Perceptual Analysis for Music Segmentation

Min-Hong Jian, Chia Han Lin and Arbee L.P. Chen*

Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan 300, R.O.C.
Email : alpchen@cs.nthu.edu.tw

Abstract

In this paper, a music segmentation framework is proposed to segment music streams based on human perception. In the proposed framework, three perceptual features corresponding to four perceptual properties are extracted. By analyzing the trajectory of feature values, the cutting points of a music stream can be identified. According to the complementary characteristics of the three features, a ranking algorithm is designed to achieve a better accuracy. We perform a series of experiments to evaluate the Complementary Characteristics and the effectiveness of the proposed framework.

Keywords: Music content analysis, perceptual features, harmonic melody, music segmentation

1. Introduction

Due to a great progress of computer technologies and the mature development of Internet, more and more audio data are distributed in the Internet. The management of the audio data, including indexing, classifying and retrieving the audio data, has become an important issue. Most of the audio data management researches are based on the content of the audio data. In [15], the research issues of audio content analysis are categorized into four directions: *audio segmentation and classification*, *content-based audio retrieval*, *audio analysis for video indexing*, and *integration of audio and visual information for video segmentation and indexing*. Since the audio data contains a lot of information, the first step of most audio content analysis researches is the segmentation of audio data. Most of audio segmentation researches [5][11][14][15] are based on the variation of sound types. In these works, the sound type, such as speech, music, and environment sounds, is identified by analyzing the audio data based on various methods, such as zero-crossing rate, FFT, short-term energy function.

Recently, among various audio data, music data have become more popular because of the rich information contained in the music data and the interesting music styles. Therefore, more and more researchers have focused on the content analysis of music data. The music data can be divided into two groups based on the representation form. The first one is the symbolic representation such as MIDI format, and the other one is the acoustic signal representation such as Wave and Mp3 formats. Because of the limited music quality and the lack of human voice representation for the symbolic music format, most music data are distributed as acoustic signal formats. Therefore, we focus on the music segmentation of the music data with the acoustic signal type. In [1], the audio segmentation algorithms are divided into two categories: *Model-based algorithms* and *Novelty-based algorithms*. The model-based algorithms match the trajectory of feature values with a model for identifying and labeling the audio segment, and the novelty-based algorithms identify abrupt changes in the trajectory of features values.

There are some model-based methods for music segmentation. Herrera[4] presented useful strategies to build an music content analysis system and reviewed different model-based methods for the music segmentation based on musical instruments, such as SVM, Neural network, Bayesian classifiers. Rauber[8] used Hidden Markov Models to segment the musical signals into continues sequence based on notes and rests. However, these model-based methods need a supervised learning. Obviously, both modeling and training are time consuming.

As for novelty-based methods, there exit some general methodologies of music segmentation. Tzanetakis[13] implemented some schemes to segment audio streams, such as spectral centroid, spectral flux and Zero-Crossing

* To whom all correspondence should be sent.

Rate(ZCR). In addition, he also developed a method to evaluate the segmentation performance. Foote[3] used acoustical parameters in Slaney's auditory toolbox[12] to calculate the local self-similarity in the music, and he also defined a kernel correlation to calculate the audio novelty for music segmentation. However, the human perception is not considered in these approaches.

Martin[6] argued that the music content analysis system should advocate more psycho-acoustic perspectives rather than music theory and note-level transcription. There are few approaches considering human auditory system and take advantage of *psycho-acoustics*. Psycho-acoustics is the relationship between human acoustic and perception. It's also a measurement of the sensitivity to sounds in human auditory system. Scheirer[10] built music listening system including psycho-acoustics, music psychology, and signal processing models. Thus, two music pieces can be compared based on these computational models. However, these models are too complicated and only pitch, loudness and beats are considered. Pampalk[7] and Rauber[9] found rhythm patterns based on specific loudness sensation to perform genre classification. The loudness sensation of human depends on different frequencies of sounds. They used the equal loudness contours of Fletcher-Munson to calculate the loudness sensation of various music pieces. However, only the loudness of psycho-acoustics is considered and this framework cannot apply to various styles of music.

