MODELING SUBJECTIVE EVALUATION OF MUSIC SIMILARITY USING TOLERANCE

Shota Kawabuchi, Chiyomi Miyajima, Norihide Kitaoka and Kazuya Takeda

Nagoya University Graduate School of Information Science

ABSTRACT

In order to improve the automated retrieval of similar songs, we need to develop an estimation method which can measure their subjective similarity. In this study, we assume that subjective similarity of songs is determined by both the acoustical similarity of the songs and the individuality of the listener. We focus mainly on the individuality of listeners, and use knowledge about this individuality to develop a subjective similarity estimation model.

The results of our previous study suggest that the likelihood of someone judging two songs (a musical pair) as “similar” is influenced by the individual characteristics (individuality) of the listener. In this paper we refer to the likelihood of judging songs to be similar as the “tolerance” of the listener, and propose a model of subjective similarity evaluation which takes individual tolerance into account. In our experiment, we estimate listeners’ tolerance using subjective musical similarity evaluation data. We also conduct an experiment using a much smaller amount of similarity evaluation data to estimate tolerance, as this would be desirable for practical applications.

Index Terms**—** music similarity, subjective similarity, similarity evaluation, individuality

1. INTRODUCTION

With the emergence of mass storage media and the improvement of data compression technology, users can experience difficulty finding desired songs from a large database due to the quantity of data. Estimating subjective music similarity, and using this information for the retrieval of songs, is one possible solution to this problem.

Estimation of subjective similarity could be realized if we could extract acoustic features that are important for music perception and compare these features as humans do. However, it is difficult to estimate subjective music similarity from acoustic similarity because human music perception is not well understood. Furthermore, there is some individual variation among listeners regarding similarity perception, i.e., some listeners may feel two songs are similar, while other listeners may feel the same two songs are dissimilar. In this study we assume that subjective similarity of music is determined by two factors, the acoustical similarity of the song pairs and the individuality of the listener. The goal of our study is to shed light on the acoustic features which contribute to musical similarity, and on the individuality of listeners, and use this information to develop a method of subjective similarity estimation.

Studies to estimate music similarity have been conducted widely in the field of music information processing. For example, Pampalk [1] extracted Mel-Frequency Cepstrum Coefficients (MFCC) for short time frames, fitted Gaussian distributions or Gaussian Mixture Models (GMMs) to each song, and calculated Kullback-Leibler divergences between the songs as a measure of timbre similarity. Other acoustic features which have been used widely are spectral centroids and zero crossing rates as timbre features, chroma vectors as tonal features, fluctuation patterns [2], and rhythm histograms as rhythm features [3].

There have been several studies related to individuality for the purpose of music retrieval. These studies combined listeners’ preferences or individuality of similarity evaluation with objective similarity. Hoashi et al. [4] proposed a music recommendation method which employs users’ preferred songs or genres, and constructs vectors that reflect users’ preferences. Vignoli and Pauws [5] measured musical similarity by weighting a combination of five features; timbre, genre, tempo, mood and year. They tried to reflect individual preferences in their system, letting users set parameters manually. Lampropoulos et al. [6] proposed a system that retrieved similar songs using a neural network. Their system uses the acoustical features of songs as input and optimizes neural networks with the users’ rankings and similarity ratings for the retrieved songs. They assumed that the features that contribute to music perception are different for each individual, so they used feature subsets to allow users to choose optimal subsets. Kawabuchi et al. [7] conducted an experiment in which subjects evaluated the musical similarity of song pairs as similar or dissimilar, and the collected data was then used to train distance functions between songs that reflected each subject’s individual similarity judgment.

In [7], they indicated that the frequency of making a “similar” judgment varied widely between subjects (Fig. 1). This result implies that there is a large amount of individual variation in the likelihood of judging a musical pair to be “similar”. However, the musical similarity model used in that paper did

This work was supported by JSPS KAKENHI Grant Number 25540168.

