EMOTION-BASED MUSIC RETRIEVAL USING CONSISTENCY PRINCIPLE AND MULTI-QUERY METHOD

Song-Yi Shin, Joonwhoan Lee
Department of Computer Engineering, Chonbuk National University, Jeonju, Chonbuk, South Korea

Kyoung-bae Eum
Department of Computer Eng., Kunsan National University, Kunsan, Chonbuk, South Korea

Eun-Jong Park
Electronics and Telecommunications Research Institute, Daejeon, South Korea

Keywords: Features of MPEG-7, Emotion-based music retrieval, Multi-query method, Consistency principle.

Abstract: In this paper, we propose the construction of multi - queries and consistency principal for emotion-based music retrieval. Existing content-based music retrievals which use a single query reflect the retrieval intention of user by moving the query point or updating the weights. However, these methods have the limitation to represent the complicated factors of emotion. In the proposed method, additional queries of music are taken in each feedback process to express the user's emotion. We classify the music by the emotions. And the music is clustered by the MKBC(Mercer Kernel-Based Clustering) method. After that, the inclusion degree of each descriptor is obtained. This means the weight that represents the importance of each descriptor for each emotion in order to reduce the computation. We got the excellent result within the 2nd retrieval through the feedback. In the feedback process, we used the consistency principle and multi- queries.

1 INTRODUCTION

Much audio data are supplied by diverse electronic devices in modern society. It is hard to find the music the user want to listen because much audio data are supplied. Content-based retrieval system has researched to solve this problem. In content-based retrieval method, the physical features of information corresponding to query are determined in advance and extracted. The user can fast and easily retrieve when comparing with the retrieval method by the conventional web-robot. This method is convenient because the supervisor only select and extract the physical features.

The emotion-based retrieval system has researche d in accordance with increasing the need on one's own information. It is retrieved by the emotion which is felt in the music. It is more advance method than the content-based retrieval. Recent approaches in music information retrieval have studied the relationship between low-level perceptual/acoustic features and semantic emotional/affective concepts such as mood (Lu et al., 2006; Leman, 2007; Leman et al., 2004). Lu et al. (Lu et al., 2006) has presented a mood detection system, where a GMM(Gaussian Mixture Model) was adopted to classify the mood of music clips into four cluster.

In this system, intensity, timbre, and rhythm were extracted from acoustic music data. Leman et al. (Leman et al., 2004) presented some low level acoustic features such as loudness, centroid, and interonset interval, and studied their correlation with three factors of emotion expressions. But, the conventional emotion-based music retrieval has only classified the mood and used the MIDI(Musical Instrument Digital Interf ace), MFCC(Mel-frequency cepstral coefficients), an d LPC(Linear prediction coding) data, which is diffic ult for standardization.

Shin S., Lee J., Eum K. and Park E.
EMOTION-BASED MUSIC RETRIEVAL USING CONSISTENCY PRINCIPLE AND MULTI-QUERY METHOD.
DOI: 10.5220/0002802802070213
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

Recently, the timbre data of MPEG-7 has been used. However, it is also difficult for standardization because it is mixed with MFCC and LPC. In this paper, we use only audio features supported in MPEG-7 to solve the problem.

We use the optimal features for each emotions to reduce the computation for retrieval process. The consistency is maintained during the feedback. If the music is discriminated as relevant by user, it should be retrieved constantly. Otherwise, it should not be retrieved any more. The performance of retrieval is improved through the feedback using consistency principle and multi-queries.

2 MANUSCRIPT PREPARATION

2.1 Emotion Model of Thayer

In our method, we use the emotion based on the model of Thayer (Thayer, 1989). According to the emotion model, the emotion can be divided by the two elements of “Arousal” and “Valence”. Arousal means the strength of emotion. We can feel triumphant when its value becomes large. We can feel calm when its value becomes small. Valence means the degree of positive/negative of emotion. It means the negative emotion when becoming small by (-).

Fig. 1 shows that the emotions are classified by the (+) degree and (-) degree of “Arousal” and “Valence”.

