A Musical Mood Trajectory Estimation Method Using Lyrics and Acoustic Features

Naoki Nishikawa† nnishika@kuis.kyoto-u.ac.jp
Katsutoshi Itoyama† iytoyama@kuis.kyoto-u.ac.jp
Hiromasa Fujihara‡∗ h.fujihara@aist.go.jp
Masataka Goto‡ m.goto@aist.go.jp
Tetsuya Ogata† ogata@kuis.kyoto-u.ac.jp
Hiroshi G. Okuno† okuno@kuis.kyoto-u.ac.jp

† Dept. of Intelligence Science and Technology, Grad. School of Informatics, Kyoto University

‡ National Institute of Advanced Industrial Science and Technology (AIST)

∗ School of Electronic Engineering and Computer Science, The Queen Mary University of London

ABSTRACT

In this paper, we present a new method that represents an overall musical time-varying impression of a song by a pair of mood trajectories estimated from lyrics and audio signals. The mood trajectory of the lyrics is obtained by using the probabilistic latent semantic analysis (PLSA) to estimate topics (representing impressions) from words in the lyrics. The mood trajectory of the audio signals is estimated from acoustic features by using the multiple linear regression analysis. In our experiments, the mood trajectories of 100 songs in Last.fm's Best of 2010 were estimated. The detailed analysis of the 100 songs confirms that acoustic features provide more accurate mood trajectory and the 21% resulting mood trajectories are matched to realistic musical mood available at Last.fm.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Modeling

General Terms

Algorithm, Design

Keywords

musical mood representation, musical mood estimation, time-varying impression, mood trajectory, lyrics and audio signals music information retrieval

1. INTRODUCTION

Musical mood estimation is gaining increasing attention in recent years in musical information retrieval (MIR) research [3]. Human listeners recognize music not only from the acoustic features but also from the expression of the emotions. MIR systems based on the similarity of musical moods are expected to provide more intuitive retrieval results with human sensibility than systems based on only acoustic features [1, 2]. Various methods for musical mood estimation [4–12] have been reported, and these methods are still being developed. This is because musical moods are highly subjective and difficult to quantify [3].

We believe that the following two aspects are important for musical mood representation:

1. musical moods depend on both lyrics and audio signals
2. musical moods have time-varying characteristics.

Regarding the former, lyrics have a large influence on musical moods because lyrics represent these moods linguistically. Some musical pieces have contradicting moods in their lyrics and audio signals, e.g., a happy melody with melancholic lyrics. To represent these musical moods, we need both lyrics and audio signals. In fact, the accuracy of musical mood detection is improved by using both lyrics and audio signals [8–10]. Regarding the latter, humans receive time-varying impressions from music [11,12]. Some pieces of music have highly different moods between sections such as the verses and choruses. To represent these musical moods, we need to model their time-varying aspect.

In previous MIR research [4–12], there was no research dealing with these two points to the authors' knowledge. Research using only lyrics or audio signals [4–7] cannot reflect musical moods depend on both lyrics and audio signals. Research using both lyrics and audio signals [8–10] focused on static musical moods. Research focusing on time-varying musical moods [11, 12] used only audio signals.

In this paper, a time-varying musical mood (a mood trajectory) estimation method using both lyrics and acoustic features is proposed. We estimate the mood trajectories of lyrics and audio signals separately. In other words, we represent time-varying musical moods not as a mixture of lyrics and audio trajectories but as a pair of these two trajectories. This multilateral musical mood representation can describe conflicts between the complex time-varying musical moods of the lyrics and audio signals. We assume that humans feel a constant mood in a certain musical section defined as a phrase. Under this assumption, the mood trajectories

MIRUM'11, November 30, 2011, Scottsdale, Arizona, USA.