In this paper, a music segmentation framework is proposed to segment music data based on human perception. In the proposed framework, three perceptual features corresponding to four perceptual properties are extracted. A set of cutting points of a music data based on each perceptual feature can be identified by analyzing the corresponding trajectory of feature values. According to the complementary characteristics of the three features, a ranking algorithm is designed and applied on the sets of cutting points to achieve a better accuracy. We perform a series of experiments to evaluate the Complementary Characteristics and the effectiveness of the proposed framework.

The rest of this paper is organized as follows. The overview of the proposed framework is described in Section 2. Three psycho-acoustic features and some demonstrated experiments are discussed in detail in Section 3. The feature analysis for boundary detection and the ranking algorithm are discussed in Section 4. Experiments results are shown in Section 5. Finally, Section 6 concludes this paper and point out our future works.

2. The proposed framework

According to [6], there is a stronger evidence that most people are more sensitive to the perceptual features rather than structures of the music. Cook[2] discussed perceptual properties in psycho-acoustics, such as pitch, loudness, timbre and beats. The relationships between these properties and human perception of listening are also discussed. Therefore, we define the "*Harmonic melody*" and "*Cutting point*" of a music stream and propose a music segmentation framework based on the perceptual properties.

Definition: *Harmonic melody*

A harmonic melody is a music scene within which the pitch, loudness, timbre and beats are similar.

Definition: *Cutting point*

A cutting point occurs when a harmonic melody is changing to another harmonic melody.

Based on the definition, two melodies A and B are considered as different if the transposition between A and B is large enough or the timbre, loudness or beats of A and B are different.

Figure 1 shows the overview of the proposed framework. The preprocessing stage transforms music archives into .Wav file format. Then, the *roughness*, *periodicity pitch* and loudness of a music object will be extracted in the feature extraction stage. In the boundary detection stage, some heuristic schemes of digital signal processing are used to identify the abrupt changes in the trajectory of the feature values. The cutting points corresponding to each feature can therefore be identified.

3. Feature extraction

Three features are used in the proposed framework to identify the cutting points.

3.1 Roughness

The roughness of a sound depends on the frequency distribution and the waveform of the sound. If the frequency distributions of two sounds are different, the beats of the two sounds we feel based on the psycho-acoustics will be different. Moreover, if the waveforms of two sounds are different, the timbre of these two sounds will be different. Therefore, if the roughnesses of two music scenes are different, the beat or the timbre of the two music scenes will be different too.

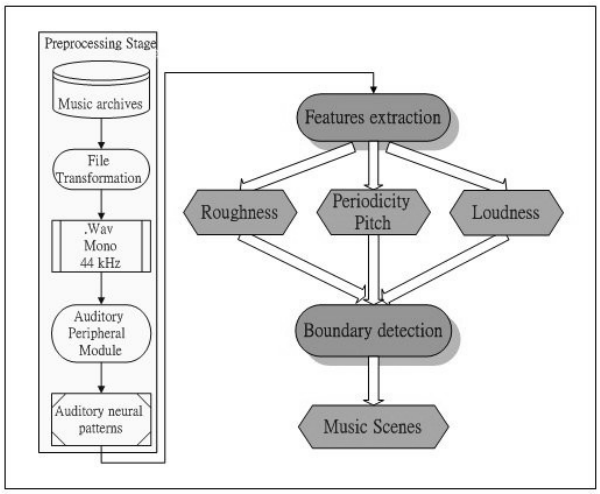

Figure 1. Overview of the proposed framework

Figure 2 shows an example that a music piece has regular strong beats before 8 second, and a soft melody after 8 second. Two music scenes are distinguished at 8 second based on beats. Figure 3 shows an example that the music piece has one timbre before 2.5 second and another timbre after 2.5 second. Therefore, two music scenes can be distinguished by using the roughness based on the timbres.

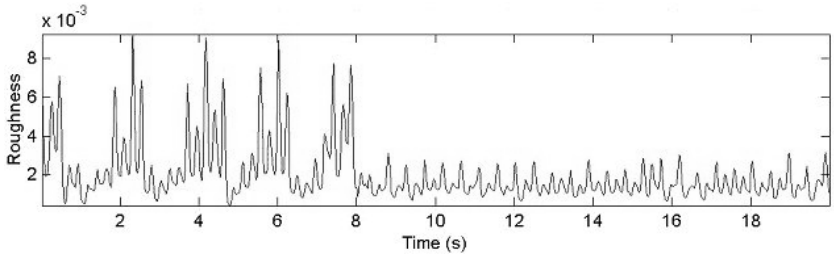

Figure 2. The roughness of a music piece with two different beats.