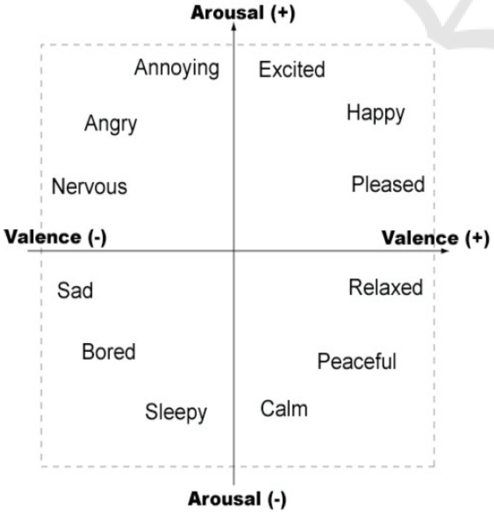

Figure 1: Emotion Based of Thayer.

In this paper, we use the emotional adjectives such as “Annoying”, "Angry", "Nervous", "Sad", "Bored", "Sleepy", "Calm", "Peaceful", "Relaxed", "Pleased", "Happy", and "Excited". Those emotional adjectives are comprised of six-pairs having opposite emotions.

2.2 Low level Descriptor of MPEG-7 Audio

The low level descriptor of MPEG-7 Audio used in the paper can express various features of audio signal (I. T. M. C. D. I. Part 4). It expresses the timbre of music and musical instrument. It is composed by the eight groups. They include the 18 temporal and spectral descriptors.

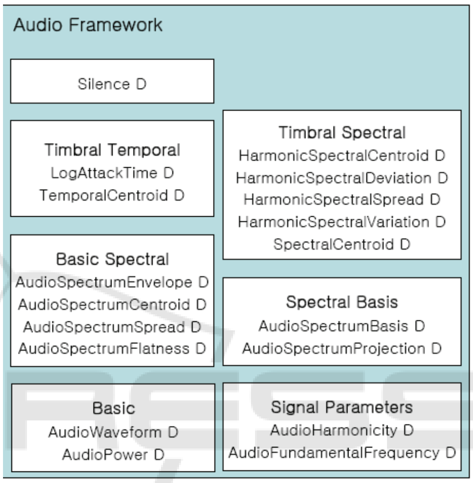

Figure 2: Low level descriptor of MPEG-7 Audio.

Fig. 2 shows the framework of low level descriptor of MPEG-7 Audio. The “D” means the descriptor (Overview of the MPEG-7 Standard). The 17 features are as follows.

AudioSpectrumBasis Type: The *AudioSpectrumBasis D* contains basis functions that are used to project high-dimensional spectrum descriptions into a low-dimensional representation

AudioSpectrumCentroid Type: The *AudioSpectrumCentroid D* describes the center of gravity of the log-frequency power spectrum. The *SpectrumCentroid* is defined as the power weighted log-frequency centroid.

AudioSpectrumEnvelopeType: The *AudioSpectrumEnvelope D* describes the spectrum of the audio according to a logarithmic frequency scale.

AudioSpectrumFlatnessType: The *AudioSpectrumFlatness D* describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands.

AudioHarmonicityType: The *AudioHarmonicity D* describes the degree of harmonicity of an audio signal.

AudioSignatureType: A structure containing a condensed representation as a unique content identifier for an audio signal for the purpose of robust automatic identification of audio signals (contains statistical summarization of data of *AudioSpectrumFlatnessType*)

AudioWaveformType: Description of the waveform of the audio signal.

AudioFundamentalFrequencyType: The *AudioFundamentalFrequency D* describes the fundamental frequency of the audio signal.

DcOffsetType: The *DcOffset* is each channel of maximum *AudioSegment* relative to the average technology.

InstrumentTimberType: The I*nstrumentTimbreType* is a set of Timbre Descriptors established in order to describe the timbre perception among sounds belonging simultaneously to the Harmonic and Percussive sound families

HarmonicSpectralCentroidType: The *HarmonicSpectralCentroid* is computed as the average over the sound segment duration of the instantaneous *HarmonicSpectralCentroid* within a running window. The instantaneous *HarmonicSpectralCentroid* is computed as the amplitude (linear scale) weighted mean of the harmonic peaks of the spectrum.

HarmonicSpectralDeviationType
The *HarmonicSpectralDeviation* is computed as the average over the sound segment duration of the instantaneous *HarmonicSpectralDeviation* within a running window. The instantaneous *HarmonicSpectralDeviation* is computed as the spectral deviation of log-amplitude components from a global spectral envelope.

HarmonicSpectralSpreadType: The *HarmonicSpectralSpread* is computed as the average over the sound segment duration of the instantaneous *HarmonicSpectralSpread* within a running window. The instantaneous *HarmonicSpectralSpread* is computed as the amplitude weighted standard deviation of the harmonic peaks of the spectrum, normalized by the instantaneous *HarmonicSpectralCentroid*.