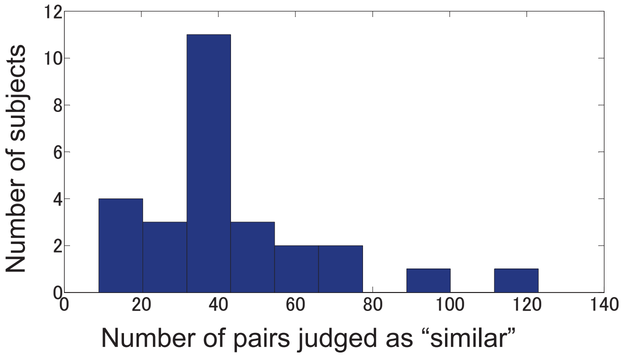

Fig. 1. Histogram of number of pairs judged as "similar". This indicates how many pairs were judged as similar in 200 trials (average 43.4 times). Most of the subjects are concentrated around the average, however, outlier subjects also exist (least 9 times, most 123 times).

not make use of this observation. In this paper, we refer to the likelihood of judging a musical pair to be similar as the " tolerance " of the listener, and propose a model of subjective similarity evaluation which takes this tolerance into account.

Our subjective similarity evaluation model is proposed in Section 2. In Section 3, we adapt the proposed model by taking into account subjective evaluation data, and estimate each subject's tolerance. The paper is concluded in Section 4.

2. SUBJECTIVE EVALUATION MODEL

We assume that the outcome of a similarity evaluation is determined by the acoustic similarity of a musical pair and by the tolerance of the listener. If a pair of songs is judged to be "similar" by most listeners, a listener who has a high level of tolerance should also judge that musical pair to be similar. Musical pairs which are near a "decision boundary" are sometimes judged as "similar", but at other times judged as "dissimilar". In light of these facts, we model the probability that a listener i judges a musical pair j to be "similar" with a logistic function:

$$p(e_{ij}=1|s_i,p_j)=\frac{1}{1+e^{-(s_i+p_j)}} \qquad (1)$$

where e_{ij} is the result of a similarity evaluation by listener i in regards to musical pair j. $e_{ij}=1$ means that listener i judged musical pair j to be "similar" and $e_{ij}=0$ means that listener i judged musical pair j to be "dissimilar". s_i and p_j represent the tolerance of listener i and the similarity of musical pair j, respectively. If s_i is large, listener i is likely to judge a pair of songs to be similar. If p_j is large, musical pair j is likely to be judged to be similar.

2.1. Parameter estimation

Suppose we have data on the similarity judgments of M subjects in regards to N musical pairs. If we apply our subjective evaluation model to the data, the likelihood of the model can be expressed as follows:

$$\begin{aligned} & p(\mathbf{E}|\mathbf{s},\mathbf{p}) \\ = & \prod_{i=1}^{M}\prod_{j=1}^{N} p(e_{ij}=1|s_i,p_j)^{e_{ij}}p(e_{ij}=0|s_i,p_j)^{1-e_{ij}} \end{aligned} \qquad (2)$$

where $\mathbf{E}=(e_{11},\dots,e_{ij},\dots,e_{MN}),\mathbf{s}=(s_1,\dots,s_M)$ and $\mathbf{p}=(p_1,\dots,p_N)$. By maximizing this likelihood, we can estimate the optimal values for $\mathbf{s}$ and $\mathbf{p}$.

Notice that this model has a problem: if musical pair j is evaluated as " dissimilar " by all listeners (our experimental data includes many such pairs, in fact), the optimal p_j is clearly $-\infty$. To avoid this problem, we assume that parameters $\mathbf{s}$ and $\mathbf{p}$ follow some prior distributions, $p(\mathbf{s})$ and $p(\mathbf{p})$, and we estimate $\mathbf{s}$ and $\mathbf{p}$ which maximize posterior probability $p(\mathbf{s},\mathbf{p}|\mathbf{E})$. In this paper, we assume that the prior distributions of tolerances $s_i(i=1,\dots,M)$ and similarities $p_j(j=1,\dots,N)$ follow independent and identically distributed Gaussian distributions $\mathcal{N}(s_i|\mu_s,\sigma_s^2)$ and $\mathcal{N}(p_j|\mu_p,\sigma_p^2)$, respectively. To maximize the posterior probability $p(\mathbf{s},\mathbf{p}|\mathbf{E})$, we should maximize the following equation:

$$\begin{aligned} & p(\mathbf{s})p(\mathbf{p})p(\mathbf{E}|\mathbf{s},\mathbf{p}) \\ = & \prod_{i=1}^{M}\mathcal{N}(s_i|\mu_s,\sigma_s^2)\prod_{j=1}^{N}\mathcal{N}(p_j|\mu_p,\sigma_p^2) \\ & \prod_{i=1}^{M}\prod_{j=1}^{N} p(e_{ij}=1|s_i,p_j)^{e_{ij}}p(e_{ij}=0|s_i,p_j)^{1-e_{ij}}. \end{aligned} \qquad (3)$$

To estimate the values for $\mathbf{s}$ and $\mathbf{p}$ which maximize (3), we set the objective function as follows:

$$\begin{aligned} L &= \log\{p(\mathbf{s})p(\mathbf{p})p(\mathbf{E}|\mathbf{s},\mathbf{p})\} \\ &\propto -\sum_{i=1}^{M}\frac{(s_i-\mu_s)^2}{2\sigma_s^2}-\sum_{j=1}^{N}\frac{(p_j-\mu_p)^2}{2\sigma_p^2} \\ &\quad -\sum_{i=1}^{M}\sum_{j=1}^{N}\Big[(1-e_{ij})\,(s_i+p_j) \\ &\qquad +\log\left\{1+e^{-(s_i+p_j)}\right\}\Big]. \end{aligned} \qquad (4)$$

Partially differentiating L with respect to $s_i(i=1,\dots,M)$ and $p_j(j=1,\dots,N)$,

$$\begin{aligned} \frac{\partial}{\partial s_i}L &= \sum_{j=1}^{N}\left[\frac{e^{-(s_m+p_j)}}{1+e^{-(s_m+p_j)}}-(1-e_{ij})\right] \\ &\quad -\frac{1}{\sigma_s^2}(s_i-\mu_s), \end{aligned} \qquad (5)$$

$$\frac{\partial}{\partial p_j}L = \sum_{i=1}^{M}\left[\frac{e^{-(s_i+p_n)}}{1+e^{-(s_i+p_n)}} - (1-e_{ij})\right] - \frac{1}{\sigma_p^2}(p_j-\mu_p). \qquad (6)$$

Setting (5) and (6) to equal 0, the optimal parameters for $\mathbf{s}$ and $\mathbf{p}$ can be obtained as the solution of the system of equations.

2.2. Parameter estimation algorithm

To calculate the values of $\mathbf{s} = (s_1, \ldots, s_M)$ and $\mathbf{p} = (p_1, \ldots, p_N)$ which maximize objective function L, we update the parameters iteratively as follows:

$$s_i \leftarrow s_i + \eta\frac{\partial}{\partial s_i}L \qquad (i = 1, 2, \ldots, M) \qquad (7)$$

$$p_j \leftarrow p_j + \eta\frac{\partial}{\partial p_j}L \qquad (j = 1, 2, \ldots, N) \qquad (8)$$

where η is the learning rate. By repeating this procedure until objective function L converges, estimators of $\mathbf{s}$ and $\mathbf{p}$ can be obtained.

3. APPLYING MODEL TO SUBJECTIVE EVALUATION DATA

By applying our subjective evaluation model to subjective musical similarity evaluation data, the tolerance of subjects and the similarity of musical pairs were estimated. We also estimated tolerance using a smaller number of evaluations, because it is desirable in practice if the tolerance of listeners can be estimated with less evaluation data.