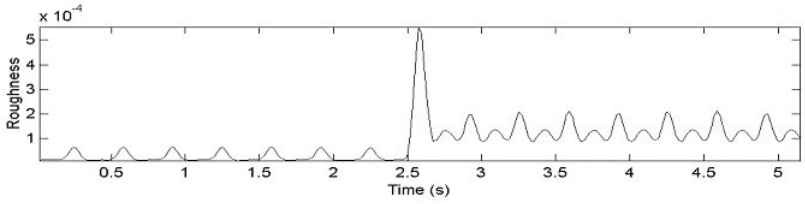

Figure 3. The roughness of a music piece with two different timbres.

3.2 Periodicity pitch

Most musical instruments have a clear pitch which is associated with the periodicity of the sound they produce. Moreover, the periodicity pitch is an $m \times 1$ matrix which records energies of m different period levels at each sampling frame. Thus, two music scenes can be distinguished if their pitches are different, as shown in figure 4. And the cutting point is at 1.9 second.

3.3 Loudness

Loudness is the intensity of the sound. Since human is sensitive to the sound intensity, loudness is an important cue for the change of music scenes. RMS (root-mean-square) level in decibels is used to measure the intensity, which is the square root of the sum of the squares of the windowed sample values. Figure 5 shows an example, and the cutting points are at 1 and 1.9 seconds since the intensities of the sound are different. The loudness is also an $m \times 1$ matrix which records intensities of m different frequency levels at each sampling frame.

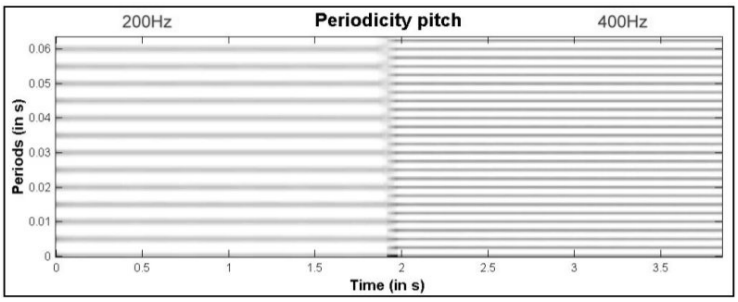

Figure 4. Two music scenes distinguished by different periodicity pitch.

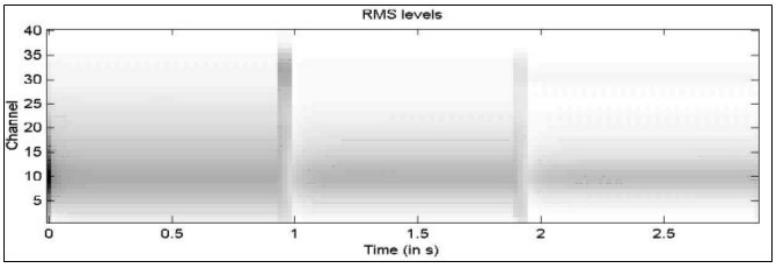

Figure 5. Different RMS levels in a music piece.

3.4 Complementary Characteristics

Although the three features discussed in the previous subsection are useful to most situations, there still exist some limitations.

As shown in Table 1, for roughness, the frequency distribution of the music is considered. The change of music scene based on the timbre and beats can be distinguished by using roughness, which may not be found by the other features. For example, two music scenes with the same pitch, 500Hz and 503Hz mixed, performed by different instruments can be distinguished by roughness based on the timbre, as shown in Figure 6. For periodicity pitch, the precise frequencies of the music are considered. The change of music scene based on the pitch can be distinguished by using periodicity pitch, which may not be found by the other features. For example, although the roughness of A from 400Hz and 405Hz is similar with B from 600Hz and 605Hz based on the frequency distribution, A and B can be distinguished by different pitch combinations, as shown in Figure 7. For loudness, the intensity of music is considered. The change of music scene based on the loudness can be distinguished by using loudness, which may not be found by the other features since the pitch or the timbre may not change.

