Subjective Similarity of Music: Data Collection for Individuality Analysis

Shota Kawabuchi* and Chiyomi Miyajima* and Norihide Kitaoka* and Kazuya Takeda*
* Nagoya University, Nagoya, Japan
E-mail: shota.kawabuchi@g.sp.m.is.nagoya-u.ac.jp Tel: +81-52-789-4432

***Abstract*—We describe a method of estimating subjective music similarity from acoustic music similarity. Recently, there have been many studies on the topic of music information retrieval, but there continues to be difficulty improving retrieval precision. For this reason, in this study we analyze the individuality of subjective music similarity. We collected subjective music similarity evaluation data for individuality analysis using songs in the RWC music database, a widely used database in the field of music information processing. A total of 27 subjects listened to pairs of music tracks, and evaluated each pair as similar or dissimilar. They also selected the components of the music (melody, tempo/rhythm, vocals, instruments) that were similar. Each subject evaluated the same 200 pairs of songs, thus the individuality of the evaluation can be easily analyzed. Using the collected data, we trained individualized distance functions between songs, in order to estimate subjective similarity and analyze individuality.**

I. Introduction

With the emergence of mass storage media and the progression of data compression technology, users can experience difficulty finding desired songs in a large database due to the quantity of data. Estimating subjective music similarity and using this information for the retrieval of songs is one of the possible solutions to this problem. As a result, there are many studies and systems for retrieving songs from such databases (e.g. [1], [2]), and for recommending songs using the preferences of users (e.g. [3], [4]). In order to further develop such systems, there are many discussions about methods of calculating music similarity. Methods which use acoustic features to calculate similarity have long been used by many music retrieval systems. The usual method is to calculate acoustic features that correspond to spectral shapes, such as Mel Frequency Cepstrum Coefficients (MFCC) for each time frame, and then calculate similarity between distributions of the features. For example, Pampalk [5] calculated MFCC for each time frame and then represented the distribution of MFCC as a Gaussian Mixture Model (GMM) or single Gaussian distribution, and calculated dissimilarity between distributions using Kullback-Leibler divergence.

In this paper, we examine estimating subjective music similarity from acoustic music similarity. If we could extract acoustic features which contribute to the perception of music similarity, and calculate similarity between those features as humans do, we could discover a useful subjective similarity estimation method. However, there are many factors which are not clear yet in the field of music similarity perception. Thus, it is still too difficult to estimate subjective music similarity from acoustic music similarity. In our study, we collect similarity evaluation data which were provided by humans, and investigate promising acoustic features and similarity measures between the features, using the collected data as "ground-truth". However, the existence of limitations on the performance of the method, which uses conventional acoustic features, was suggested in [6]. Individualizing subjective similarity estimation is one simple approach to improving retrieval performance, because the method referred above doesn't consider individualization of the features. As shown in some studies (e.g. [4]), similarity perception varies between individuals, and an optimal acoustic similarity measure for each individual should be used in retrieval systems.

In this study, subjective evaluation data for 200 musical pairs was collected, each pair consisting of 30 second sections of two songs. By using the collected data, we can study which acoustic features contribute to the perception of similarity. However, it can be assumed that there is individual variation in judging subjective music similarity. Thus it is important to investigate what kinds of differences in perception cause this variation. In the experiment, each subject evaluated the same 200 pairs of songs. Differences in evaluation results among the subjects should reflect the individuality of the subjects well.

In this paper, data collection is explained in section 2. In section 3 we describe how the collected data is used to develop a method of calculating acoustic music similarity which is able to reflect the individuality of the subjects. In section 4 an experiment was conducted to measure how the method described in section 3 improve the performance of subjective similarity estimation. Section 5 concludes the paper.

II. Data Collection

A. Subjects

The number of subjects who participated in the experiment was 27 (13 male and 14 female). All of the subjects were in their twenties.