3.1. Data used for experiment

For this experiment, we used subjective similarity evaluation data[7][1] for 200 musical pairs chosen from 80 popular music songs in the RWC music database [8]. The experimental procedure was as follows. First, each subject listened to two songs (a musical pair) and then evaluated their similarity as "similar" or "dissimilar". The subject then selected musical components (melody, tempo/rhythm, vocals, instruments) which they felt as similar. Each subject evaluated the same 200 musical pairs. The number of subjects who participated in the experiment was 27 (13 males and 14 females).

We only used the overall similarity evaluation data in this study, i.e., data collected on the similarity of the musical components of the songs was not used. Using this data as $\mathbf{E}$, we estimated the tolerance of the subjects and the similarity of the musical pairs.

3.2. Applying subjective evaluation model

We applied the subjective evaluation model to the data described in 3.1. We estimated parameters $s_i (i = 1, \ldots, M)$ and $p_j (j = 1, \ldots, N)$ for each subject and pair using the parameter estimation algorithm described in 2.2. We used the Gaussian random numbers ($\mu = 0$ and $\sigma^2 = 1$) as the initial values of $s_i (i = 1, \ldots, M)$ and $p_j (j = 1, \ldots, N)$ for the iterative algorithm. Parameters of prior distributions $p(\mathbf{s})$ and $p(\mathbf{p})$ were $\mu_s = 0, \sigma_s^2 = 1, \mu_p = 0$ and $\sigma_p^2 = 1$. Iterations were performed until the objective function L converged.

The estimated tolerances and similarities are shown in Fig. 2. The lines in the figure represent decision boundaries (the planes on which the probability of a pair being judged similar or dissimilar is 50%). The ratio of correct discrimination to the total number of evaluations (ratio of the number of "similar" points at which $s_i + p_j > 0$ plus the number of "dissimilar" points at which $s_i + p_j \leq 0$) was 0.842. From these results we can see that while "dissimilar" points could be discriminated with a high degree of accuracy (the discrimination rate was 95.8%), "similar" points could not be discriminated as well (the discrimination rate was 42.5%). This result was probably caused by the large difference in the number of "dissimilar" and "similar" points. The number of pairs judged to be "similar" was 1171, while the number of pairs judged to be "dissimilar" was 4229, out of 5400 (27×200) total evaluations.

3.3. Estimation of tolerance using a small number of evaluations

In order to predict subjective evaluations using our model, the tolerance of the subject and similarity of a musical pair should be known. Similarity of the pairs can be estimated using the acoustic features of the songs, while the tolerance of listeners can be estimated using a listener's prior evaluations. However, in practice, requiring listeners to evaluate a large number of musical pairs is not realistic. Therefore, it would be better if tolerance estimation could be done using the evaluation data of only a small number of musical pairs. In 3.2 we estimated tolerance and similarity by applying a subjective evaluation model to subjective evaluation data for 27 subjects $\times$ 200 musical pairs. In this section, we conduct an experiment to estimate the tolerance of subjects using evaluations and given similarities of smaller numbers of pairs chosen from the original 200 pairs. Before we describe the experiment, we will explain the two estimation methods used.

3.3.1. Maximum likelihood

Suppose tolerance of subject i is s_i and similarities of n musical pairs chosen from 200 musical pairs are $\mathbf{p} = (p_1, \ldots, p_n)$. $\mathbf{p}$ has already been estimated using the acoustic features of the songs. The optimal parameters for s_i which maximize posterior probability $p(\mathbf{s}, \mathbf{p}|\mathbf{E})$ can be calculated by maximizing

