Similarity Measures for Chinese Pop Music Based on Low-Level Audio Signal Attributes

Chun-Man Mak; Tan Lee; Suman Senapati; Yu-Ting Yeung; Wang-Kong Lam
Department of Electronic Engineering
The Chinese University of Hong Kong
{cmmak, tanlee, ssuman, ytyeung, wklam}@ee.cuhk.edu.hk

ABSTRACT

In this article a method of computing similarity of two Chinese pop songs is presented. It is based on five attributes extracted from the audio signal. They include music instrument, singing voice style, singer gender, tempo, and degree of noisiness. We compare the computed similarity measures with similarity scores obtained with subjective listening by over 200 human subjects. The results show that rhythm and mood related attributes like tempo and degree of noisiness are most correlated to human perception of Chinese pop songs. Instrument and singing style are relatively less relevant. The results of subjective evaluation also indicate that the proposed method of similarity computation is fairly correlated with human perception.

1. INTRODUCTION

With the rapidly increasing popularity of digital music and related technologies, thousands of new songs are made available over the Internet everyday. The convenience of low-cost digital storage also promotes the increase in personal collection of music files. However, a user may find it more difficult to look for a song that he or she likes to listen. This calls for music recommendation systems to sort and find music efficiently.

A recommendation system aims to identify songs that match a user's taste and recommend these songs to the user. There are two types of music recommendation systems: Content-based and metadata-based recommendation. A content-based system mainly exploits audio features extracted from the music file itself to make recommendation [2], [6]. Metadata-based systems, like the *Last.fm* music network [11], make use of textual metadata associated with music documents, such as the artist's name, the song title, or the album name. These systems are often combined with collaborative filtering techniques to capture users' preferences. Lyrics of a song can also be used [1]. There are also hybrid systems that combine audio content and metadata for recommendation, [3],[4],[10].

In content-based recommendation system, the recommended songs are those that sound similar to the existing songs that the user likes to listen to. Recommendation becomes a task of matching in this case. A kind of similarity between a query song and a set of candidate songs needs to be computed. The candidates with the highest similarity measure will be returned as the results of recommendation. The computation of an effective similarity measure is therefore a crucial step in many recommendation systems.

Similarity of two songs can be computed from the contents or textual metadata embedded in the songs. In this article, we focus on studying content-based music similarity computed from songs in a specific genre, namely, Chinese pop songs. Most recommendation systems previously proposed were developed in a cross-genre environment [2],[10]. The song database contains a large number of songs of different genres, and genre classification is performed to put songs into the right group. Recommendation of songs within the same genre also has important applications. For example, a pop music producer may wish to promote his/her new productions by matching the taste of potential listeners. It is necessary to understand the aspects that humans consider when they decide if two songs are similar in intra-genre case. In addition, to the best of authors' knowledge, there does not exist any content-based recommendation system for Chinese pop songs. Our investigation results can provide valuable information in building such system in the future.

To facilitate the computation of a similarity measure, the important attributes of a song must be represented numerically. There are many attributes that can be used to represent a song, such as melody structure and chord. Such features, however, are difficult to extract accurately. Therefore, in the proposed method, we use a set of low-level audio descriptors, i.e., instrument identity, singing style, gender of the singer, tempo, and degree of noisiness (DoN) to represent the songs. The usefulness of these features is verified via subjective evaluation.

The paper is organized as follows. Section 2 describes the audio attributes that are used to represent Chinese pop songs, and the method of computing similarity from these attributes. Section 3 explains the process of subjective evaluation. Section 4 gives the experimental results, and the correlation between the proposed similarity measure and the findings of subjective evaluation. Conclusions are drawn in section 5.

2. MUSIC ATTRIBUTES

Prior to signal analysis, all audio files are down-sampled to 16 KHz. Most digital music are recorded at sampling rate above 16 KHz. By unifying to this common frequency, we try to minimize the discrepancies of signal quality among the many possible data sources. In the case of stereo recording, the two tracks are averaged to obtain a monaural wave signal. In short, all signal analysis operations are performed based on monaural wave signal at 16 kHz sampling rate.