	Merits	Limitations
Roughness	Timbre, beats	Pitch and loudness
Periodicity pitch	Pitch	Timbre, beats and intensity
Loudness	loudness	Timbre, beats and pitch

Table 1. The merits and limitations are for three features.

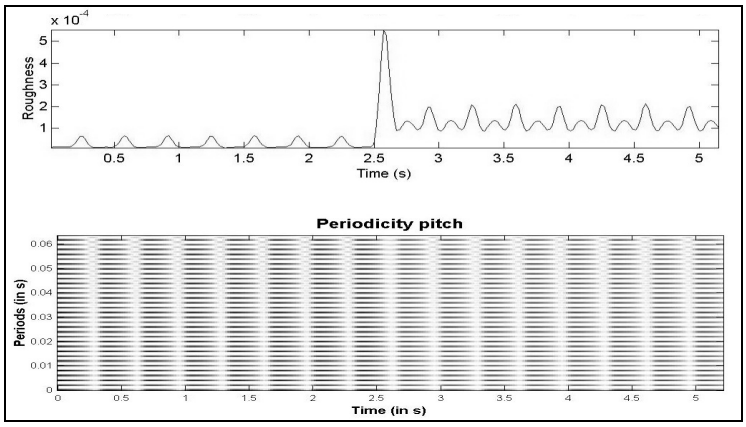

Figure 6. There are two music scenes in the sound and the cutting point is at 2.5 seconds.

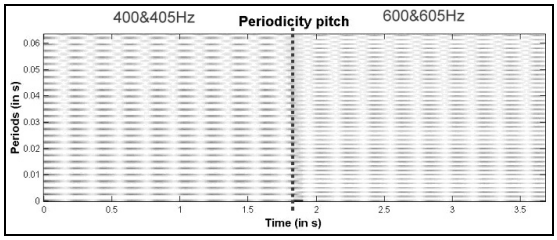

Figure 7. Different pitch combinations result in two music scenes.

4. Feature analysis for Boundary detection

Some heuristic methods, such as Spectrum Flux, variance and STFV are used to analyze the feature values for the boundary detection. Additionally, a novel ranking algorithm is designed to find the cutting points, such that the "Complementary Characteristics" is considered. There are four steps in the boundary detection as shown in figure 8.

Step1. Spectral Flux[11]

It is the difference between two feature vectors evaluated at two successive sound frames. It can also be expressed as follows.

$$Flux = \left\| |X_i| - |X_{i+1}| \right\|$$

By calculating the flux of feature vector, the $m \times 1$ matrix can be reduced to a single value at each sampling frame. Therefore, the abrupt changes of feature values can be easily detected as shown is Figure 9. The calculation of flux will be applied to the features except the roughness which is a single value at each sampling frame.

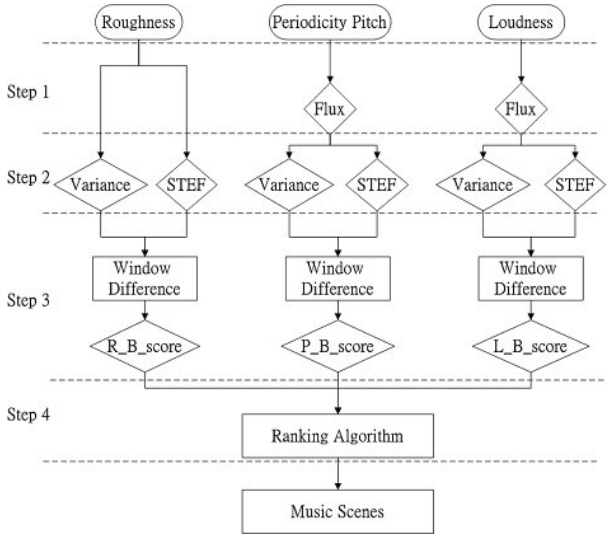

Figure 8. The boundary detection flowchart.

