Automatic Mood Detection from Acoustic Music Data[1]

Dan Liu
Department of Automation
Tsinghua University
Beijing 100084 China
aryaya00@mails.tsinghua.edu.cn

Lie Lu
Microsoft Research Asia
Sigma Center, Haidian District
Beijing 100080 China
llu@microsoft.com

Hong-Jiang Zhang
Microsoft Research Asia
Sigma Center, Haidian District
Beijing 100080 China
hjzhang@microsoft.com

Abstract

Music mood describes the inherent emotional meaning of a music clip. It is helpful in music understanding, music search and some music-related applications. In this paper, a hierarchical framework is presented to automate the task of mood detection from acoustic music data, by following some music psychological theories in western cultures. Three feature sets, intensity, timbre and rhythm, are extracted to represent the characteristics of a music clip. Moreover, a mood tracking approach is also presented for a whole piece of music. Experimental evaluations indicate that the proposed algorithms produce satisfactory results.

1 Introduction

As there are more and more music databases in personal computer and the Internet at present, people start to realize the importance of creating metadata that allow users to access musical works easily. Although traditional information such as the name of the artist or the title of the work remains important, these tags have limited applicability in many music-related queries. Nowadays, users expect more semantic metadata to archive music, such as similarity, style and mood (Huron, 2000). However, compared to the first two, few works have focused on mood detection.

One common opinion objecting to mood detection is that the emotional meaning of music is subjective and it depends on many factors including culture. Music psychologists now agree that culture is of great importance in people's mood response to music, as well as other factors including education and previous experiences. However, it is also found that, within a given cultural context, there is agreement among individuals as to the mood elicited by music (Radocy and Boyle, 1988). Krumhansl (Krumhansl, 2002) also pointed out that musical sounds might inherently have emotional meaning. For example, some music patterns represent contentment or relaxing, while some others make an individual feel anxious or frantic. Therefore, it is possible to build a mood detection system in a concrete environment, for example, for classical music in western culture.

Few works have touched this field. Liu (Liu, Zhang and Zhu, 2003) has presented a mood recognition system, where a fuzzy classifier was adopted to classify the mood of Johann Strauss's waltz centos into five clusters. In this system, tempo, loudness, pitch change, note density and timbre were extracted from MIDI file and used as the primitives to recognize the mood of music. Katayose (Katayose, Imai and Inokuchi, 1988) also presented a sentiment extraction system for pop music, where monophonic acoustic data was firstly transcribed into music codes. Then, primitives of music such as melody, rhythm, harmony and form were extracted from these music codes. These works have led to some impressive results, but they both concentrated on MIDI or symbolic representations, due to the difficulty of extracting useful features from acoustic data. However, most music in real world is not in symbolic form and there is no existing transcription system that can translate it into symbolic representations well (Scheirer, 2000). Therefore, it is necessary to deal with the acoustic data directly.

In this paper, we present a mood detection algorithm for classical music from acoustic data.

1.1 Mood Taxonomy

One issue of mood detection is on mood taxonomy. In music psychology, the traditional approach to describing mood response is using adjective descriptors, such as pathetic, hopeful and gloomy. However, these adjectives varied quite freely in different researches (Liu, Zhang and Zhu, 2003; Katayose, Imai and Inokuchi, 1988). There is not a standard mood taxonomy system accepted by all currently. Hevner's adjective checklist (Hevner, 1935) presented in 1930s has served as the basis for some subsequent research on mood response to music. This checklist is composed of 67 adjectives from eight clusters, which include Sober, Gloomy, Longing, Lyrical, Sprightly, Joyous, Restless and Robust. However, since adjectives in the same cluster are actually of approximately the same meaning, it is very difficult to discriminate one from others. This ambiguity makes it

[1] This work was performed in Microsoft Research Asia

difficult to obtain the "ground truth". Meanwhile, it doesn't indicate any underlying stimulus that influences these responses, which will be of great importance for computational modeling. In the late 1990s, Thayer (Thayer, 1989) proposed a two-dimensional mood model. Unlike Hevner's checklist that uses individual adjectives which collectively form a mood pattern, this dimensional approach adopts the theory that mood is entailed from two factors: Stress (happy/anxious) and Energy (calm/energetic), and divides music mood into four clusters: Contentment, Depression, Exuberance and Anxious/Frantic as shown in Fig. 1.