2.1 Vocal / Non-vocal Segmentation

A typical Chinese pop song is about 3 to 4 minutes in duration and is composed of four parts. It starts with an instrument intro, followed by vocal verse, and instrument interlude, and finally the chorus. The vocal part is sung in the Cantonese dialect or Mandarin. Assuming this song structure, we divided each song into the vocal part (with human singing voice) and the non-vocal part (instruments only) in our analysis. The extraction of instrument attribute is done on the non-vocal part, while the attributes of singing style and gender are extracted from the vocal part. Tempo and DoN are estimated from the whole song. This is illustrated in Figure 1.

The wave signal is divided into short-time frames of 100ms long with frame shift of 50ms. Short-time Fourier Transform (STFT) of 2048 points is applied to each frame and 13 Mel-frequency cepstral coefficients (MFCC) are computed. A statistical classifier is built for vocal / non-vocal segmentation. About 500 songs are manually segmented into vocal and non-vocal parts. These songs, along with other songs used in our experiments, are all Chinese pop songs purchased from CD stores. These segments are used as the training data for the vocal / non-vocal classifier. The minimum segment length was set to be 1 second. A vocal class and a non-vocal class are modeled by two Gaussian mixture models (GMM), each with 64 mixtures.

The segmentation is performed by dividing the audio signals into non-overlapping segments of 1 second long, i.e., 20 frames, and classifying these segments as either vocal or non-vocal. For each segment, the log-likelihood with respect to the vocal class is computed as

$$L_{vocal} = \sum_{n=1}^{20} \log\left(Prob\left(\mathbf{\Theta}_n \middle| \mathbf{w}_{vocal}, \mathbf{\mu}_{vocal}, \mathbf{\sigma}_{vocal}\right)\right) \qquad (1)$$

where $\mathbf{\Theta}_n$ is the MFCC feature vector of the n^{th} frame in the segment, and $\mathbf{w}_{vocal}, \mathbf{\mu}_{vocal}, \mathbf{\sigma}_{vocal}$ are respectively the mixture weights, means, and covariance matrices of the vocal GMM. The log-likelihood with respect to the non-vocal class, denoted by $L_{non\text{-}vocal}$, is computed in the same way. The segment is classified to the class with the higher log-likelihood. This method of statistical classification is also used in the extraction of instrument, singing style, and gender attributes. The classification accuracy of the vocal / non-vocal classifier is 90%. The test data used in our experiments are not included in the training set and this applies to all other classification accuracy reported in this article. Readers may also notice that the numbers of mixtures used vary for different classifiers described in latter sections. These numbers are determined empirically to achieve the best trade-off between computation time and accuracy.

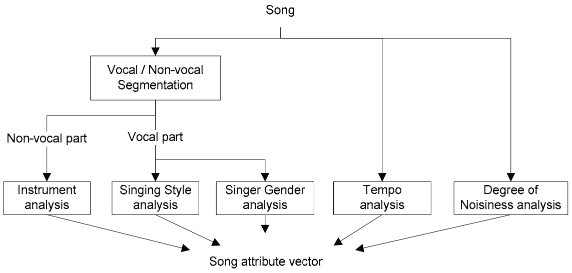

Figure 1. Segments of songs

2.2 Extraction of Music Attributes

Five attributes are used to describe a Chinese pop song: instrument, singing style, gender of singer, tempo, and degree of noisiness. Instrument, singing style, and gender are computed based on MFCC features and the statistical classification technique.