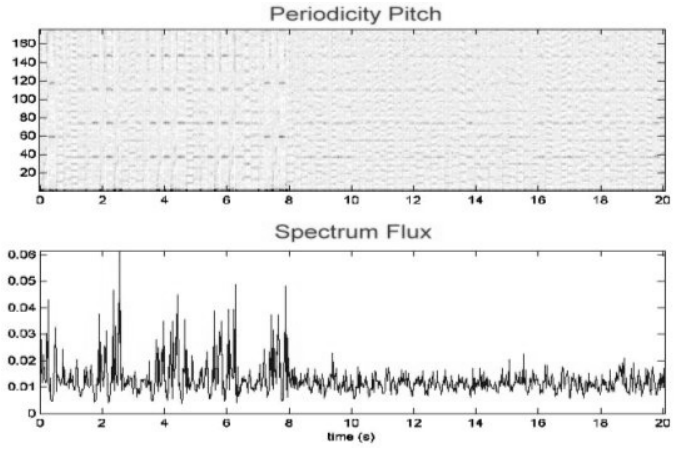

Figure 9. The spectrum flux of the periodicity pitch.

Step2.Variance and STFV (Short-Term Feature Value)

Variance and STFV of feature arrays are used to figure out the abrupt change of feature values based on the local and global views. Variances of feature array within a window are computed, as shown in figure 10. There exists a cutting point if an abrupt change of variances occurs. The STFV calculate the sum of feature values within a window which is a sliding window starting from each sampling frame, as shown in figure 11, which provides more global view of patterns in feature vectors. The formula of STFV is shown as follows.

$$STFV_n = \frac{1}{N} \sum_{m=1,\ldots,N} x(n+m)$$

where

n = the number of the sampling frame
N= the size of sliding window
x(m) = the feature value at sampling frame *m*

The variance and STEV are calculated for every sampling frame, and the size of the sliding window is 75 sampling frames.

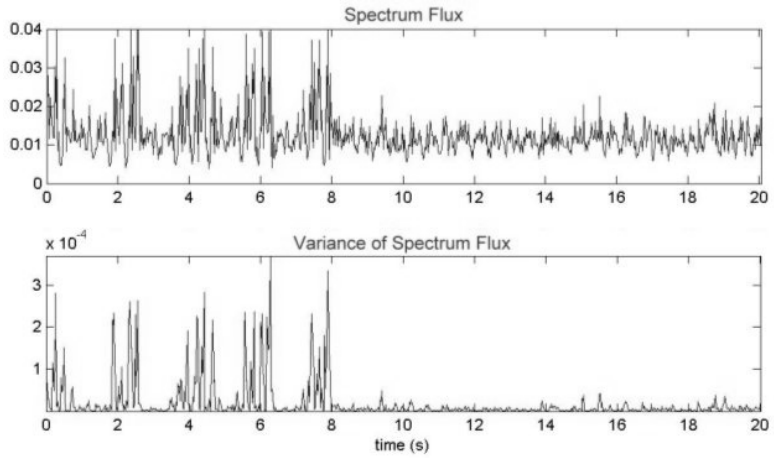

Figure 10. The variance of spectrum flux.

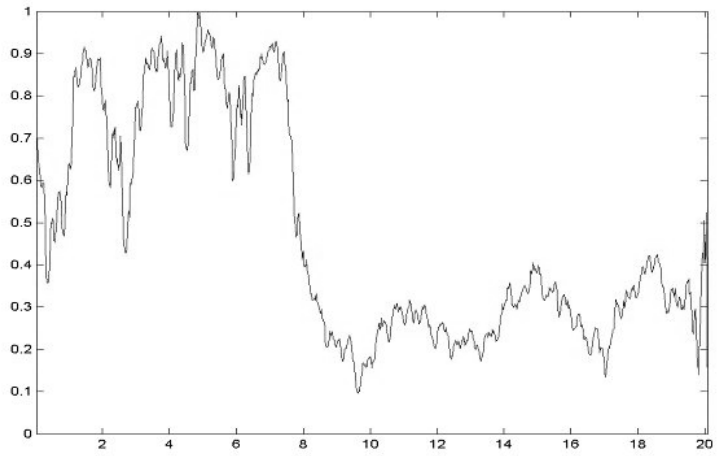

Figure 11. The STFV of spectrum flux.

Step3. Window Difference

The difference between two adjacent window of the variance and STFV is calculated for each sampling frame to obtain the window differences. Figure 12 shows an example. For the sampling frame A, the difference between the sum of STFV values within window1 and window2 is calculated. The window differences of variance and of STFV are calculated to obtain the B_score (Boundary score) for each feature. The B_score for each feature will be used in the ranking algorithm to find the cutting point. The formula of B_score is shown as follows:

$$B\_score = w1 * win\_diff\_of\_Variance + w2 * win\_diff\_of\_STEF$$

where w1,w2 are the weights.