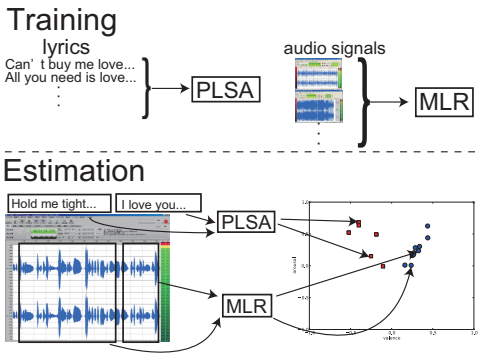

Figure 1: Overview of mood trajectory estimation method

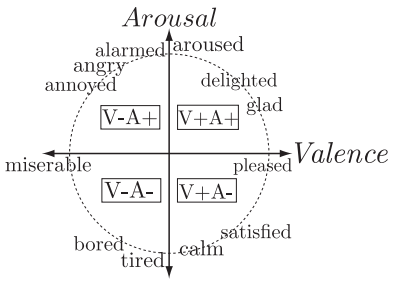

Figure 2: Russell's circumplex model

are estimated as sequences of coordinates on V-A space [13] from lyrics and audio signals divided into phrases. The V-A space has two axes: valence (positive-to-negative) and arousal (high-to-low energy).

A mood trajectory of lyrics is estimated by the probabilistic latent semantic analysis (PLSA) [14]. As the normal PLSA focuses on the co-occurrence data of documents and words, estimated latent topics are not certain to represent moods. To obtain emotional representation from a PLSA, topics are mapped to V-A space by using prior knowledge. A mood trajectory of an audio signal is estimated by multiple linear regression (MLR) [11,12]. Multi-dimensional acoustic features for each musical phrase are mapped to V-A space.

The rest of this paper is organized as follows. In Section 2, we take a general view of our method. Section 3 and Section 4 explain a mood trajectory estimation method for lyrics and audio signals. In Section 5, we analyze results of an experimental evaluation and confirm effectiveness of our method. Finally, Section 6 conclude and summarize the main findings.

2. OVERVIEW OF A MOOD TRAJECTORY ESTIMATION

The goal of this paper is to estimate a musical mood trajectory, i.e. the time-varying musical moods of lyrics and audio. The input are lyrics and an audio signal of a song, and the output are the mood trajectories of lyrics and an audio signal. Though our method can be applied to songs from arbitrary musical genres with lyrics, we used popular songs for our experimental evaluation. Figure 1 shows an overview of our method.

We define "a phrase" as a certain section in which humans feel a constant mood. Note that boundaries between phrases and those between structural sections (e.g., verses, choruses) are not always the same though they tend to be similar to each other. In addition, the numbers of phrases in audio signals do not necessarily coincide with those in lyrics. In this paper, we assume that the input lyrics and audio signals were manually divided into phrases in advance.

The musical mood of each phrase is defined as the coordinates on V-A space [13]. V-A space is a psychological model that represents an emotional state as coordinates on two-dimensional space (see Figure 2). The horizontal axis (valence) ranges from positive to negative, and the vertical axis (arousal) ranges from high to low energy. Therefore, the mood trajectories of lyrics and an audio signal are estimated as sequences of coordinates on V-A space from lyrics and an audio signal divided into phrases. The musical mood of an input song is represented by combining the mood trajectories of lyrics and an audio signal.

The mood trajectory of lyrics is estimated by using a PLSA [14] constrained by affective words as prior knowledge. The PLSA is a probabilistic generative model based on the co-occurrence data of documents (lyrics) and words. This model is used to estimate semantic relations between words and semantic topics of documents. As the PLSA is based only on the co-occurrence data of documents (lyrics) and words, a lot of training data can be used without labeling. Because of this, the PLSA is widely used in natural language processing [15,16].

As the normal PLSA focuses on the co-occurrence data of documents and words, the estimated semantic relations and latent topics are not certain to represent moods. By using affective words as prior knowledge and estimating parameters of the PLSA by maximum a posterior (MAP) estimation, the estimated semantic relations and latent topics are mapped on V-A space.

The mood trajectory of an audio signal is estimated by using a multiple linear regression (MLR). The input variables are the phrases' acoustic features, and the output variables are the phrases' V-A coordinates. By using a MLR, acoustic features can be mapped on V-A space. In a training phase, combination of phrases and V-A coordinates of phrases are made manually as training data, and the MLR regressor is trained to relations between acoustic features and V-A coordinates of phrases. In an estimating phase, the V-A coordinates of phrases are calculated from the acoustic features of phrases by a trained MLR regressor, and the estimated V-A coordinates construct a mood trajectory.