2.2.1 Instrument

For simplicity, we assume that only a single instrument is present or dominant at a particular time instant. Eight instruments commonly used in Chinese pop songs are modeled: reedy, electronic-guitar, piano, strings, synthesizer, guitar, flute, and percussion. Each instrument is represented by a GMM of 256 mixtures. The training data includes the 500 manually annotated songs and part of the RWC database [7]. Following an approach similar to vocal/non-vocal segmentation, each non-vocal segment in a song is assigned to an instrument class. Then the instrument attribute is represented by a vector with eight elements, i.e.,

$$\mathbf{F}_{inst} = \left\{I_1, I_2, \ldots, I_8\right\} \qquad (2)$$

where I_i is the percentage of segments in the song that are assigned to the i^{th} instrument class. In our experiments, the classification accuracy was 72%.

2.2.2 Singing Style

It is not trivial to describe the singing style in a Chinese pop song. In our study, singing style is defined with reference to a few famous singers of Chinese pop songs. For a given song, we try to measure the degree of similarity between the singing voice in the song and the voice of each reference singer. Among the 500 training songs, we choose 6 male and 6 female singers with distinct voices and styles. Each singer has about 30 training songs. A separate class is established for children's voice. As a result, we have a total of 13 singing style classes. Each class is represented by a GMM of 64 mixtures. The singing style attribute, $\mathbf{F}_{sing}$, is defined as

$$\mathbf{F}_{\text{sing}} \in \{S_1, S_2, ..., S_{13}\} \tag{3}$$

where S_i is the percentage of segments in the song that are assigned to the i^{th} class. In our experiments, the classification accuracy of singing style is 82%.

2.2.3 Gender

An explicit gender classifier is built. A male voice and a female voice model are trained to be 128–mixture GMM. The gender attribute, $\mathbf{F}_{gend}$, is defined as:

$$\mathbf{F}_{gend} \in \{G_{male}, G_{female}\} \tag{4}$$

where G_{male} and G_{female} are the percentage of male and female voice segments in the song, respectively. The song-level gender classifier has an accuracy of 97%.

2.2.4 Tempo

The tempo detection method proposed in [9] is used to generate the tempo information we need. The tempo of a given song is estimated by the Fourier analysis of the beat onset pattern. Complex domain spectral difference is used as the detection function for the beat onset. To find the spectral difference, we first estimate the instantaneous spectral change $\eta(m)$ at the m^{th} short-time frame, which is defined as

$$\eta(m) = \sum_k \left| \hat{Y}_m(k) - Y_m(k) \right| \tag{5}$$

where $Y_m(k)$ is the spectral value (DFT coefficient) at frame m and frequency bin k, and $\hat{Y}_m(k)$ is the corresponding value predicted from the immediately preceding frame, i.e.

$$\hat{Y}_m(k) = |Y_{m-1}(k)| e^{j\hat{\Phi}_m(k)}. \tag{6}$$

The expected phase $\hat{\Phi}_m(k)$ is predicted from phase in previous two frames as follows:

$$\hat{\Phi}_m(k) = \varphi_{m-1}(k) + (\varphi_{m-1}(k) - \varphi_{m-2}(k)) \tag{7}$$

where $\varphi_{m-1}(k)$ and $\varphi_{m-2}(k)$ are the observed phases for frame m-1 and m-2 respectively. Frame size of 100ms and frame shift of 10ms are used. The frame shift is much shorter than the 50ms shift as used in the extraction of other attributes because of the need to detect fast tempo. To obtain a beat onset pattern with clear and well-defined peaks, $\eta(m)$ is subtracted by the moving average threshold $\bar{\eta}(m)$ with window size $W = 10$, i.e.

$$\bar{\eta}(m) = \frac{1}{W} \sum_{i=m-\frac{W}{2}}^{m+\frac{W}{2}} \eta(i). \tag{8}$$

Half-wave rectification is then performed on the difference to obtain $\hat{\eta}(m)$, the beat onset value of frame m, i.e.

$$\hat{\eta}(m) = HWR(\eta(m) - \bar{\eta}(m)) \tag{9}$$

where $HWR(x) = (x + |x|)/2$.