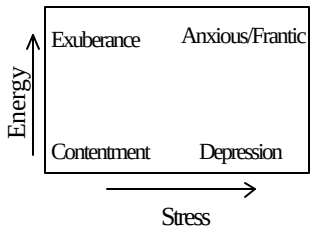

Figure 1: Thayer's model of mood

In Fig. 1, Contentment refers to happy and calm music, such as Bach's "Jesus, Joy of Man's Desiring"; Depression refers to calm and anxious music, such as the opening of Stravinsky's "Firebird"; Exuberance refers to happy and energetic music such as Rossini's "William Tell Overture"; and Anxious/Frantic refers to anxious and energetic music, such as Berg's "Lulu". Such definitions of the four clusters are explicit and discriminatable, and the two-dimensional structure also gives importance cues for computational modeling. Therefore, it is applied in our mood detection system.

1.2 Hierarchical Framework

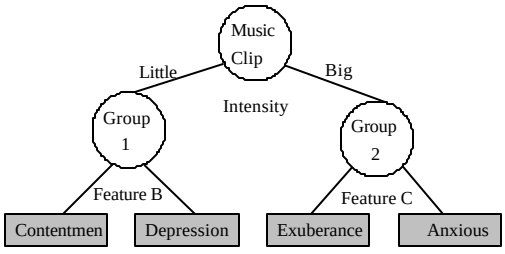

Figure 2: The hieratical mood detection framework

Based on Thayer's hierarchical model of mood (Thayer, 1989), a hierarchical framework is proposed for mood detection, as illustrated in Fig. 2. Huron (Huron, 1992) pointed out that of the two factors in Thayer's model of mood, Energy is more computationally tractable and can be estimated using simple amplitude-based measures. In fact, energy for Contentment and Depression is usually much less than that of Exuberance and Anxious. Therefore, features representing energy are firstly used to classify all these four mood clusters into two groups. If the energy is little, it will be classified into Group 1 (Contentment and Depression); otherwise, it is classified into Group 2 (Exuberance and Anxious). Then, other features are used to determine which exactly the mood type is. This framework is accordant to the music psychological theory. Meanwhile, since the performance of different features is not the same in discriminating different mood clusters pairs, this framework is advantaged in making it possible to use the most suitable features for different tasks. Moreover, like other hierarchical methods, it can make better use of sparse training data than its non-hierarchical counterparts (McCallum, 1998).

This paper is structured as follows. Section 2 describes the extraction of features. Detailed mood detection process is presented in Section 3. In Section 4, an automatic segmentation approach is presented for mood tracking in a whole piece of music. Section 5 deals with empirical experiments and performance evaluations of the proposed algorithms and Section 6 with conclusions and future directions.

2 Feature Extraction

It was indicated that mode, intensity, timbre and rhythm are of great significance in arousing different music moods (Hevner, 1935; Radocy, and Boyle, 1988; Krumhansl, 2002). For example, major keys are consistently associated with positive emotions, whereas minor ones are associated with negative emotions. However, mode is very difficult to obtain from acoustic data (Hinn, 1996). Therefore, only the rest three features are extracted and used in our mood detection system. Compared to the two dimensions in Thayer's model of mood, intensity is corresponding to "energy", while both timbre and rhythm are corresponding to "stress".

Each input music clip is first down-sampled into a uniform format: 16000Hz, 16 bits, mono channel, and divided into non-overlapping 32ms-long frames. In each frame, an octave-scale filter-bank is used to divide the frequency domain into several sub-bands:

$$[0,\frac{w_0}{2^n}),[\frac{w_0}{2^n},\frac{w_0}{2^{n-1}}),...[\frac{w_0}{2^2},\frac{w_0}{2^1}] \quad (1)$$

where w_0 refers to the sampling rate and n is the number of sub-band filters. In real implementation, 7 sub-bands are used. Then, timbre features and intensity features are extracted from each frame. Their means and variances are calculated across the music file and thus make up of timbre and intensity features sets. Meanwhile, rhythm features are also extracted directly from the music clip. In order to remove the relativity among these raw features, Karhunen-Loeve transform is performed on each feature set. After K-L transform, each of the three feature vectors is mapped into an orthogonal space, and each covariance matrix also becomes diagonal in the new feature space. This procedure helps to achieve a better classification performance with GMM classifier later. Detailed feature extractions are as follows.