3. MOOD TRAJECTORY ESTIMATION FOR LYRICS

We estimate a mood trajectory of lyrics by estimating the V-A coordinates of each word in the lyrics, calculating the V-A coordinates of phrases, and plotting the V-A coordinates of phrases. We estimate the V-A coordinates of each word in the lyrics by using a PLSA [14] based on a MAP estimation using prior knowledge. Our method assumes that the lyrics were manually segmented to phrases before using a PLSA since it requires certain amount of lyrics to estimate an appropriate mood stably. It is one of our future works to preclude this presumption. PLSA estimates mood of each word of lyrics and moods of phrases are calculated by summing up moods of words of segmented phrases.

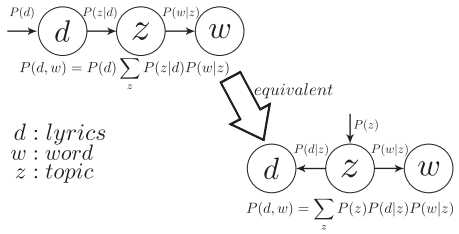

Figure 3: Graphical model of a probabilistic latent semantic analysis (PLSA)

3.1 Probabilistic Latent Semantic Analysis

We estimate a mood trajectory of lyrics by using a PLSA. The PLSA is a probabilistic generative model used to estimate the semantic relations between words and semantic topics of documents. The co-occurrent probabilities of documents and words are associated with latent variables (Figure 3). A latent variable z is assumed to be observed from a document d, and a word w is assumed to be generated from z. A latent variable z can be understood as a topic of documents. Using d, w, and z, a co-occurrence probability $P(d, w)$ is defined as:

$$P(d, w) = P(d) \sum_z P(z|d)P(w|z)$$
$$= \sum_z P(z)P(d|z)P(w|z).$$

$P(w|z)$ is the observed probability of words from topics, and words having a high $P(w|z)$ represent topics of z.

3.2 MAP Estimation of Parameters of PLSA

For a mood trajectory estimation for lyrics, it is necessary for z to represent the moods of the lyrics. However, z does not always represent a mood because normal PLSA focuses on the co-occurrent data of documents and words, and estimates the semantic relations between words and semantic topics of documents. We condition on z representations by maximum a posterior (MAP) estimation of PLSA parameters by using affective words such as happy and sad of which the V-A coordinates are already known.

Here, z is defined as:

$$z \in \{\text{V+A+}, \text{V+A}-, \text{V}-\text{A+}, \text{V}-\text{A}-\}.$$

Each z represents a distinct quadrant of V-A space. Prior distribution of $P(w|z)$ is defined as:

$$P(w, z|\alpha_{w,z}) \propto \prod_w \prod_z P(w|z)^{\alpha_{w,z}-1}.$$

By setting $\alpha_{w,z}$ for affective words and increasing $P(w|z)$ for affective words, each z can represent each quadrant of V-A space. For example, if $z_1 = \text{V+A+}$, we set $\alpha_{w,z}$ of words on the V+A+ quadrant such as happy and glad.

As prior knowledge, ANEW [17] and WordNet [18] are used. ANEW is constructed during psycholinguistic experiments and contains 1,034 words that have coordinates on V-A space. WordNet is a lexical database for English. This links English words to sets of synonyms called synsets, and synsets are linked to each other through semantic relations. By using ANEW and WordNet, $\alpha_{w,z}$ is set according to the following procedure:

1. By searching the synsets of each word of ANEW and extending ANEW with synonyms, ANEW is extended from 1034 to 9757 words.

2. $\alpha_{w,z}$ of words of ANEW on any quadrant of V-A space is set. If a word of ANEW is on the origin of V-A space, $\alpha_{w,z}$ is set to 1. According to the distance between each word of ANEW and origin of V-A space, $\alpha_{w,z}$ is set ranging from 1 to 1.01. We compare the estimated V-A coordinates of words of ANEW in the training data and original coordinates of ANEW, and decide the max of $\alpha_{w,z}$ for accurate estimation.