HarmonicSpectralVariationType: The *HarmonicSpectralVariation* is defined as the mean over the sound segment duration of the instantaneous *HarmonicSpectralVariation*. The instantaneous *HarmonicSpectralVariation* is defined as the normalized correlation between the amplitude of the harmonic peaks of two adjacent frames.

LogAttackTimeType: The *LogAttackTime* is defined as the logarithm (decimal base) of the time duration between the times the signal starts to the time it reaches its stable part

SpectralCentroidType: The *SpectralCentroid* is computed as the power weighted average of the frequency of the bins in the power spectrum.

TemporalCentroidType: The *TemporalCentroid* is defined as the time averaged over the energy envelope.

In our method, we selected several descriptors which express the emotion well among these 18 low-level descriptors. We got the retrieval result for each emotion by the 1 to 5 descriptors. In order to reduce the computation time, we determined the number of descriptors showing best retrieval result for each emotion.

2.3 Feedback using Consistency Principle and Multi-queries Method

Most EBMR/CBMR(Emotion Based Music Retrieval / Content Based Music Retrieval) systems use only physical features that can be expressed as vectors (Wold et al., 1996; Foote, 1997). These methods make it easy to update weights or to construct query points. However, these methods are difficult to expect the consistent retrieval results because the complex emotion is expressed by the vector feature space extracted from the simple query music.

In this paper, we suggest the principle of consistency for relevance feedback. The emotion query of the user is made by constructing multiple-queries in every feedback process.

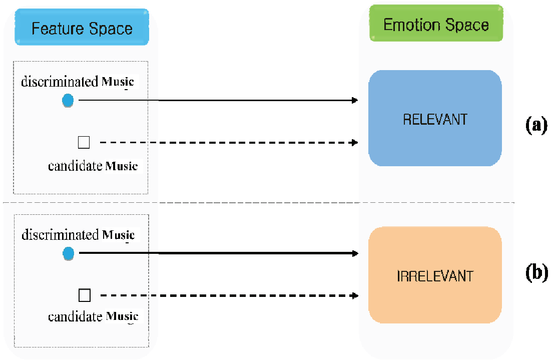

Figure 3: Principle of consistency.

Fig.3 shows the principle of consistency. If the music discriminated as suitable by the user and candidate music are similar in feature space as Fig. 3(a), those should be similar in emotion space. If the music discriminated as unsuitable by the user and candidate music is similar in the feature space as Fig. 3(b), the candidate music should be distinguished as the irrelevant in emotion space to maintain consistency.

During the feedback process, one more important principle to maintain this consistency is as follows. If the music is discriminated as relevant by user, it should be retrieved continuously. Otherwise, it should not be retrieved any more.

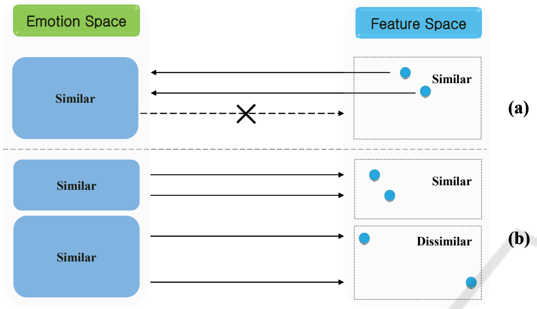

Figure 4: Necessity of multiple queries.

As Fig. 4(a), the music that is similar in the feature space must have the similar emotion. But, the inverse is not true. As Fig. 4(b), there is much music having the similar emotion even though they are different in the feature space. Therefore, we can say that the emotion of human can't be explained by just limited number of music features. If we use only the single query, we can not sufficiently express the complex emotion of human.

In this paper, we regard relevant or irrelevant music evaluated by user as the new query during the feedback. We construct multiple-queries to express the user's emotion by maintaining the principle of consistency.