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SSimRWC/

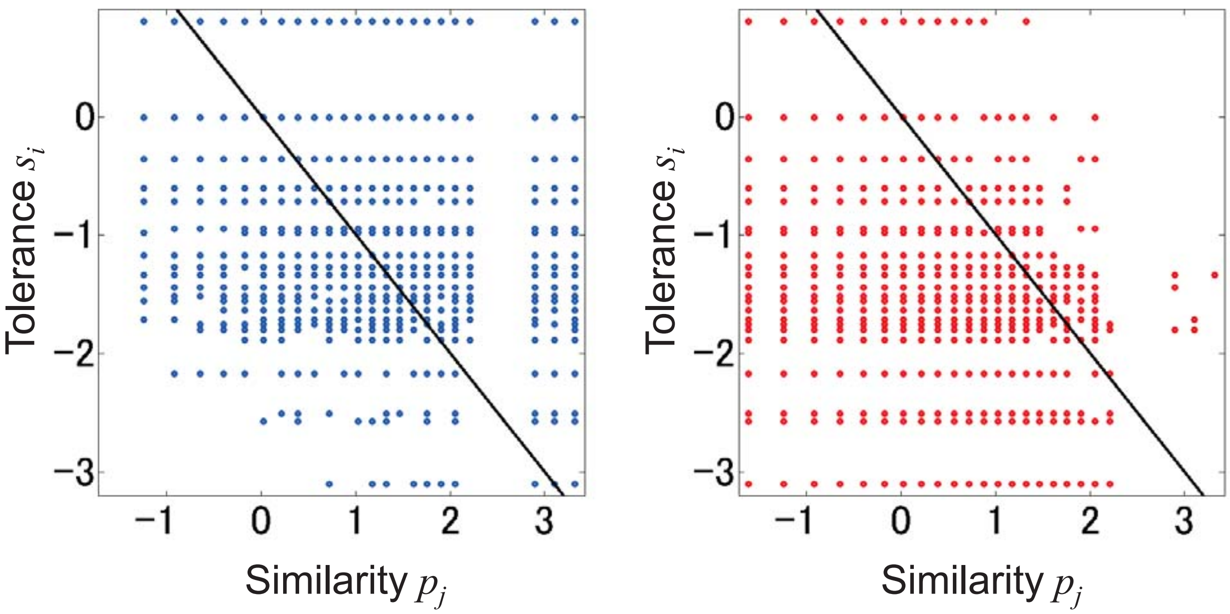

Fig. 2. Scatter plots of “similar” points (left figure) and “dissimilar” points (right figure) on the tolerance-similarity plane. Vertical axis is subject’s tolerance, and horizontal axis is a pair’s similarity. Lines on the figure represent discrimination surfaces $s_i + p_j = 0$, and it is desirable that “similar” points are on the right side of this surface and “dissimilar” points are on the left side.

the objective function (4) as in 2.2.

3.3.2. Evaluation patterns

We can express the similarity evaluations of subject i for n musical pairs as $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in})$, where $e_{ij} = 1$ for “similar” evaluations and $e_{ij} = 0$ for “dissimilar” ones. If we have evaluation data for a number of listeners and their tolerance has already been calculated, we can estimate the tolerance of an unknown listener using these evaluation patterns. If we calculate the Hamming distance between the evaluation patterns of the unknown listener and the patterns of known listeners, we can then use the tolerance of the known listener with the most similar Hamming distance to the unknown listener as the estimated tolerance of the unknown listener.

3.4. Prediction of subjective evaluation

An experiment to estimate tolerance with a small number of subjective evaluations was conducted. The experimental procedure was as follows. First, 200 musical pairs were divided into 10 groups. Then one of these groups of pairs was chosen for testing. Next, varying numbers of pairs were chosen randomly from the remaining nine groups for estimating tolerance. A subject’s tolerance was then estimated using the methods explained in 3.3.1 and 3.3.2. After tolerance was estimated, subjective evaluations for 20 test pairs were predicted and prediction accuracy was calculated. The numbers of pairs used for tolerance estimation were 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. This procedure was repeated for all subjects and all groups of musical pairs (27 subjects × 10 groups of song pairs).

As similarities $\mathbf{p}$ are necessary for prediction, the following two similarities were used:

- Euclidean distances between acoustic feature vectors (logarithmic VQ histogram) [7]
- Similarities p_j estimated using data from 26×200 evaluations (for comparison; these can be considered as the ideal similarities)

The method of calculating acoustical feature vectors was as follows. Mel Frequency Cepstrum Coefficients (1-13 coefficients), intensity [9], spectral centroid, spectral flux, spectral roll-off, and high frequency energy [10] were extracted as short-term features. For each feature, first and second order temporal differentials were calculated and used as short-term features. The short-term features referred to above are dimensionally reduced through Principal Component Analysis (PCA) to capture 95% of the variance and then 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.