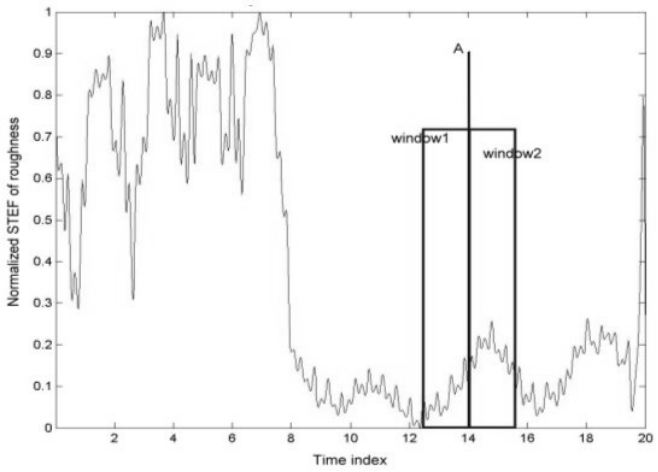

Figure 12. The difference of adjacent windows.

Step4. Ranking Algorithm

After step3, three B_scores corresponding to roughness, periodicity pitch and loudness are calculated. Most segmentation methods use the summation of these scores to figure out cutting points or use some weighting functions to give different weights to each feature. However, the relationships between features are not considered, and the desired cutting points may be lost. Figure 13 shows an example. There are four cutting points ordered by the cutting scores and only three cutting point will be selected. Based on the summation of scores, the most outstanding cutting point corresponding to some feature, such as cutting points D, may not be found. However, D is a good cutting point since it can be identified by roughness, although it may not be identified by using periodicity pitch and loudness.

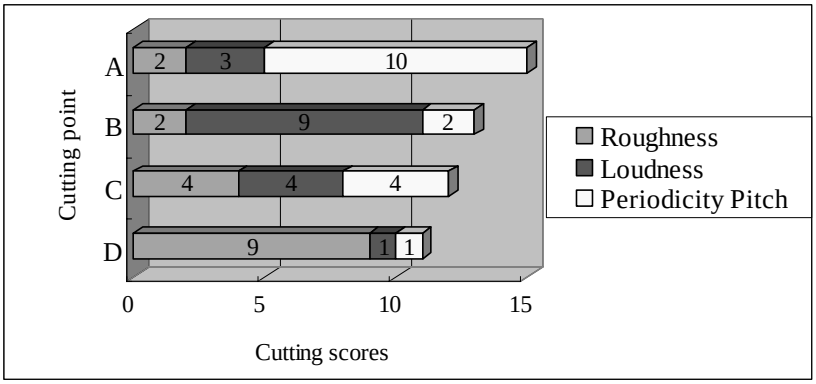

Figure 13. An example of cutting point selection bases on the summation of B_scores of each feature.

Therefore, the ranking algorithm considering each feature individually is proposed to find the cutting points. Each feature has own peaks which are the candidate of cutting points, as shown in figure 14. The peak with the highest B_score of each feature will be selected by the ranking algorithm. According to the feature priority, the time stamps of these selected peaks will be inserted into an array of cutting points. If the time stamp between the selected peak and some peaks in the array is within the minimum distance, the incoming peak will be filtered out.

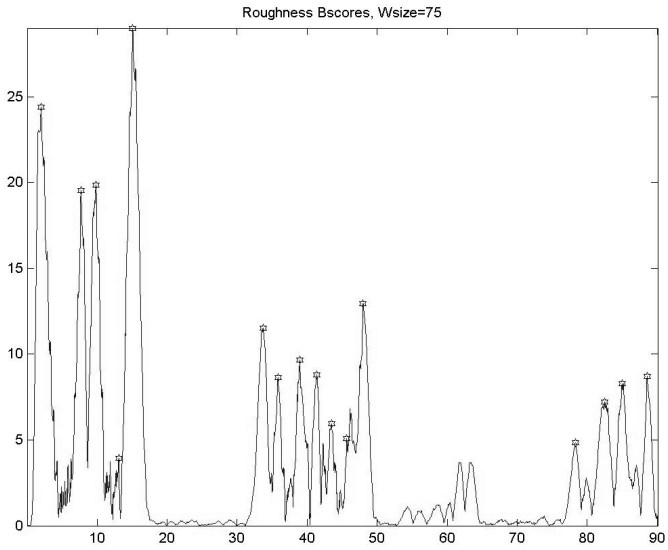

Figure 14. The roughness peaks order by B_score of roughness.