We calculated the similarities for all music in DB by using the equation (1) and found all music that satisfies the condition $Sim^i_R \geq Sim^i_{IR}$. Finally, we got the N-result music by ascending order.

$$\begin{aligned} Sim^i_R &= 1.0 \quad if\ i \in R \\ & else\ \frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{DN} S_c^{i,k}\cdot W^c \\ Sim^i_{IR} &= 1.0 \quad if\ i \in IR \\ & else\ \frac{1}{m}\sum_{k=1}^{m}\sum_{c=1}^{DN} S_c^{i,k}\cdot W^c \end{aligned} \tag{1}$$

where n = Number of relevant Music
m = Number of irrelevant Music
DN = Number of used MPEG − 7 Descriptor s
$S_c^{i,k}$ = Similarity of between i and k Music
w .r.t c descriptor
W^c = Weight of descriptor c

The user can always retrieve the candidate music that closes to relevant query music and remote to irrelevant query music. The total retrieval procedure is as follows.

```
Select representative query Music of
each adjective queries.
Select features of music for each
adjective query.
k=1, retrieval_rate=0,
total_retrieval_num=10.
Repeat
   k-th retrieval;
    Evaluate the music by user;
    Reconstruct multiple queries for
    relevant/irrelevant music by (1);
    Calculate retrieval_rate;
    k = k + 1;
Until retrieval_rate >= 0.9.
```

2.4 Weights Decision using Inclusion Degree of Audio Descriptors

The W^C of equation (1) is a parameter. It decides which physical feature provides more influence on the emotion in any emotion query music. We get the weights of the physical features. After that, we selected necessary features by these weights of the physical features. So, we can prevent not only computation time but also complicate retrieval system.

In this paper, for adjective pairs respectively, we calculated the inclusion degree for the physical feature of MPEG-7. We used MKBC(Mercer kernel-based clustering) algorithm as the algorithm for clustering (Cirolami, 2002; Park, 2008). There are several MPEG-7 audio features. The number of properties is also different. So, this similarity measure will be fallen to the curse of the dimension. The MKBC is kernel-based clustering. Therefore, the MKBC is the best algorithm-m that does not fall into the curse.

$$\begin{aligned} & R = \{HC_1^{adj}, HC_2^{adj}\} \\ & v_k^i = \frac{card(X^k \in HC_i^{adj})}{card(X^k)}, i = 1,2 \end{aligned} \tag{2}$$

HC_i^{adj} means the i-th cluster information classified in a view of emotion adj by human. X^k indicates the cluster information classified by the system using the feature ofD^j.

v_k^i is a standard representing how it explains the classified cluster information. If we assume that cluster information classified by human is an ideal result, we can calculate the inclusion degree for any physical feature D^j.

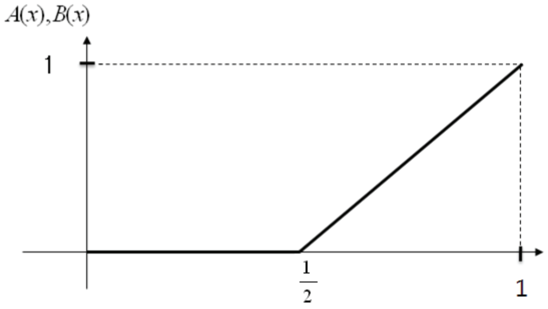

Figure 5: Membership function of V_k^i.

In equation (3), x represents v_k^i

$$A(x) = degree\ \ X^k\ belongs\ \ to\ HC_1^{adj}$$

$$B(x) = degree\ \ X^k\ belongs\ \ to\ HC_2^{adj}$$

$$A(x),B(x) = \begin{cases} 0 & when \quad x < \frac{1}{2} \\ 2(x-0.5) & when \quad \frac{1}{2} \le x \le 1 \end{cases} \tag{3}$$

The equation (4) means how much correspond between the cluster classified by human and the cluster classified by the system.

$$M_c = s_c \frac{\sum_{k=1}^{c} (v_k^1 + v_k^2)}{c}, where \tag{4}$$

$$S_c = \begin{cases} -0.1c + 1.2, when\ 2 \le c \le 10 \\ \quad 0 \qquad\quad, when\ c \ge 10 \end{cases}$$

$$ID^j = \sum_{c=2}^{5} M_c \tag{5}$$

$$\omega^j = \frac{ID^j}{\sum_{all\ j} ID^j}$$

3 EXPERIMENT

Fig 6 shows the overall flow of the proposed emotion-based music retrieval system. We used the 360 songs in the experiment. They were not classified by genre and classified by the emotion using Thayer model. The music has the emotion such as “Annoying”, "Angry", "Nervous", "Sad", "Bored", "Sleepy", "Calm", "Peaceful", "Relaxed", "Pleased", "Happy", and "Excited". A song has one emotion and it can have over one emotion. Therefore, the duplicate retrieval is possible by another emotion. We used the temporal and spectral descriptors.