B. Songs used

A total of 80 songs were used in the experiment, which were obtained from the RWC music database "Popular music" [7], which include 100 songs. The songs used were all Japanese popular music (J-Pop) songs (songs No. 1-80 were used). The length of the recordings used was 30 seconds, beginning from the starting point of the first chorus section. The RWC music database was annotated (AIST annotation [8]) and this

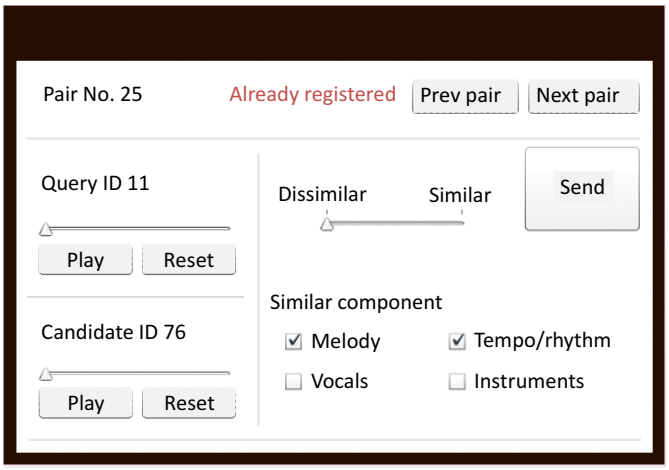

Fig. 1. Data collection interface.

annotation was used to obtain the starting point of the chorus sections.

C. Procedure

A subjective similarity data collecting system that works on a web browser was created for the experiment (Fig. 1). First, the system presented two songs ('query' and 'candidate', which we referred to as a "pair") to the subject. The presented pair was randomly chosen from 200 pairs selected before the experiment in each trial ('query' and 'candidate' were randomly chosen from the two songs of the pair). Each subject evaluated whether the pair was similar or dissimilar overall, and then whether each musical component (e.g., melody, tempo/rhythm, vocal, instruments) of the songs was similar or not. For example, if the subject felt the pair was dissimilar, but felt that the melody and tempo/rhythm were similar, the subject chose 'dissimilar' and checked 'melody' and 'tempo/rhythm' as similar components on the interface. On the other hand, if the subject felt the pair was similar and also thought 'tempo/rhythm' and 'instruments' were similar, the subjects chose 'similar' and checked 'tempo/rhythm' and 'instruments' on the interface. During the experiment, each subject repeated this process 200 times with 200 different pairs.

The subjects could replay the songs and re-evaluate the pair repeatedly. Time allowed for the experiment was 200 minutes (50 minutes $\times$ 4 sets). Subjects rested 10 minutes between sets.

D. Resulting data

Fig. 2 shows a histogram of the number of subjects who evaluated the pairs as similar. The number of subjects who evaluated the pairs as similar for each component (melody, tempo/rhythm, vocals, instruments) is also shown in Fig. 2. For each of the 200 pairs, the number of subjects who evaluated the pair as similar was calculated. This figure shows that many pairs were evaluated as dissimilar by many subjects, but that only a few pairs were evaluated as similar by many subjects. The existence of individuality is also shown by the existence

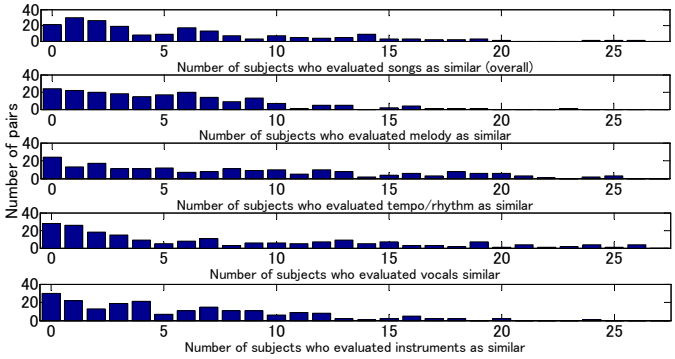

Fig. 2. Histograms of number of subjects who evaluated a pair as similar.

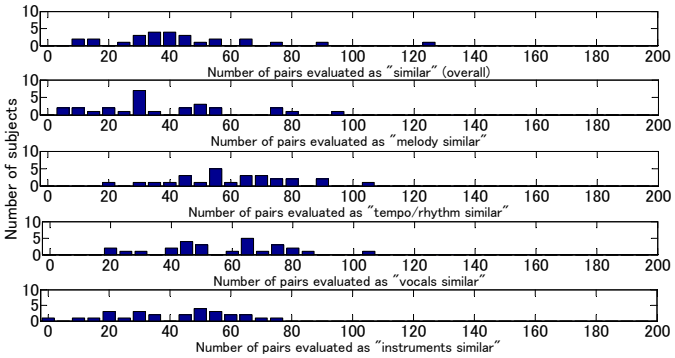

Fig. 3. Histograms of number of pairs evaluated as similar.

of pairs evaluated as similar by some subjects but as dissimilar by other subjects.