To handle tempo variations over the entire duration of a song, we divide a song into segments of 12s long, with time shift of 4s. Each of these segments contains 1200 frames. Fourier analysis is performed on beat onset pattern of each segment, and the frequency axis is mapped into tempo values. The analysis result is represented by a tempogram, which is a two-dimensional time-tempo representation of the strength of tempo values in local segments. An example of tempogram is shown in Figure 2. The tempo value is limited to the range of 30 beat per minute (bpm) to 300 bpm, which covers most of the pop music. The tempo value with the strongest impulse is picked as the local tempo value for each segment. The local tempo values of all segments in the song form a distribution, from which the tempo attribute is derived as

$$\mathbf{F}_{tempo} = \{T_1, T_2, T_3\} \tag{10}$$

where T_1, T_2, and T_3 corresponds to the 25%, 50%, and 75% percentile tempo values.

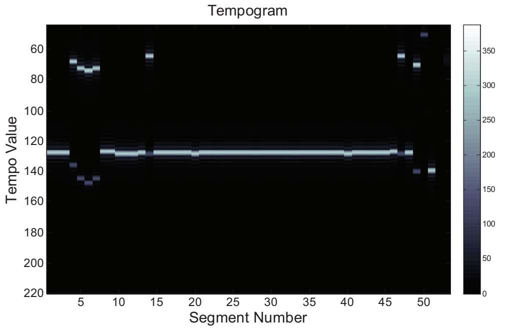

Figure 2. An illustration of the tempogram

2.2.5 Degree of Noisiness

The mood of a song is an important factor to be considered when computing similarity of songs. The energy of a song is in some way related to the mood of the song. According to Thayer's emotion model [8], songs with low intensity are usually associated with calm, relax, or depressed emotions, while songs with high intensity, are associated with exciting or angry emotion. This is illustrated in Figure 3. In addition, songs with thick texture, i.e., many instruments and voices playing simultaneously, are generally associated with more intense and excited feeling. On the other hand, songs with thin texture, i.e., a few instruments or voices, are commonly found in calm and relax music. Audio signals in thick-texture music are likely to have flatter spectra, compared to those in the thin-texture music. Thus we propose to use both the signal intensity and spectral flatness features to compute a "degree of noisiness" (DoN) attribute, which is related to the mood of the song.

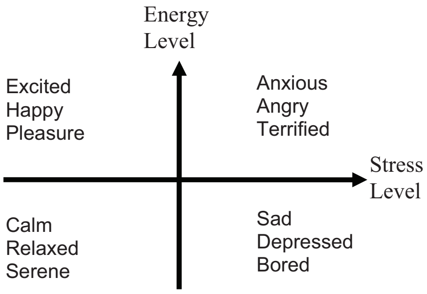

Figure 3. Thayer's model of mood [8]

Three values are defined as the DoN label: 0. 0.5, and 1, which correspond to "*low*", "*medium*", and "*high*" respectively. A low DoN means that the song is perceived as quiet and calm, while a high DoN means loud and noisy. We found that in many pop songs, the DoN level varies noticeably between the first and the second half of the song. Therefore, two DoN labels are used to describe a song, one for the first half of the song and one for the second half.

Prior to the DoN analysis, the audio signal is normalized by its maximum amplitude. Since the beginning and ending of the songs usually contains silence, the first 10% and the last 10% of the signal in each song are discarded. The remaining signal is then divided into two halves as described above. For each frame of 100 ms long, the signal intensity is computed by

$$P(m) = 10\log_{10}\left(\frac{1}{T}\sum_{t=0}^{T-1} X_m^2(t)\right) \tag{11}$$

where $P(m)$ is the power of m^{th} frame, X_m is the signal of m^{th} frame in time domain, and T is equal to 1600 for 16KHz sampling and 100ms frame size. From the frame-level signal intensity values, we compute the mean and standard deviation of the whole segment (half of the song), which are denoted by M_{power} and SD_{power}, respectively. Similarly, we compute the mean and standard deviation of the spectral flatness over each segment, which are denoted by M_{SF} and SD_{SF}, respectively. The spectral flatness, $SF(m)$, is defined by