2.1 Timbre Features

Many existing results show that the timbre of sound is determined primarily by the spectral information in different sub-bands (Zhang and Kuo, 1998). In this paper, both spectral shape features and spectral contrast features are used. The

Feature Name		Definition
Spectral Shape Features	Centroid	Mean of the short-time Fourier amplitude spectrum.
	Bandwidth	Amplitude weighted average of the differences between the spectral components and the centroid.
	Roll off	95^{th} percentile of the spectral distribution.
	Spectral Flux	2-Norm distance of the frame-to-frame spectral amplitude difference.
Spectral Contrast Features	Sub-band Peak	Average value in a small neighborhood around maximum amplitude values of spectral components in each sub-band.
	Sub-band Valley	Average value in a small neighborhood around minimum amplitude values of spectral components in each sub-band.
	Sub-band Average	Average amplitude of all the spectral components in each sub-band.

Table 1: Definition of Timbre Features

detail features used are listed in Table 1. Spectral shape features, which include centroid, bandwidth, roll off and spectral flux, are widely used to represent the characteristics of music signals (Tzanetakis and Cook, 2002). They are also important for mood detection. For example, centroid for music of Exuberance is usually higher everywhere than that of Depression, since Exuberance is generally associated with a high pitch whereas Depression is with a low pitch. Meanwhile, octave-based spectral contrast features are also used to represent relative spectral distributions, due to their good properties in music genre recognition (Jiang, Lu, Zhang, Tao and Cai, 2002).

2.2 Intensity Features

Intensity is approximated by the signal's root mean-square (RMS) level in decibels (Erling, 1996). It is essential for mood detection, because intensity in music of Contentment and Depression is usually little, but that of Exuberance and Anxious is usually big. In this system, intensity in each sub-band and the sum of them are used.

2.3 Rhythm Features

Three aspects of rhythm are closely related with people's mood response: strength, regularity and tempo. For example, in the Exuberance cluster, the rhythm is usually strong, steady and the tempo is fast; while in Depression, music is usually slow and with no distinct rhythm pattern. Therefore, these three features are extracted accordingly. Since drum or some bass instruments are the most important components to represent rhythm, and they show their properties mainly in the lower sub-bands, in our system, only the lowest sub-band is used to extract rhythm features.

After amplitude envelope is extracted from this sub-band by using a half hamming (raise cosine) window, a Canny estimator is used to estimate its difference curve, which is used to represent the rhythm information. The peaks above some threshold in such a rhythm curve are detected as bass instrumental onsets. Then, three features are extracted as follows:

- Average Strength: the average strength of bass instrumental onsets.
- Average Correlation Peak: the average of the maximum three peaks in the auto-correlation curve. The more regular the rhythm is, the higher the value is.
- Average Tempo: the common divisor of the peaks of the auto-correlation curve.

3 Mood Detection

Based on the three feature sets extracted in Section 2, the mood detection process is performed through a hierarchical framework, as illustrated in Fig. 3. Compared to its non-hierarchical counterpart as shown in Fig. 4, such hierarchical framework can stress on different features for different classification tasks, it can also make better use of sparse training data (McCallum, 1998).

In our system, Gaussian Mixture Model (GMM) is utilized to model each feature set. In constructing each GMM, the Expectation Maximization (EM) algorithm is used to estimate the parameters of the Gaussian component and mixture weights. The initialization is performed using the K-means algorithm.

For a given music clip X, it is firstly classified into Group 1 (Contentment and Depression) or Group 2 (Exuberance and Anxious) based on its intensity information; and then classification is performed in each group based on timbre and rhythm features, as the Fig. 3 illustrates.

To classify the music clip into differnet groups, simple Bayesian criteria is employed, as

$$\frac{P(G_1 \mid I)}{P(G_2 \mid I)} \begin{cases} \geq 1, & Select \quad G_1 \\ < 1, & Select \quad G_2 \end{cases} \tag{2}$$

where G_i represents different mood group, I represents the intensity feature set.