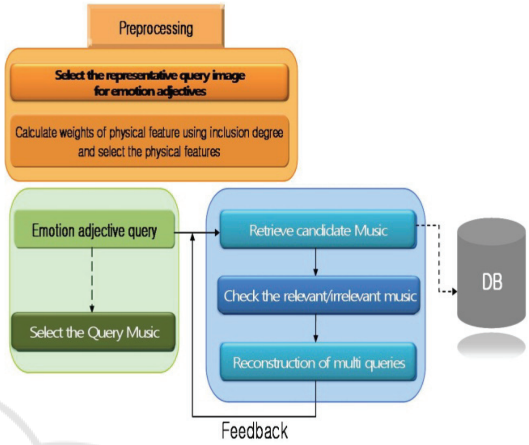
Preprocessing
Select the representative query image for emotion adjectives
Calculate weights of physical feature using inclusion degree and select the physical features
Emotion adjective query
Select the Query Music
Retrieve candidate Music
Check the relevant/irrelevant music
Reconstruction of multi queries
DB
Feedback

Figure 6: Proposed emotion-based music retrieval system.

In this paper, we used the 17 descriptors with the exception of “Silence D”. The Euclidean similarity is used as the similarity for descriptors, because the audio features supported by MPEG-7 are provided as a vector. Therefore, we used the Cosine or Euclidean similarity among the vector similarity methods. Table 1 shows the result of experiment. In this paper, the performance evaluation was done by the surveys of three students. As a result, the Euclidean similarity is better than Cosine similarity.

Table 1: The result obtained by each similarity (%).

	Cosine	Euclidean
Angry	21.4	75.7
Calm	22.9	62.9
Excited	18.6	88.6
Sad	64.3	25.7

The features for retrieval are used by varying the number and importance degree for each emotion adjective. Table 2 shows the weight decided by MKBC.

Representative music used in the 1st retrieval is selected through the previous survey. The number of representative music is five. The user select whether the music is relevant or irrelevant. The relevant music is the training set and the test sets are the 360 songs stored in database. In the 1st retrieval method, we can search by using the similarity of these features.

After that, the similarity value is sorted by the descending order and the 10 music are found. From the 2nd retrieval, we used the feedback using consistency principle and multi-queries method. The candidate music is searched by the ascending order. If the music is satisfied by the $Sim^i{}_R > Sim^i{}_{IR}$.

Table 2: The weight of the features.

Adjective	Descriptor and weights	
Angry	AudioFundamentalFrequency	1.000
Annoying	AudioBasis, AudioFlatness, AudioSignature AudioWave, AudioEnvelope, AudioCentroid	0.884 ,0.876, 0.807, 0.639, 0.618, 0.584
Bored	SpectralCentroid	0.813
Calm	AudioBasis, AudioFlatness, AudioSignature, AudioWave, AudioEnvelope, AudioCentroid, InstrumentTimbre, HarmonicSpectralSpread, Dcoffset, HarmonicSpectralCentroid, SpectralCentroid	0.884, 0.876, 0.807, 0.639, 0.618, 0.584, 0.484, 0.481, 0.477, 0.450, 0.442
Excited	AudioEnvelope	0.959
Happy	SpectralCentroid, HarmonicSpectralCentroid	0.813, 0.789
Nervous	HarmonicSpectralSpread	1.000
Peaceful	AudioFundamentalFrequency	1.000
Pleased	HarmonicSpectralCentroid, SpectralCentroid	0.860, 0.857
Relaxed	HarmonicSpectralSpread, AudioFlatness, Dcoffset, HarmonicSpectralCentroid, TemporalCentroid, AudioSignature, AudioEnvelope, HarmonicSpectralDeviation, AudioBasis, AudioFF	1.000, 0.875, 0.872, 0.856, 0.823, 0.816, 0.747, 0.728, 0.663, 0.659
Sad	HarmonicSpectralCentroid, SpectralCentroid, AudioBasis	0.860, 0.857, 0.663
Sleepy	AudioEnvelope, AudioFlatness, AudioFF	0.959, 0.951, 0.912

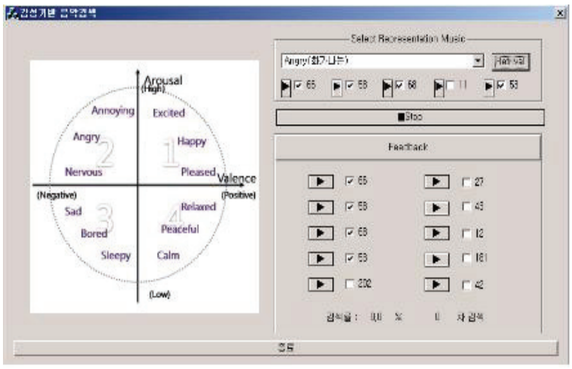

Figure 7: Emotion based Music retrieval system.