$$SF(m) = \frac{\sqrt[K]{\prod_{k=1}^{K}|F_m(k)|}}{\frac{1}{K}\sum_{k=1}^{K}|F_m(k)|} \tag{12}$$

where $|F_m(k)|$ is the magnitude spectrum of X_m computed by 2048-point DFT, K is equal to 1023. The DC term (k=0) is ignored in the computation.

Each DoN class is represented by a single-mixture Gaussian model, which models the four-element feature vector $\{M_{power}, SD_{power}, M_{SF}, \text{and } SD_{SF}\}$. The DoN attribute vector of a song is defined as:

$$\mathbf{F}_{DoN} = \{N_1, N_2\} \tag{13}$$

where N_1 and N_2 corresponds to the label in the first and second half of the song, respectively.

2.3 Computation of Similarity Score

Let A and B denote two songs. The overall similarity between A and B is the weighted sum of the similarities computed for the five audio attributes. The similarity value of each attribute has the range of 0 (most dissimilar) to 1 (identical). The superscript in the attribute vectors and elements denotes the song from which the attribute is computed. For instrument and singing style attributes, the similarities $s_{inst}(A,B)$ and $s_{sing}(A,B)$ are computed by cosine similarity, i.e.

$$s_{inst}(A,B) = \frac{\mathbf{F}_{inst}^{A} \bullet \mathbf{F}_{inst}^{B}}{\left\|\mathbf{F}_{inst}^{A}\right\|\left\|\mathbf{F}_{inst}^{B}\right\|} \tag{14}$$

and

$$s_{sing}(A,B) = \frac{\mathbf{F}_{\text{sing}}^{A} \bullet \mathbf{F}_{\text{sing}}^{B}}{\left\|\mathbf{F}_{\text{sing}}^{A}\right\|\left\|\mathbf{F}_{\text{sing}}^{B}\right\|}. \tag{15}$$

For gender, the similarity, $s_{gend}(A,B)$, is defined as:

$$s_{gend}(A,B) = \frac{\min\left(G_{male}^{A}, G_{male}^{B}\right)}{\max\left(G_{male}^{A}, G_{male}^{B}\right)} \tag{16}$$

Tempo similarity, $s_{tempo}(A,B)$, is computed in a similar way as the gender, except that we take the average over the three components in the tempo attribute vector, i.e.,

$$s_{tempo}(A,B) = \frac{1}{3}\sum_{i=1}^{3}\frac{\min\left(T_i^A, T_i^B\right)}{\max\left(T_i^A, T_i^B\right)} \tag{17}$$

For the discrete class labels in the DoN attribute vectors, a normalized form of Euclidean distance is used for $s_{DoN}(A,B)$, which is defined as,

$$s_{DoN}(A,B) = 1 - \frac{\left\|\mathbf{F}_{DoN}^{A} - \mathbf{F}_{DoN}^{B}\right\|}{\sqrt{2}} \tag{18}$$

The overall similarity score is the weighted sum of the similarities by all attributes,

$$s(A,B)=\sum_{x\in music\ attribute} w_x s_x(A,B) \qquad (19)$$

where w_x denotes the weight for attribute x. The weights are determined empirically based on the observations from subjective evaluation.

3. SUBJECTIVE EVALUATION

Subjective evaluation is carried out to obtain a set of reference similarity scores that can be considered as the ground truth. These similarity scores will be used for two purposes: (a) to analyze which attributes are more important to human listeners when comparing two songs, and (b) to determine the optimal weights in the proposed objective similarity computation.