There could be many aspects of individuality related to subjective similarity. The number of times a subject evaluated pairs as similar is one of those aspects (i.e., some subjects thought many of the pairs, or components of the pairs, were similar, while others did not). Fig. 3 shows the histograms of the number of pairs evaluated as similar in 200 trials for each subject. Subjects who evaluated around 40 pairs as similar among 200 pairs were the majority (average 43.4 pairs). The average number of pairs evaluated as similar for each component were 38.7 pairs for melody, 61.1 pairs for tempo/rhythm, 56.6 pairs for vocals, and 41.5 pairs for instruments.

III. Calculating Acoustical Similarity

In this section, a method of calculating acoustic music similarity that corresponds well with subjective music similarity is considered, using the data collected in section II. It can be assumed that the relationship between subjective music similarity and acoustic music similarity differs from subject to subject. Therefore, in this section a method that can represent such differences is proposed. First, the acoustic characteristics of each song are represented as a feature vector. Second, the difference between the two feature vectors is calculated using weighted Euclidean distance:

$$||\mathbf{v}_i - \mathbf{v}_j||_{\mathbf{W}} = \sqrt{(\mathbf{v}_i - \mathbf{v}_j)^T \mathbf{W} (\mathbf{v}_i - \mathbf{v}_j)} \qquad (1)$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are d-dimensional feature vectors, and $\mathbf{W}$ is a $d \times d$ positive semi-definite weighting matrix. By optimizing weighting matrix $\mathbf{W}$ for each subject, the formula can be customized for each individual. To train this weighting matrix, metric learning techniques are adopted.

A. Feature vector

In order to calculate Euclidean distances between songs, each song should be represented by a vector. The methods used to represent the acoustic characteristics of each song are explained below.

First, short term features are extracted for each song. The extracted short term features are Mel Frequency Cepstrum Coefficients (13 coefficients are used), intensity [9], spectral centroid, spectral flux, spectral roll-off, and high frequency energy (also known as "brightness") [10]. In order to extract short time features other than intensity and spectral flux, MIR toolbox 1.3.2 [11] was used. For each feature, temporal differentials on time frame n were calculated by regression coefficient on small section:

$$\Delta \mathbf{x}(n) = \frac{\sum_{l=-L}^{L} l \cdot \mathbf{x}(n-l)}{\sum_{l=-L}^{L} l^2} \tag{2}$$

where $\mathbf{x}(n)$ is the feature vector at time frame n, and L is a parameter which determines the number of points used to calculate the regression coefficient. In this experiment, $L = 2$ was used. Second order differentials were also calculated as regression coefficients of first order differentials. These differentials are also used as short term features. The conditions of feature extraction are shown in TABLE I.

TABLE I
CONDITIONS OF SHORT TERM FEATURE EXTRACTION

Sampling frequency	16000 Hz
Window function	Hanning window
Window length	50 ms
Shift length	25 ms

Second, in order to represent each song as a vector, short term features of each song are summarized as global feature vectors. In this paper, two methods are used to summarize short term features. One method is by using vector quantization (VQ), the other is by using long term statistics.

1) Logarithmic relative histogram of VQ: First, the short term features are quantized using an LBG algorithm. By obtaining a relative histogram of centroids for each song, each song can be represented as a unique feature vector. Then, feature vectors are converted by calculating the logarithm of each bin.

In order to calculate VQ logarithmic relative histograms, the short term features referred to above are dimensionally reduced through Principal Component Analysis (PCA) to capture 95% of the variance. In this paper, dimensionally reduced short term features are quantized with a codebook size of 512.

2) Long term statistics: Another way of representing songs with vectors is by using long term statistics, based on the method presented in [12]. First, we calculate the K-point moving average $m(n,d)$ and standard deviation $s(n,d)$ for each time frame:

$$m(n,d) = \frac{1}{K} \sum_{k=1}^{K} x(n-k+1,d) \tag{3}$$

$$s(n,d) = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \{x(n-k+1,d) - m(n,d)\}^2} \tag{4}$$

where n is the time frame number, d is the dimensionality of the feature vector, and $x(n,d)$ is a short term feature. By calculating the time averages and standard deviations of $m(n,d)$ and $s(n,d)$, we get 4 d-dimentional vectors (time averages of $m(n,d)$ and $s(n,d)$ and standard deviations of $m(n,d)$ and $s(n,d)$). Concatenating these vectors into a vector, a song can be represented as a vector.