5. Experiment results

In this paper, the evaluated method proposed by Tzanetakis and Cook[13] is used to show the effectiveness of the proposed music segmentation framework. Ten popular music clips including various styles are used in this experiment. The experiment results are shown in the following Tables. The value of AG, FB and BE are calculated as follows:

$$AG = \frac{|CP_H|}{|\cup CP_i|}, \quad FB = \frac{|CP_A \cap CP_H|}{|CP_H|}, \quad BE = \frac{|CP_A' \cap CP_H|}{|CP_H|}$$

where CP_i : the cutting point identified by user i

CP_H : a set of cutting points identified by most users (commonly selected cutting points)

CP_A : k cutting points identified by the proposed framework, k = $|CP_H|$

CP_A': k cutting points identified by the proposed framework, k =24

AG is the measure of human consistency about cutting points, i.e., the ratio of the number of the commonly selected cutting points. We assume recall refers to the percentage of commonly selected cutting points identified by the automatic segmentation approach. Therefore, FB is the recall when the automatic segmentation approach identify the same number of cutting points as $|CP_H|$. BE is the best recall for the automatic segmentation approach by increasing the number of the identified cutting points up to 24. Moreover, the value of MX is the number of the cutting points necessary to reach the best recall.

According to our experiments, the average FB of roughness, periodicity pitch and loudness is 49%, 53.9 and 51.7%, respectively. Therefore, the priority of the three features is P>L>R (Periodicity pitch > loudness > roughness). This priority will be used in the ranking algorithm. The ranking algorithm results in encouraging results better than each single feature, as shown Table 2~5. Table 5 shows that the FB can reach 62.8% and best effort can reach 92.1% for popular music

.

Popular Music		Human Agreement		Automatic				
				Fixed Budget		Best Effort		
Style		AG	%	FB	%	BE	MX	%
1.	Pop	6/9	67	3/6	50	6/6	13	100
2.	Pop	5/7	71	3/5	60	5/5	19	100
3.	Pop	6/9	67	3/6	50	6/6	16	100
4.	Electronic	8/9	89	5/8	63	7/8	18	88
5.	Electronic	6/8	75	3/6	50	5/6	24	83
6.	Rock	7/9	78	3/7	43	4/7	17	57
7.	Rock	6/9	67	4/6	67	5/6	24	83
8.	Rock	6/9	67	1/6	17	4/6	22	67
9.	Folk	6/10	60	2/6	33	6/6	23	100
10.	Folk	7/11	64	4/7	57	6/7	8	86
Average			71.2		49		18.4	86.4

Table 2. The segmentation results of roughness

Popular Music		Human Agreement		Automatic				
				Fixed Budget		Best Effort		
Style		AG	%	FB	%	BE	MX	%
1.	Pop	6/9	67	5/6	83	6/6	9	100
2.	Pop	5/7	71	1/5	20	3/5	17	60
3.	Pop	6/9	67	4/6	67	6/6	23	100
4.	Electronic	8/9	89	7/8	88	8/8	10	100
5.	Electronic	6/8	75	1/6	17	5/6	13	83
6.	Rock	7/9	78	3/7	43	6/7	18	86
7.	Rock	6/9	67	2/6	33	5/6	24	83
8.	Rock	6/9	67	3/6	50	5/6	9	83
9.	Folk	6/10	60	4/6	67	5/6	11	83
10.	Folk	7/11	64	5/7	71	6/7	8	86
Average			71.2		53.9		14.2	86.4