In each group, the probability of being an exact mood given timber and rhythm features can be calculated as

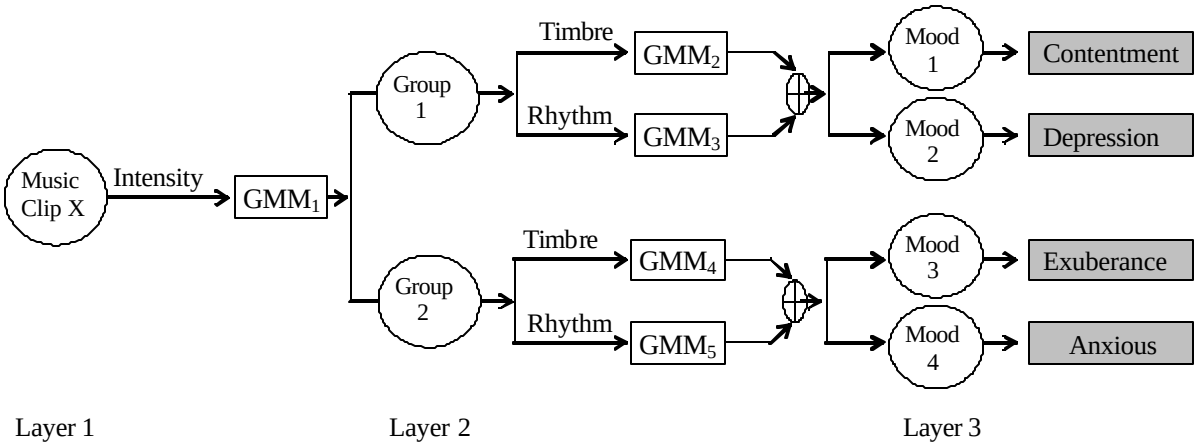

Figure 3: The hierarchical mood detection framework

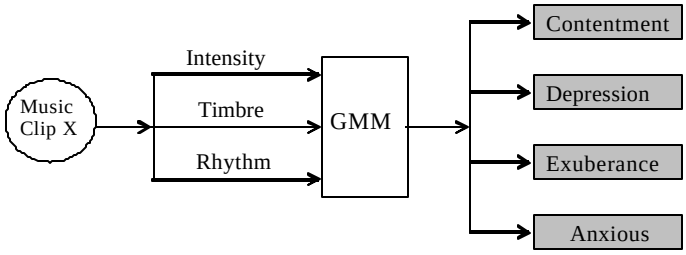

Figure 4: The non-hierarchical mood detection framework

$$P(M_j|G_1,T,R) = l_1 \times P(M_j|T) + (1-l_1) \times P(M_j|R) \qquad j = 1,2$$
$$P(M_j|G_2,T,R) = l_2 \times P(M_j|T) + (1-l_2) \times P(M_j|R) \qquad j = 3,4 \qquad (3)$$

where M_i is the mood cluster, T and R represent timbre and rhythm features respectively; l_1 and l_2 are two weighting factors to emphasize different features for the mood detection in different mood groups.

Actually, in the Group 1, the tempo of both mood clusters is usually slow and the rhythm pattern is generally not steady, while the timbre of Contentment is usually much brighter and more harmonic than that of Depression. Therefore, the timbre features are more important than the rhythm features in the classification in Group 1. On the contrary, in Group 2, rhythm features are more important. Exuberance usually has a more distinguished and steady rhythm than Anxious, while their timbre features are similar, since the instruments of both mood clusters are mainly brass. Based on these facts, l_1 is usually set as larger than 0.5, while l_2 is less than 0.5. Their detailed values are given in the experiments in Section 5.1. After each probability is obtained, Bayesian criterion, similar to Equation 2, is again employed to classify the music into exact mood cluster.

4 Mood Tracking

In the previous sections, we present an algorithm on mood detection in a given music clip, where the mood type is consistent. However, since the mood is usually changeable in a whole piece of classical music (Kamien, 1992), it is not appropriate to detect the mood in the range of the whole song. In fact, it is necessary to divide the music into several independent segments, each of which contains a constant mood, and then to detect the mood type in each segment respectively. In this way, mood is tracked in a whole piece of music.