In order to calculate long-term statistics, the short-term features referred to above are used without PCA (unlike the logarithmic relative histogram of VQ). In this paper, the numbers of points K of moving average and standard deviation in equation (3), (4) were both 20 (which corresponds to 0.5 s).

B. Metric learning

To train optimally weighted Euclidean distance equations (1) for each subject, methods of metric learning are used. Metric learning techniques train the weighting matrix $\mathbf{W}$, so that pairs that were labeled as similar have a small distance function value, and pairs that were labeled as dissimilar have a large distance function value. In this study, two metric learning techniques were adopted, Metric Learning to Rank (MLR) [13] and Information-Theoretic Metric Learning (ITML) [14]. MLR trains the optimal distance function based on rank, that is, it sorts the songs according to trained distance function, so that "similar" songs are ranked higher than "dissimilar" songs. To train the optimal distance function with MLR, we used the MATLAB implementation of MLR (mlr-1.0)[1]. ITML trains the distance function, which is smaller than the given upper bound for similar pairs and larger than the given lower bound for dissimilar pairs, and regularizes weighting matrix $\mathbf{W}$ to be as close as possible to the identity matrix. To train the optimal distance function with ITML, we used the MATLAB implementation of ITML (itml-1.2)[2].

IV. EXPERIMENT

An experiment was conducted to measure how metric learning techniques improve the performance of subjective similarity estimation. The two types of feature vectors that were explained in section III-A and the two types of metric

[1] http://cseweb.ucsd.edu/~bmcfee/code/mlr/
[2] http://www.cs.utexas.edu/~pjain/itml/

learning techniques that were introduced in section III-B are used.

Using the subjective similarity evaluation data referred to in section II as the similarity label, an optimal distance function for each subject was trained. The experiment was conducted using 10-fold cross validation, i.e., we divided 80 songs into 10 sets averagely, trained the optimal distance function with 9 sets (72 songs), tested with the remaining set (8 songs) and then repeated this procedure 9 more times, each time changing which sets were used for training and test data. To measure the performance of the trained distance function, Area Under the ROC Curve (AUC) was used. In order to confirm how metric learning techniques improve the performance of subjective similarity estimation, Euclidean distance was used for comparison. For the purpose of confirming individuality of the optimal distance functions (relevance between subjective similarity and acoustic similarity), the trained distance functions were tested not only with the subject's own data (the same subject's data used for training), but also with another subject's data.

To train the distance function with MLR or ITML, we have to set the slack trade-off parameter. In our results, the AUC values shown were achieved using the best values of parameter $C \in \{10^{-2}, 10^{-1}, \ldots, 10^{6}\}$ in the metric learning algorithm for each case. To train the optimal distance function, we not only used the overall similarity evaluation data, but also the similarity data for each component (melody, tempo/rhythm, vocals, instruments).

A. Results

Fig. 4 shows the results of using VQ logarithmic relative histograms as feature vectors. It seems that there is no difference in subjective similarity evaluation performance before training (Euclidean) and after training (MLR and ITML) except for vocal similarity. For vocal similarity, the AUC value of "ITML" is greater than that of "ITML (tested with other subject's data)". This result leads us to believe that individuality in judging vocal similarity exists, and that this individuality can be represented using the weighted Euclidean distance between VQ logarithmic relative histograms. However, when we used MLR, the improvement in the AUC value was very small. Thus, it is shown that the selection of a metric learning algorithm is an important issue.

Fig. 5 shows the results of using long term statistics as feature vectors. In all categories (overall, melody, tempo/rhythm, vocals, instruments), subjective similarity evaluation performance was improved for MLR and ITML, in comparison with using Euclidean distance. However, it seems that there is no difference between "tested with the same subject's data" and "tested with other subject's data" for both MLR and ITML, except for overall and vocal similarity. For overall and vocal similarity, AUC values of "MLR" and "ITML" are greater than of "MLR (test with other subject's data)" and "ITML (test with other subject's data)", respectively. This leads us to believe that individuality judging overall and vocal similarity exists, and that this individuality can be represented using

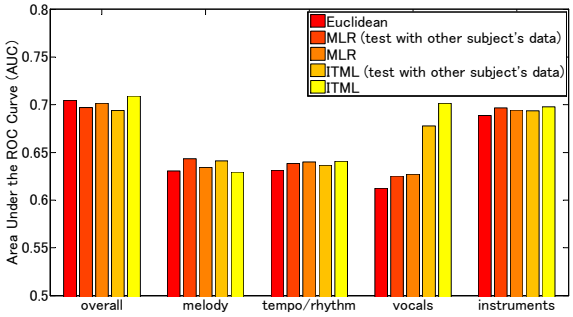

Fig. 4. Results for VQ logarithmic relative histogram.