A total of 215 subjects participated in the listening tests. 43 Chinese pop songs, each with duration of about 4 minutes, were selected as the test materials. These songs were not included in the training set of 500 songs. They are chosen in a way that within this limited number of test songs, the varieties of instruments, singers, tempo, and mood are maximized. For the test song Q, 10 candidate songs were manually selected from the remaining songs in the test set so that we have various degrees of differences in the attributes between the candidates and the query songs. Five subjects were asked to listen to Q and the 10 candidate songs, denoted as C_i, i={1,2,…,10}. The subject was asked to rate the similarity between Q and C_i, on the scoring scale as shown in Figure 4.

A test session was divided into two parts, in the first part, a subject was asked to first listen to song Q, and 5 of the candidate songs. The subject was allowed to repetitively listen to Q if he/she liked to. There was a short break after the first part. Then the second part starts by listening to Q again, and then the remaining 5 candidates. The orders of playing the candidate songs are different from one listener to another.

4. EXPERIMENTAL RESULTS

We first investigated on the importance of each music attribute in the proposed similarity measure. This is done by assigning the weight of one attribute to be 1 and the others to be zero. For the query song Q in the set of subjective test songs, we compute the objective similarity score of Q and each of the 10 candidates using (19) and obtain a similarity vector $\mathbf{S}_Q$ = {$s(Q,C_1)$, $s(Q,C_2)$, … , $s(Q,C_{10})$}. From the results of subjective evaluation, we compute the average subjective similarity scores from all subjects that were tested with this set of songs. The resulted scores form the subjective similarity vector denoted as $\mathbf{E}_Q$={$e(Q,C_1)$, $e(Q,C_2)$, … , $e(Q,C_{10})$}. The Spearman's rank correlation between $\mathbf{S}_Q$ and $\mathbf{E}_Q$, denoted by ρ_Q, is computed for each query song. Subsequently, the overall average of correlation is computed over the 43 songs. The results are shown in Table 1. It is observed that the importance of the attributes in human listening is highly uneven. Instrument and gender are the least relevant attributes. The correlation values are close to zero, indicating that subjective scores are almost uncorrelated to the objective similarity. Singing style and tempo are more important, with a small but positive correlation. DoN is the most important attribute with the highest correlation between objective similarity and subjective scores.

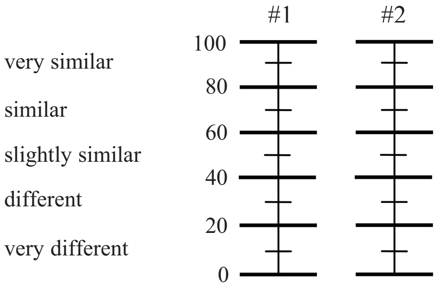

Figure 4. Scoring scale

Attribute	Average Correlation
Instrument	0.07
Singing style	0.20
Gender	0.06
Tempo	0.24
Degree of Noisiness	0.49

Table 1. Correlation of each attribute to subjective scores

Next we try to determine a set of optimal weights for similarity computation in (19) by maximizing the correlation to subjective listening. Exhaustive search is performed. The attribute weights are varied from 0 to 10 with an increment step size of 1. It was found that the optimal set of weights are 1, 2, 0, 3, and 4 for instrument, singing style, gender, tempo, and DoN, respectively. The average correlation achieved with this set of weights is equal to 0.544. Although it is not a very high correlation, the result shows that these audio attributes are still useful in modeling subjective similarity judgment of Chinese pop songs. The weights are in agreement with our observations on the study of single attribute.

Figure 5 shows the distribution of the correlation values for all test songs when the optimal weights are applied. Among the 43 songs, 5 songs have negative values of correlation, and 27 have correlations higher than 0.6. This indicates that the proposed objective similarity measure can fairly model the subjective similarity for most of the songs.