Experimental results are shown in Fig. 3. Mean discrimination rates increase as the number of pairs used for training increases under all conditions. With the maximum likelihood method, we need 20 to 30 musical pairs for training to reach the highest level of prediction accuracy, while we need less than 10 pairs using the evaluation pattern method to

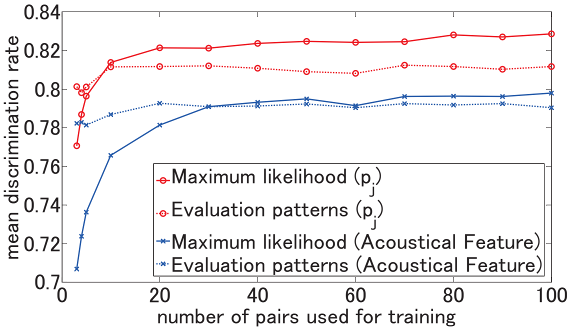

Fig. 3. Plot of the mean discrimination rates according to the number of pairs used for training. Solid lines represent results obtained using maximum likelihood method to estimate tolerance. Dotted lines represent results obtained using the evaluation pattern method. Red lines with circular markers represent results obtained using similarity p_j and blue lines with x-markers represent results obtained using acoustical features.

achieve an acceptable level of accuracy. However, the method which uses evaluation patterns achieved lower prediction accuracy than the maximum likelihood method. We believe this is because there are a few outlying subjects whose subjective evaluation data was not similar to that of any other subjects. Therefore, their evaluations could not be predicted with a high degree of accuracy using the evaluation pattern method.

4. CONCLUSION

In this paper we proposed a model for estimating the subjective similarity of songs by incorporating a listener's tolerance as a parameter. This model assumes that subjective similarity evaluations are determined by the tolerance of the listener and by the acoustic similarity of the musical pair. We explained how we formulated this model and how we developed our parameter estimation algorithm. We used subjective similarity evaluation data to estimate the parameters of our model, and the resulting model was able to achieve an evaluation prediction rate of 84.2%. We also conducted an experiment that estimated listeners' tolerance using a smaller number of evaluations. We proposed two methods to estimate tolerance, maximum likelihood and a method which uses evaluation patterns. Results showed that the method using evaluation patterns could achieve a high level of prediction accuracy even when the number of pairs used for training was small (i.e., less than 10).

The subjective similarity evaluation model proposed in this paper assumes that all listeners perceive one identical type of music similarity, and that the difference between listeners is based only on differences in tolerance. This is probably an unrealistic assumption, however. Actual listeners are likely to perceive multiple types of similarities (e.g., similarities in timbre, rhythm, and so on) and tolerances are likely to vary in regards to each perceptual similarity. Therefore, in our future work it is important to extend our subjective similarity evaluation model to include multiple similarities and tolerances.

5. REFERENCES

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

[2] E. Pampalk, A. Rauber, and D. Merkel, "Content-based organization and visualization of music archives," in *ACM Multimedia*, December 2002, pp. 1–6.

[3] T. Lidy and A. Rauber, "Evaluation of feature extractions and psycho-acoustics transformations for music genre classification," in *the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005, pp. 34–41.

[4] 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.

[5] 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, pp. 272–279.

[6] A.S. Lampropoulos, D.N. Sotiropoulos, and G.A. Tsihrintzis, "Individualization of music similarity perception via feature subset selection," *IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, pp. 552–556, 2004.

[7] S. Kawabuchi, C. Miyajima, N. Kitaoka, and K. Takeda, "Subjective similarity of music: Data collection for individuality analysis," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012, pp. 1–5.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, October 2002, pp. 287–288.

[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 Experimental Psychology: Human Perception and Performance*, vol. 26, no. 6, pp. 1707–1813, 2000.