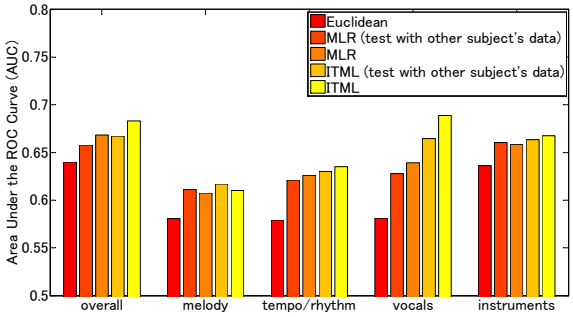

Fig. 5. Results for long term statistics.

weighted Euclidean distance between long term statistics. As with VQ logarithmic relative histograms, greater performance was confirmed when ITML was used as the metric learning algorithm.

In all the similarity categories, resulting AUC values for VQ logarithmic relative histograms were greater than for long term statistics. This result suggests that VQ logarithmic relative histograms are more appropriate for subjective similarity estimation than long term statistics.

V. Conclusions

In this paper, we described a method of data collection for the purpose of measuring subjective music similarity. The collected data confirmed that individuality in the judging of subjective music similarity exists. In particular, individuality was explicitly revealed by the wide variation in the number of times a subject evaluated song pairs as similar (e.g., Fig. 3).

Using the collected data, we adopted a method of metric learning to confirm the existence of individuality. When we used VQ logarithmic relative histograms as feature vectors, we could confirm individuality for vocal similarity. When we used long term statistics as feature vectors, we could confirm individuality for overall song and vocal similarity.

In future work, we should try to develop a more sophisticated model in order to improve subjective similarity estimation performance and better represent individuality; e.g., a model which considers the frequency of "similar" evaluations, uses acoustic features which weren't used in this paper (features related to rhythm, chords, etc.), and so on.

REFERENCES

[1] A. Raubel, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity," in *the 3th International Conference on Music Information Retrieval (ISMIR 2002)*, 2002, pp. 71–80.

[2] M. Goto and T. Goto, "Musicream: New music playback interface for streaming, sticking, sorting, and recalling musical pieces," in *the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, September 2005, pp. 404–411.

[3] K. Hoashi, K. Matsumoto, and N. Inoue, "Personalization of user profiles for content-based music retrieval on relevance feedback," in *ACM Multimedia*, 2003, pp. 110–119.

[4] F. Vignoli and S. Pauws, "A music retrieval system based on user-driven similarity and its evaluation," in *the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005.

[5] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph. D. thesis, Vienna University of Technology, Austria, 2006.

[6] J.J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, pp. 1–13, 2004.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, October 2002, pp. 287–288.

[8] M. Goto, "AIST annotaiton for the RWC music database," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, October 2006, pp. 359–360.

[9] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[10] P. N. Juslin, "Cue utilization in communication of emotion in music performance: relating performance to perception," *Journal of Exoerimental Psychology: Human Perception and Performance*, vol. 26, no. 6, pp. 1707–1813, 2000.

[11] O. Lartillot and P. Toiviainen, "MIR in MATLAB (II): A toolbox for musical feature extraction from audio," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, September 2007, pp. 127–130.

[12] G. Tzanetakis, "Marsyas submissions to MIREX 2010," in *the 6th Music Information Retrieval Evaluation eXchange (MIREX 2010)*, 2010.

[13] B. McFee and G. Lanckriet, "distance function learning to rank," in *Proceedings of the 27th Annual International Conference on Machine Learning*, J. Fürnkranz and Joachims. T., Eds., Haifa, Israel, June 2010, pp. 775-782.

[14] J. V. Davis, B. Kulis, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of International Conference on Machine Learning*, Corvallis, Oregon, USA, 2007, pp. 209–216.