Each musical phrase's coordinates on V-A space are calculated by:

$$V = \frac{1}{K} \sum_{k=1}^{K} ((P(\text{V+A+}|w_k) + P(\text{V+A}-|w_k)) - (P(\text{V}-\text{A+}|w_k) + P(\text{V}-\text{A}-|w_k)))$$

$$A = \frac{1}{K} \sum_{k=1}^{K} ((P(\text{V+A+}|w_k) + P(\text{V}-\text{A+}|w_k)) - (P(\text{V+A}-|w_k) + P(\text{V}-\text{A}-|w_k))).$$

Here, K is the number of words of musical phrases, and $P(z|w)$ can be calculated from $P(w|z)$, $P(z)$, and $P(w)$. Calculated V-A coordinates are plotted on V-A space, and a mood trajectory of lyrics are described.

4. MOOD TRAJECTORY ESTIMATION FOR AUDIO SIGNALS

We estimate a mood trajectory of an audio signal by using a multiple linear regression (MLR). We train a MLR regressor based on acoustic features and V-A coordinates of each phrase by using a prepared training data set. The V-A coordinates of phrases are estimated from acoustic features by a trained regressor, and mood trajectories are described by plotting estimated V-A coordinates. As with the case of mood trajectory for the lyrics, it is necessary to segment the audio signals to phrases beforehand. We input acoustic features of each phrase to a trained regressor to estimate each V-A coordinates of each phrase.

4.1 Training data collection

As a training data set, V-A coordinates and phrase switching times are needed. We developed the graphical interface seen in Figure 4 for data collection. We referred to the interactive game developed by Kim et.al. [19] to develop this interface. The horizontal axis represents valence (positive vs. negative), and the vertical axis represents arousal (high vs. low energy). Users listen to music and click on this interface when they recognize that moods are switching. The collected data are saved as matrices, and the V-A coordinates and phrase switching time are stored in each row.

4.2 Acoustic Features

The accuracy of musical mood detection is increased by using multiple acoustic features [3]. With this knowledge, we select significant acoustic features from the multiple features by using a principal component analysis (PCA).

First, we extract some acoustic features from each of the musical phrases according to the previous mood detection

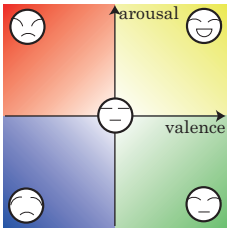

Figure 4: Graphical interface for manual annotation

Table 1: Extracted 71 acoustic features

Acoustic features	Description
Statistical spectrum descriptors	Includes spectral centroid, flux, roll-off, and flatness. They represent spectrum shape and reflect timbral features (27 dimensions).
Mel-frequency cepstral coefficients (MFCCs)	Short-term and low-dimensional features based on spectrum. These features are used in the modeling of audio signals etc. [21] (13 dimensions).
Chroma vector	The sum of the magnitude at some octaves divided into 12 divisions corresponding to 12 pitch classes [22] (12 dimensions).
Line spectral pair (LSP)	The feature used to represent linear prediction coefficients (18 dimensions).
Zero crossing	The number of times of crossing zero points of a waveform (1 dimension).

methods [8–12]. The extracted features and descriptions are in Table 1. These features are extracted using a 32-ms frame (no overlap), and the means and variances of these features are calculated as a 142 dimensional feature vector. We extract acoustic features from stereo audio signals at a sampling rate of 44.1 kHz using MARSYAS [20]. The 142 dimensional vectors are normalized (mean = 0, variance = 1) and compressed to 18 dimensions by PCA (cumulative contribution ratio = 90%). We use these 18 dimensional vectors as acoustic features.

First, second and third principal components and the contributing acoustic features are in Table 2. The acoustic features of which the absolute value of the principal component axis is in the top 5 are selected as contributing features. According to Table 2, MFCC, LSP, chroma vector, and spectral centroid contribute principal components that have a high contribution ratio.

