio signal. We used *MIRtoolbox*[7] to extract well-known frame level feature parameters listed in Table 2.

In our experiments, sampling frequency, window size (frame length), and window hop size was 44.1 kHz, 44.66 msec, and 20 msec, respectively. Feature parameters for each music piece are calculated as the mean and variance of all over the frames.

It is assumed that an optimal feature parameter depend on each axis of emotion space. We conducted stepwise procedure to determine optimal feature parameters for each axis independently. Variable selection is made variance ratio F until less than 2 in this experiment. The variable selection process was repeated until variance ratio less than 2. As the result, 16 dimensional and 17 dimensional feature parameters were selected independently in the 1st and 2nd axis, respectively.

5 Experiment

The subjective listening tests were conducted to evaluate this method. It was verified that whether the selected music piece meets the emotion of query word. We selected 50 words as query that can recall certain emotion, and searched music piece from 100 classical music pieces that are same as Sec. 2. Subjects were asked to agree or disagree and write down the number of statements on a 5-point scale (5: Agree, 1: Disagree). 8 subjects were men and women aged 20s. Table 3 shows the results of subjective evaluation. We confirmed that most of subjects agree with the result of query.

We also examined the relationship between the evaluation value of the subjective test and the distance between the (x_{w_1}, x_{w_2}) from the origin on the emotion space. The correlation coefficient between two variables was 0.36. This result implies that accuracy of mapping becomes high when the degree of the emotion is strong. However, some query words failed to mapping the emotion space. The concrete examples of such problem are “pure” and “sleepiness”, etc.. We need improve mapping method that is not dependent on a particular word.

6 Conclusion

In this paper, we proposed an emotion based MIR system which is capable to query by any Japanese word. 2-dimensional emotion space that indicates both of any word and music was calculated using SD method and factor analysis. In order to mapping any query words, some words whose emotion is known are selected, and any words are translated into these words. Therefore, the words which are similar to the adjective words are set in the space using the co-occurrence probability. And, the coordinates of any words are decided by these words and the co-occurrence probability. On the other hand, music is located by music feature parameters because impression of music appears in them. When a word is an input to the system, it outputs music which is close to the word in the impression space. Through a subjective evaluation experiment, selected music is suitable for the corresponding input query word Evaluations of several music collections showed that our approach achieves encouraging results in terms of recommendation satisfaction. It was also showed that emotion of music is more complicated, we extend our system which can cope with multiple words or sentence as query. This is direction of our future work.

References

1. Sato, A., Ogawa, J., Kitakami, H.: An impression-based retrieval system of music collection. In: Proc. of 4th International Conference on knowledge-Based Intelligent Engineering Systems & Allied Technologies, pp. 856–859 (2000)
2. Levy, M., Sandler, M.: A semantic space for music derived from social tags. In: Proc. of 8th International Society for Music Information Retrieval Conference (ISMIR 2007), pp. 411–416 (2007)
3. Kumamoto, T., Ohta, K.: Design, implementation, and opening to the public of an impression-based music retrieval system. Transactions of the Japanese Society for Artificial Intelligence 21, 310–318 (2006)
4. Iwatsuki, Y., Sako, S., Kitamura, T.: An estimation method of musical emotion considering individ. In: Proc. of International Conference on Kansei Engineering and Emotion Research (KEER 2012), pp. 456–461 (2012)
5. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Popular, classical, and jazz music databases. In: Proc. of the 3rd International Society for Music Information Retrieval (ISMIR 2002), pp. 287–288 (2002)
6. NICT MASTAR Project: ALAGIN Language Resources and Voice Resources Site (accessed May 18, 2014)
7. Lartillot, O., Toiviainen, P.: MIR in Matlab(II): A Toolbox for Musical Feature Extraction form Audio. In: Proc. of the 7th International Conference on Music Information Retrieval (ISMIR 2007), pp. 287–288 (2002)