Fig 7 shows the emotion based music retrieval system about "Angry". The one of retrieval music is checked like "√" by the user in case of relevant.

Fig 8 is the survey of 12 students and shows the result that retrieve until 90% when searching the 10 songs by the feedback using consistency principle and multi-queries. The 90% retrieval results are satisfied within 2.5 times as Fig. 8. We did not listen to these kinds of music in case of "Angry" or "Nervous". It is hard to define these kinds of music. So, these emotions have the high number of times for retrieval than other emotions. For these reasons, there is the highest number of times for retrieval in case of searching the "Nervous". There are the lowest number of times for retrieval in case of searching the "Peaceful".

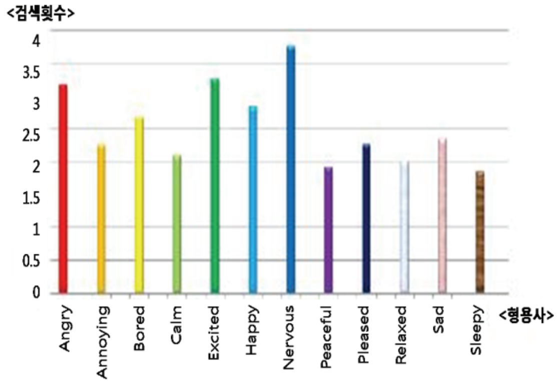

Figure 8: Until the 90% average number.

4 CONCLUSIONS

The existing researches could distinguish only genres. The data-types such as MIDI, MFCC, LPC, and the MPEG-7 timbre data with them are difficult to standardize. In this paper, only MPEG-7 descriptors are used to solve this problem. The emotion based music retrieval can be performed in the proposed method. We classify the music by the emotions. And the music is clustered by the MKBC method. After that, the inclusion degrees of the features are obtained by comparing above two results. These mean the weight that represents the importance of each descriptor for each emotion in order to reduce the computation. These are used to select the optimal descriptors for retrieval by each adjective.

We got the excellent result within the 2nd retrieval through the feedback using consistency principle and multi-queries method.

In this paper, we used the 12 emotions based on Thayer model. The emotion adjectives used in music are the adjectives such as “Rainy day”, “Exciting day” besides Thayer model. In the future, we will retrieve by using the diverse emotion adjectives besides 12 adjectives and study on the method for improving the 1st retrieval result.

REFERENCES

En-jong Park. (2008). Emotion Based Image Retrieval Using Multi-Query Images and Consistency Principle, A doctor degree dissertation of Chonbuk National University.

Information Technology Multimedia Content Description Interface Part 4: Audio, ISO/IEC FDIS 15938-4.

Jonathan T. Foote. (1997). Content-Based Retrieval of Music and Audio, *Multimedia Storage and Archiving Systems* II, Proceedings of SPIE, Vo. 3229, pages 138-147.

L. Lu, D. Liu, and H. J. Zhang. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio Speech Language Process*, vol. 14, no.1.

Mark Cirolami. (2002). Mercer kernel-Based Clustering in Feature Space, *IEEE TRANSACTIONS ON NEURAL NETWORKS*, Vol.13, No.3, pages 780-784.

M. Leman. (2007). Embodied Music Cognition and Mediation Technology, Cambridge, MA; MIT Press. 2007

M. Leman, v. Vermulen, and M. Lesaffre. (2004). Correlation of gestural musical audio cues and perceived expressive qualities, *in Gesture Based Communication in Human Computer Interaction*, A. Camurri and G. Volpe, Eds. New York: Springer Verlag.

Overview of the MPEG-7 Standard (version 6.0), ISO/IEC JTC1/SC29/WG11/N4509.

R. E. Thayer. (1989). The Biopsychology of Mood and Arousal, New York, Oxford University Press.

Wold, E., Blum, T., Keislar, d., and Wheaton, J. (1996). Content-based Classification, Search, and Retrieval of Audio, *IEEE Multimedia*, Vol. 3, No. 3, pages 27-36.