4.3 Multiple Linear Regression

The V-A coordinates of each phrase are estimated from acoustic features by a multiple linear regression (MLR) using a polynomial basis function. The input variables are the acoustic features of phrases, and the output variables are the V-A coordinates of phrases. The regression function is defines as:

$$V = \sum_{i=0}^{M-1} \sum_{j=1}^{22} v_{ij}(x_j)^i,\ A = \sum_{i=0}^{M-1} \sum_{j=1}^{22} a_{ij}(x_j)^i.$$

Here, x_j is the jth element of a 22 dimensional feature vector, v_{ij} and a_{ij} are parameters of MLR, and M is the number of parameters. According to preliminary experiments, M is set to four for the smallest prediction error. By using training data, we estimate v_{ij} and a_{ij} and train an MLR regressor. With the trained MLR regressor, we estimate the V-A coordinates of the phrases of input songs from the acoustic features. The estimated V-A coordinates are plotted on V-A space, and the audio's mood trajectories are described.

Table 2: Three major principal components, their contribution ratios of three components, and their contributing acoustic features

	Principal components		
	First	Second	Third
Contribution ratio	48.5%	10.6%	8.06%
Contributing features	Centroid MFCC	Chroma vector LSP	Chroma vector LSP

5. EXPERIMENTAL EVALUATION

We conducted an experimental evaluation to confirm the validity of our method by

1. investigating how songs are classified by mood trajectories,
2. validating that mood trajectories can represent time-varying musical moods, and
3. confirming that mood trajectories can be matched to mood tags selected from social tags.

We classified songs based on the forms and positions of mood trajectories by clustering based on the similarity of the trajectories. The variance of acoustic features in each class was compared to validate the time-varying mood expressiveness of the mood trajectories. Moreover, the mood trajectories and mood tags of each class were compared to confirm the accuracy of the mood trajectories.

As a dataset, we used 100 songs taken from the top 20 albums of "Last.fm's Best of 2010" (`http://www.last.fm/bestof2010/about`). We first segmented the lyrics and audio signals of these songs into phrases. For audio signals, three university students listened to the songs and labeled the VA coordinates of each of the phrases and phrase switching times by using GUI. Lyrics were segmented into phrases based on line feed positions of the lyrics sheets. Note that these manually-annotated phrase boundaries were used not only for training but also for testing. The V-A labels of the phrases of the audio signals were used for training a MLR regressor.

We then collected the social tags of the songs before the experiment. Social tags are frequently used in some pieces of research [23, 24] as they are tagged to songs by an unspecified large number of people based on the metadata of songs (for example mood, composer, genre) and tend to contain mood tags. We selected 106 tags from last.fm (`http://www.last.fm/`) that were tagged more than three times and found in the extended ANEW. Tags representing genre (e.g., rock, jazz) and musical instruments (e.g., guitar, piano) were excluded because they were considered irrelevant to moods. The average number of mood tags in a song was 5 and the maximum number of mood tags was 18.

Table 3: Mood trajectory features

Features	Description
Distance of two trajectories	Summation of Euclidean distance between each musical phrase of lyrics and an audio signal's mood trajectory.
Trajectory's variance	Variance of a mood trajectory.
Trajectory's mean	Mean of a mood trajectory.
Trajectory's transition variance	Variance of transition of a mood trajectory.

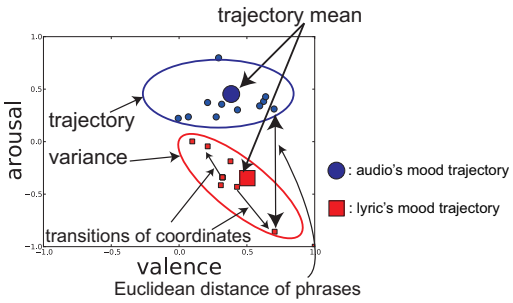

Figure 5: Mood trajectory features

5.1 Hierarchical Clustering

By performing hierarchical clustering using Ward's method [25], we clustered all of the songs. To confirm whether the songs were classified based on the complexity of time-varying musical mood, we compared variance of the acoustic features of the songs in each class. We used the *mood trajectory features* defined in Table 3 and Figure 5 for hierarchical clustering by using Ward's method.

5.2 Comparing Acoustic Features of Each Class

We compared the variance of the acoustic features of the songs in each class. Songs with complex time-varying moods (i.e., songs expected to have complex mood trajectories) were considered to have variable audio signals. Thereby, we were able to confirm the time-varying mood expressiveness of the mood trajectories by comparing the variance of the acoustic features of the songs in each class.