Since changes in intensity and timbre are main cues for new sound event and are therefore important for segmentation (Tzanetakis and Cook, 1999), both of the features are used to complement each other to improve the performance of segmentation in this method.

According to music theory (Kamien, 1992), one paragraph is usually of 16 bars and a very fast tempo is about 1 bar/second in classical music. Therefore, we assume the minimum segment length is 16 seconds, and set the basic processing unit as 16 seconds window with 1-second temporal resolution.

To find the segment boundary, divergence shape (Campbell, 1997) is used to measure the dissimilarity between two contiguous windows, supposing both the features are Gaussian distributed,

$$D = \frac{1}{2} tr[(C_i - C_j)(C_j^{-1} - C_i^{-1})] \qquad (4)$$

where C_i and C_j is the estimated covariance matrix of ith and $(i+1)$th window, respectively.

Base on the dissimilarity measure, a confidence of being a boundary is defined as

$$Conf_I = \frac{1}{A_I}\exp(\frac{D_I - m_I}{s_I}), \quad Conf_T = \frac{1}{A_T}\exp(\frac{D_T - m_T}{s_T}) \quad (5)$$

where m_I and s_I are the mean and variance of intensity dissimilarity between two contiguous windows, m_T and s_T are the mean and variance of timbre dissimilarity between two contiguous windows, A_I and A_T are used for normalization. Thus, the total confidence is

$$Conf = a \times Conf_I + (1-a) \times Conf_T \quad (6)$$

where a is a weighting factor and we set $a = 0.5$ in real implementation.

A potential chance boundary is found between *i*th and (*i+1*)th window, if the following conditions are satisfied:

$$\begin{aligned} &1) \quad Conf(i, i+1) > Conf(i+1, i+2) \\ &2) \quad Conf(i, i+1) > Conf(i-1, i) \\ &3) \quad Conf(i, i+1) > Th_i \end{aligned} \quad (7)$$

where *Conf(i, j)* is the confidence that the segment boundary is at between *i*th and *j*th window, Th_i is a threshold. The first two conditions guarantee that a local peak exists, and the last condition can prevent very low peaks from being detected. The threshold is adaptively set according to its context as:

$$Th_i = a \times \frac{1}{2 \times N} \sum_{n=-N}^{N} Conf(i-n-1, i-n) \quad (8)$$

where *N* is the number of the previous and succeeding distances to predict threshold, and a is an amplifier. In our algorithm, we set *N*= 8, $a = 1.5$ to obtain optimal result. That is, threshold is automatically set according to neighborhood of 16 second, which is assumed to be the minimum length for one segment.

The threshold works well in the whole song, but we still need to refine the boundaries if more than one potential boundaries exit in 16 second, since it contravenes our assumption on the minimum length of a segment. In this case, the distances between the current segment and its two neighbor segments are compared, and then combined with the more similar one.

5 Experiments

Two experiments are presented in this section to evaluate the proposed mood detection system. The first experiment shows the performance on the selected music clips, inside of which the mood type is consistent. In the second experiment, the mood tracking method is evaluated with some famous music works.

5.1 Mood Detection on 20s Music Clips

Our database contains about 250 pieces of music, composed mainly in the classical period and romantic period. Choir, orchestra, piano and string quartet are all included to ensure the diversity of music style in the database. Three music experts participated in selecting and annotating 200 representative music clips of 20 seconds long from the database for each of the four mood clusters: Contentment, Depression, Exuberance and Anxious. All these 800 music clips are used in the evaluation.

Among these four clusters, clips in the Contentment cluster are mainly selected from Christian music and Serenade, while Exuberance clips are mainly from Overture, March and Dancing music. As for Depression and Anxious clusters, music is selected from much broader music genres, since there are no dominant genres as in the Contentment and Exuberance clusters. Since the mood is usually changeable in a whole piece of classical music as we mentioned before, each music clip is of 20 seconds long, and selected carefully to ensure the perceived mood is consistent and representative. One example is Suppe's "Light Cavalry", which contains two clips in Depression and three clips in Exuberance.