As part of the subjective test, each human subject was asked to fill in a survey questionnaire. The subject had to rank from 1 (most important) to 5 (least important) the following attributes when he/she considers the similarity of a pair of songs: rhythm, accompaniment instrument, singing style, lyrics, and languages. The average ranks of these attributes are tabulated in Table 2. The results match with what we found from the study of correlation between objective and subjective scores. Tempo and DoN, which somehow affect the perception of the rhythm, are the most important factors. Instrument and singing style are more or less the same in importance. Lyrics and languages (Cantonese or Mandarin) are the least important factors. At the current stage, we only use attributes like tempo and DoN to model the rhythm. More complex attributes such as melody and chord information should give us a better model for the rhythm and obtain an objective similarity metric that better correlates with subjective scores. One interesting note is that some subject suggests that attributes like atmosphere of the song, style of songs in different generations, and harmony of the songs may also be important. However these attributes are more abstract and difficult to extract from low-level audio features. Metadata such as publication year, genre, etc. may be used in future system to better describe these attributes.

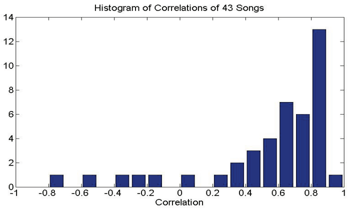

Figure 5. Histogram of correlations of 43 songs

Attributes	Average rank
Rhythm	1.3
Instrument	2.7
Singing Style	2.9
Lyrics	4.1
Language	4.4

Table 2. Average rank of attributes from survey

5. CONCLUSIONS

In this article we proposed five audio signal attributes that can be used to generate an attribute vector for a song in the Chinese pop song genre. The attribute vectors can then be used to compute similarity between songs, which is a fundamental process in content-based music recommendation system. We found that among these attributes, tempo and degree of noisiness play the most important role in approximating the subjective scores, followed by singing style and instrument. The results also indicate that rhythmic and mood information is crucial in objective similarity computation.

6. ACKNOWLEDGEMENTS

This work was supported by the Innovation and Technology Commission of the HKSAR Government.

7. REFERENCES

[1] Y. Xia, L. Wang, and K.-F. Wong: “Sentiment Vector Space Model for Lyric-Based Song Sentiment Classification,” International Journal of Computer Processing Of Languages, Vol. 21, No. 4, pp. 309-330, Dec. 2008.

[2] X. Zhu, Y.-Y. Shi, H.-G. Kim, K.-W. Eom: "An integrated music recommendation system," IEEE Transactions on Consumer Electronics,Vol.52, No.3, pp.917-925, Aug. 2006.

[3] T. Yoon, S. Lee, K.H. Yoon, D. Kim, J.-H. Lee: "A personalized music recommendation system with a time-weighted clustering," Proceedings of the 4th International IEEE Conference Intelligent Systems, Vol.2, pp.10-48-10-52, 6-8 Sept. 2008.

[4] K. Yoshii, M. Goto, K. Komatani, T. Ogata, H.G. Okuno: "An Efficient Hybrid Music Recommender System Using an Incrementally Trainable Probabilistic Generative Model," IEEE Transactions on Audio, Speech, and Language Processing, Vol.16, No.2, pp.435-447, Feb. 2008.

[5] W. Cohen and W. Fan: “Web-collaborative filtering: Recommending music by crawling the web,” Computer Networks, Vol. 33, No. 1–6, pp. 685–698, 2000.

[6] JJ Aucouturier, F Pachet, "Music similarity measures: What's the use," Proceedings of the ISMIR, 2002.

[7] RWC database: available at: http://staff.aist.go.jp/m.goto/RWC-MDB/

[8] R.E. Thayer: *The Biopsychology of Mood and Arousal*, Oxford University Press, 1989.

[9] P. Grosche and M. Muller: “A Mid-level Representation for Capturing Dominant Tempo and Pulse Information in Music Recordings,” Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 189-194, 2009.

[10] A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman: "A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures," Computer Music Journal, Vol. 28, No. 2, pp. 63-76, 2004.

[11] *Last.fm* music radio website: http://www.last.fm/home