5.3 Comparing Social Tags and Mood Trajectories

We compared the mood tags selected from social tags and mood trajectories of the songs in each class. Since social tags are tagged by an unspecified large number of people as described above, mood tags taken from them can be considered as a ground-truth for mood recognition. Thus, we were able to confirm the accuracy of mood trajectories by comparing the V-A coordinates of the social tags and estimated mood trajectories.

5.4 Results and Discussion

5.4.1 Clustering Results

Figure 6 contains the result of hierarchical clustering. The horizontal line describes all songs, and the vertical line describes the distance of each cluster. Figure 6 is colored by a threshold set at 25% of the maximum distance. The figure is colored with six colors. Since songs colored with blue are less than half of the songs colored with other colors, we concluded that four classes were obtained by hierarchical clustering.

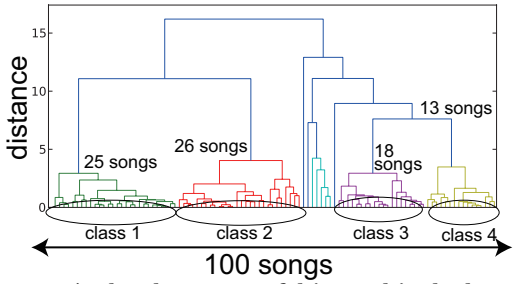

Figure 6: A dendrogram of hierarchical clustering according to the two mood trajectories

Table 4: Means of the acoustic feature variance in the songs of each class

	Principal components		
	First	Second	Third
Class 1	15.87	8.20	4.88
Class 2	**15.96**	**8.28**	**4.96**
Class 3	**15.79**	**8.13**	**4.77**
Class 4	15.91	8.19	4.83

5.4.2 Comparing Acoustic Features of Each Class

Table 4 is the means of the acoustic feature variance in the songs of each class. We compared the first, second, and third principal components. In Table 4, the principal component variances of class 2 were the largest. In contrast, the principal component variances of class 3 were the smallest. These results suggest that the songs of class 2 were classified as complex time-varying musical mood songs, and the songs of class 3 were classified as simple time-varying musical mood songs.

5.4.3 Comparing Mood Tags and Mood Trajectories

We calculated Euclidean distance on V-A space between

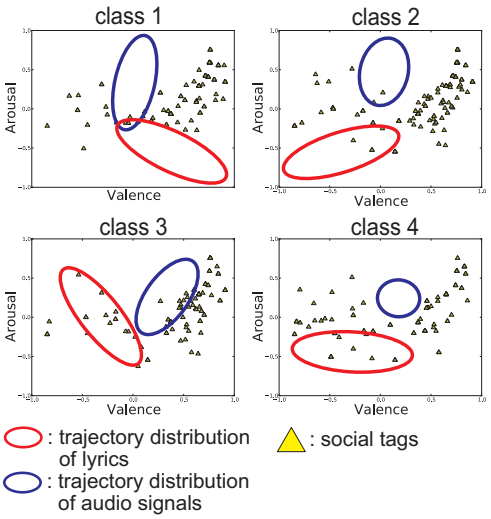

Figure 7: Distribution of mood trajectories in each class and V-A coordinates of mood tags

Table 5: Comparing results of mood tags and mood trajectories

	Class 1	Class 2	Class 3	Class 4	All songs
Matched songs	10%	13%	27%	19%	21%

each phrase of mood trajectories and mood tags of a song. The V-A coordinates of mood tags were calculated using the extended ANEW. If more than half of mood tags have the distance of 0.25 or less, we regarded that the song's mood trajectories matched the mood tags and, thus, the song's mood was correctly estimated. Table 5 shows percentages of songs that moods were correctly estimated. We can see that 21% of the songs are correctly estimated. This results, though there is still much room for improvement in it, is promising given the difficulty of the task we are tackling.

Figure 7 shows the distributions of mood trajectories in each class and the V-A coordinates of the mood tags. According to Table 5, it can be seen that class 3 yielded the best result among the four classes. This result is in line with Figure 7 where the distributions of mood tags and estimated mood trajectory of class 3 are the most similar to each other than those of the other classes.