The classification results are calculated using a cross-validation evaluation where the dataset to be evaluated is randomly partitioned so the 25% is used for testing and 75% is used for training. The process is iterated with different random partitions and the results are averaged (for Table 2 and Table 3, 10 iterations were performed). It ensures that the calculated accuracy will not be biased because of a particular partitioning of training and testing. The ± part shows the standard deviation of classification accuracy.

In order to emphasize the importance of timbre and rhythm features in different mood groups, we used different weighting factors in Equation 3 and achieved the optimal average accuracy when $l_1 = 0.8$, $l_2 = 0.4$. It confirms that timbre features are more important to classify Contentment and Depression in Group 1, and rhythm features are more important to discriminate Exuberance and Anxious in Group 2.

	Contentment	**Depression**	**Exuberance**	**Anxious**
Contentment	**76.6±7.6**	21.8±7.2	0.5±0.8	1.2±1.2
Depression	4.0±3.5	**94.5±3.4**	0±0	1.5±2.5
Exuberance	0±0	0.8±1.3	**85.5±3.2**	13.7±4.8
Anxious	0±0	0±0	11.5±6.7	**88.5±6.7**

Table 2: Mood detection confusion matrix based on hierarchical framework

	Contentment	**Depression**	**Exuberance**	**Anxious**
Contentment	**75.0±11.8**	25.0±11.8	0±0	0±0
Depression	5.8±2.6	**94.2±2.6**	0±0	0±0
Exuberance	1.5±2.6	0.7±1.3	**64.7±20.5**	33.0±18.3
Anxious	0±0	0±0	11.7±7.9	**88.3±7.9**

Table 3: Mood detection confusion matrix based on non-hierarchical framework

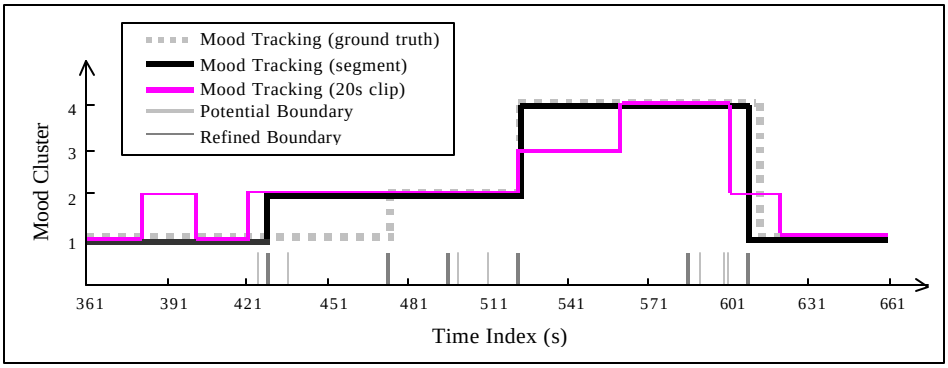

Figure 5: Mood tracking results on a part of "1812 Overture"

Table 2 shows the detailed results in the form of confusion matrix, where each row corresponds to the actual mood cluster and each column to the predicted cluster. As can be seen from Table 2, only 1.6% music in Group 1 (Contentment and Depression) is classified into Group 2 (Exuberance and Anxious), while only 0.4% music in Group 2 is classified into Group 1. That is, the accuracy rate reaches about 99% in the first step, when intensity features are used to classify all music clips into two groups. This result confirms the good performance of intensity features in discriminating the two groups of mood clusters, which severs as the basis for further classification by timbre and rhythm features.

In order to compare the performance of above hierarchical framework and its non-hierarchical counterpart, a comparative experiment is also performed on the framework shown in Fig. 4, which integrates the three feature sets and carries on classification directly on it. The corresponding results are shown in Table 3. Compared Table 2 and Table 3, it can be seen that the overall classification accuracy for the proposed hierarchical framework is up to 86.3%, about 5.7% better than the non-hierarchical framework. Meanwhile, the standard deviation of classification accuracy decreases from 10.7% to 5.2%, which indicates that our framework is more constant. It can be also seen, by adopting the proposed hieratical framework, the classification accuracies for all of the four clusters are improved, especially for Exuberance. In non-hierarchical framework, about 33.0% Exuberance clips are classified into Anxious, while it is decreased by more than 50% after using our hierarchical framework. These experimental results show that the proposed hierarchical framework has a better performance than its non-hierarchical counterpart, by using the most efficient features for different mood clusters.