Table 3. The segmentation results of periodicity pitch

Popular Music	Human Agreement		Automatic				
			Fixed Budget		Best Effort		
Style	AG	%	FB	%	BE	MX	%
1. Pop	6/9	67	3/6	50	5/6	12	80
2. Pop	5/7	71	2/5	40	3/5	6	60
3. Pop	6/9	67	5/6	83	5/6	6	83
4. Electronic	8/9	89	6/8	75	8/8	22	100
5. Electronic	6/8	75	2/6	33	4/6	12	67
6. Rock	7/9	78	2/7	29	6/7	21	86
7. Rock	6/9	67	3/6	50	5/6	12	83
8. Rock	6/9	67	3/6	50	5/6	18	83
9. Folk	6/10	60	3/6	50	4/6	18	67
10. Folk	7/11	64	4/7	57	6/7	21	86
Average		71.2		51.7		14.8	79.5

Table 4. The segmentation results of loudness

Popular Music	Human Agreement		Automatic				
			Fixed Budget		Best Effort		
Style	AG	%	FB	%	BE	MX	%
1. Pop	6/9	67	5/6	83	6/6	14	100
2. Pop	5/7	71	2/5	40	5/5	21	100
3. Pop	6/9	67	4/6	67	6/6	11	100
4. Electronic	8/9	89	7/8	88	8/8	15	100
5. Electronic	6/8	75	3/6	50	6/6	15	100
6. Rock	7/9	78	2/7	29	6/7	24	86
7. Rock	6/9	67	3/6	50	5/6	18	83
8. Rock	6/9	67	4/6	67	5/6	9	83
9. Folk	6/10	60	5/6	83	5/6	6	83
10. Folk	7/11	64	5/7	71	6/7	8	86
Average		71.2		62.8		14.1	92.1

Table 5. The segmentation results after the ranking algorithm (P>L>R)

6. Conclusion

In this paper, a music segmentation framework is proposed based on the psycho-acoustics. According to the four psych-acoustic properties, the salient cutting points can be identified by three perceptual features. The Complementary Characteristics of these features are analyzed and used in the ranking algorithm of the music segmentation system. Finally, experiments show the encouraging results in the proposed framework. In the future, we will find more perceptual features to identify the different music scenes.

7. References

[1] Aucouturier, J.-J., "A (short) review of Music Segmentation Algorithms," http://www.csl.sony.fr/~jj/segmentation.html

[2] Cook, P. (1999). "Music, cognition, and computerized sound," *First MIT Press paperback edition*, 2001.

[3] Foote, J., "Automatic Audio Segmentation using a Measure of Audio Novelty," *in Proc. of IEEE International Conference on Multimedia and Expo*, vol. I, pp. 452-455, 2000.

[4] Herrera-Boyer, P., et al., "Towards Instrument Segmentation for Music Content Description: a Critical Review of Instrument Classification Techniques," *in Proc. of ISMIR*, 2000.

[5] Lu, L., H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," *in Proc. of ACM Multimedia*, Sept. 30- Oct. 5, 2001.

[6] Martin, K., E. Scheirer, and B. Vercoe, "Musical content analysis through models of audition," *in Proc. of ACM Multimedia Workshop on Content-Based Processing of Music*, Bristol, UK, 1998.

[7] Pampalk, E., A. Rauber, and D. Merkl., "Content-based organization and visualization of music archives," *in Proc. of ACM Multimedia*, December 1-6, 2002.

[8] Raphael, C., "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, NO4, April 1999.

[9] Rauber, A., E. Pampalk, D. Merkl, "Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Sound Similarity," *in Proc. of ISMIR*, October 13-17, 2002.

[10] Scheirer, E., "Music-Listening Systems," *MIT Media Laboratory*, 2000.

[11] Scheirer, E. and M. Slaney, "Construction and evaluation of a robust multi-feature speech/music discriminator," *in IEEE Transactions on Acoustics, Speech and Signal Processing*, PP. 1331-1334, 1997.

[12] Slaney, M., "Auditory toolbox," *Technical Report*, Interval Research Corporation, Palo Alto, CA, 1998.

[13] Tzanetakis, G.. and Cook P., "Multifeature Audio Segmentation for Browsing and Annotation," *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct 1999.

[14] Tzanetakis, G., and Cook, P., "Audio Analysis using the Discrete Wavelet Transform," *in Proc. of WSES Int. Conf. Acoustics and Music: Theory and Application and Applications (AMTA2001)*.

[15] Zhang, T., and C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *in IEEE Transactions on Speech and Audio Processing*, Vol 9. No 4, 2001.