6. CONCLUSION AND FUTURE WORK

We proposed a mood trajectory estimation method for lyrics and audios signals. Mood trajectories can represent relationship between lyrics' and audio signal's moods and time-varying musical moods. In an experimental evaluation, we clustered songs based on mood trajectories, compared the acoustic features of each class, and compared social tags and mood trajectories. Comparing acoustic features of each class showed that songs could be classified according to the complexity of time-varying moods. Moreover, comparing social tags and mood trajectories showed that our method could correctly estimate the mood trajectories for 21% of the songs. We should implement MIR systems based on mood trajectories and subjective experiments to confirm the validity of systems for future work.

7. ACKNOWLEDGEMENTS

This work is partially supported by Kakenhi (S) 19100003 and JST PRESTO.

8. REFERENCES

[1] J. Bergstra et al.: Scalable Genre and Tag Prediction with Spectral Covariance, ISMIR2010, pp.507–512, 2010.

[2] C. Yang: Music Database Retrieval Based on Spectral Similarity, ISMIR2001,pp.37–38, 2001.

[3] Y. E. Kim et al.: Music Emotion Recognition: A State of the Art Review, ISMIR2010, pp.255–266, 2010.

[4] Y. Hu et al.: Lyric-Based Song Emotion Detection With Affective Lexicon and Fuzzy Clustering Method, ISMIR2009, pp.123–128, 2009.

[5] M. van Zaanen et al.: Automatic Mood Classification Using TF*IDF Based On Lyrics, ISMIR2010, pp.75–80, 2010.

[6] J. Skowronek et al.: A Demonstrator for Automatic Music Mood Estimation, ISMIR2007, pp.345–346, 2007.

[7] T. Eerola et al.: Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models, ISMIR2009, pp.621–626, 2009.

[8] D. Yang et al.: Disambiguating music emotion using software agents, ISMIR2004, pp.52–58, 2004.

[9] C. Laurier et al.: Multimodal Music Mood Classification Using Audio and Lyrics, ICMLA 2010, pp.688–693, 2008.

[10] X. Hu et al.: Lyric Text Mining in Music Mood Classification, ISMIR2009, pp.411–416, 2009.

[11] E. M. Schmidt et al.: Prediction of Time-varying Musical Mood Distributions from Audio, ISMIR2010, pp.465–470, 2010.

[12] E. M. Schmidt et al.: Feature Selection for Content-based, Time-varying Musical Emotion Regression, ACM SIGMM MIR 2011, pp.267–274, 2010.

[13] J. A. Russell: A Ciucumplex Model of Affect, JPSP, Vol.39, No.6, pp.1161–1178, 1980.

[14] T. Hofmann: Probabilistic Latent Semantic Analysis, UAI99, pp.289–296, 1999.

[15] G. Xue et al.: Topic-bridged PLSA for Cross-domain Text Classification, SIGIR '08, pp.627–634, 2008.

[16] Y. Akita et al.: Language Model Adaptation based on PLSA of Topics and Speakers, ICSLP2004, pp.602–605, 2004.

[17] M. M. Bradley et al.: Affective Norms for English Words (ANEW): Instruction Manual and Affective Rating, Technical Report, C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[18] G. A. Miller: WordNet: A Lexical Database for English, CACM, Vol.38, Issue.11, pp.39–41, 1995.

[19] Y. E. Kim et al.: MoodSwings: A Collaborative Game for Music Mood Label Collection, ISMIR2008, pp.231–236, 2008.

[20] G. Tzanetakis et al.: MARSYAS: A Framework for Audio Analysis, Organised Sound, Vol.4, Issue.3, pp.169–175, 2000.

[21] B. Logan: Mel Frequency Cepstral Coefficients for Music Modeling, ISMIR2000, 11p., 2000.

[22] M. A. Bartsch et al.: To Catch a Chorus: Using chroma-based Representations far Audio Thumbnailing, WASPAA'01, pp.15–18, 2001.

[23] K. Bischoff et al.: How Do You Feel about 'Dancing Queen'?: Deriving Mood & Theme Annotations from User Tags, JCDL'09, pp.285–294, 2009.

[24] C. Laurier et al.: Music Mood Representations From Social Tags, ISMIR2009, pp.381–386, 2009.

[25] J. H. Ward: Hierarchical Grouping to Optimize an Objective Function, JASA, Vol.58, No.301, pp.236–244, 1963.