5.2 Mood Tracking

The proposed mood tracking method is also evaluated on several pieces of classical music and achieved satisfactory result. For example, it can correctly detect that Haydn's "Serenade" is constantly Contentment; and the second movement of Beethoven's "Symphony No. 3" is mainly Depression.

Fig. 5 shows the results of mood tracking for a part of "1812 Overture" composed by Tchaikovsky (from 361s – 661s). The figure also shows the potential boundaries and refined boundaries. It can be seen that almost all of the correct boundaries are recalled, although there exist some false alarms. This ensures that the mood inside one segment is consistent. Compared the mood tracking results based on our approach, every 20s clips and the "ground truth", it can be also seen that since our approach can detect the boundaries well, the resulting mood tracking performance is better than that of detecting mood every 20 seconds.

6 Conclusion

In this paper, we present a mood detection approach for classical music from acoustic data. Thayer's model of mood is adopted for mood taxonomy, and three efficient feature sets are extracted directly from acoustic data representing intensity, timbre and rhythm respectively. A hierarchical framework is used to detect the mood in a music clip. In order to detect the mood in a whole piece of music, a segmentation scheme is presented for mood tracking. This algorithm achieves satisfactory accuracy in the experimental evaluations.

There are many future improvements in the proposed algorithm. We will work on extracting more powerful features to better represent music primitives in music perception. Furthermore, we will try more efficient ways for mood tracking. Finally, we will extent this mood detection algorithm to other styles such as pop music.

References

[1] Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proceeding of the IEEE*, 85 (9), 1437-1462.

[2] Erling, W., et al (1996). Content-based dassification, search, and retrieval of audio. *IEEE Trans. Multimedia*, 3, 27-36.

[3] Hevner, K. (1935). Expression in music: a discussion of experimental studies and theories. *Psychological Review*, 42, 186-204.

[4] Hinn, D. M. (1996). *The effect of the major and minor mode in music as a mood induction procedure*. Master

Thesis, Virginia Polytechnic Institute.

[5] Huron, D. (1992). The ramp archetype and the maintenance of auditory attention. *Music Perception*, 10 (1), 83-92.

[6] Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. *International Symposium on Music Information Retrieval* (ISMIR) 2000.

[7] Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H. & Cai, L. H. (2002). Music type classification by spectral contrast features. *Proceeding of Int. Conf. Multimedia Expo*.

[8] Kamien, R. (1992). *Music: an appreciation (5th Edition)*. McGraw-Hill Inc.

[9] Katayose, H., Imai, M. & Inokuchi, S. (1988). Sentiment extraction in music. *Proceeding of Int. Conf. Pattern Recognition*, 2, (pp. 1083-1087).

[10] Krumhansl, C. L. (2002). Music: a link between cognition and emotion. *Current Directions in Psychological Science*, 11(2), 45-50.

[11] Liu, D., Zhang, N. Y. & Zhu, H. C. (2003). Form and mood recognition of Johann Strauss's waltz centos. *Chinese Journal of Electronics*, 3. (in press)

[12] McCallum, A., et al (1998). Improving text classification by shrinkage in a hierarchy of classes. *Proceeding of. Int. Conf. Machine Learning*, (pp. 359-367).

[13] Radocy, E. & Boyle, J. D. (1988). *Psychological foundations of musical behavior.* Illinois: Charles C Thomas.

[14] Scheirer, E. D. (2000). *Music-listening systems*. Ph. D. Thesis, MIT Media Lab.

[15] Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford University Press.

[16] Tzanetakis, G. & Cook, P. (1999). Multifeature audio segmentation for browsing and annotation. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (pp. 17-20).

[17] Tzanetakis, G. & Cook, P. (2002). Music genre classification of audio signals. *IEEE Trans. Speech Audio Processing*, 10 (5), 293-302.

[18] Zhang, T. & Kuo, J. (1998). Hierarchical system for content-based audio classification and retrieval. *Proceeding of SPIE's Conference on Multimedia Storage and Archiving Systems III*, 3527, (pp. 398-409).