AUTOMATIC MUSIC SUMMARIZATION BASED ON MUSIC STRUCTURE ANALYSIS

*Xi Shao[#] *, Namunu C Maddage[#] * ,Changsheng Xu[#], Mohan S Kankanhalli**

[#]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{shaoxi, maddage,xucs}@i2r.a-star.edu.sg
*School of Computing, National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

In this paper, we present a novel approach for music summarization based on music structure analysis. From the audio signal, we first extract the note onset representing the time tempo of the song and the music structure analysis can be performed based on this tempo information. After music content has been structured into different semantic regions such as Introduction (Intro), Verse, Chorus, Ending (Outro), etc., the final music summary can be created with chorus and music phrases which are included anterior or posterior to selected chorus to get the desired length of the final summary. In this way, we can guarantee that the summaries begin and end at meaningful music phrase boundaries, which is a difficult problem for existing music summarization methods. Experiments show our proposed method can capture the main theme of the music compared to the ideal summaries selected by music experts and user subjective evaluation indicates our proposed method has a good performance.

1. INTRODUCTION

Automatic music summarization is very useful to music indexing, content-based music retrieval and on-line music distribution. Approaches aiming at automatic music summarization include two stages. The first stage is feature extraction. The music signal is cut into frames and each frame is characterized by features. Features related to texture, dynamics, rhythmic characteristics, melodic gestures and harmonic content are used. Unfortunately, some of these features are difficult to extract, and it is not always clear which features are most relevant. As a result, the first challenge in music summarization is to determine the relevant features and find a way to extract them. In the second stage (music structure analysis stage), the most repeated sections are identified based on similarity analysis using various methods. These approaches can be classified into two main categories: Machine Learning Approaches and Pattern Matching Approaches. Machine Learning Approaches [1][2][3] attempt to categorize each frame of a song into a certain cluster based on the similarity distance between this frame and other frames in the same song. Then the frame number of each cluster is used to measure the occurrence frequency. The final summary is generated based on the cluster that contains the largest number of frames. Pattern matching approaches [4][5][6] aim at matching the underlying candidate excerpts, which includes a fixed number of continuous frames, with the whole song. The final summary is generated based on the best matching excerpt. Although some good results are claimed in the previous methods, these methods can't guarantee the final summaries beginning and ending at meaningful music phrase boundaries (such as verse/chorus transitions), semantically corresponding to the linguistic uncompleted sentences in speech, which is not desirable to the listeners. In this paper, we propose a novel automatic music summarization scheme in two stages. In the first stage, the rhythm structure of the song is analyzed by note onset and the beat. Then, in the second stage, structure analysis method is used to structure the song content, and a song can be segmented into different parts according to the roles they play in the song, such as Introduction (Intro), Verse, Chorus, Ending (Outro), etc. Then, the summary is created with the chorus, which is melodically stronger than the verse [7] and the music phrases are included anterior or posterior to the selected chorus in order to get the desired length of the final summary. The rhythm information is useful for aligning musical phrases such that the generated summary has a smooth melody.

2. RHYTHM STRUCTURE ANALYSIS

As mentioned in [8], the rhythm of a song can be perceived by human listener as the beat which corresponds to the sequence of the equally spaced temporal units. Our proposed rhythm extraction approach is shown in Figure 1. We assume that the time-signature of an input song is 4/4 which is most commonly used meter in the popular songs. Another assumption is that the tempo is roughly constant for a certain song. These assumptions fit most of the popular song. As figure shows, we first decompose the song into 8 subbands, whose

frequency ranges are shown in Table 1, corresponding to octaves of music scale. The sub-band signals are segmented into 60ms windows with 50% overlap and both the frequency and energy transients are analyzed using the similar method to that in [9]. An energy-based detector is used on the upper subbands (subband 05-08) to detect the strong transient note onset, while a frequency based distance measure is formulated for use with the lower subbands (subband 01-04), because fundamental frequencies (F0s) and harmonics of music notes in popular music are strong in these sub-bands.

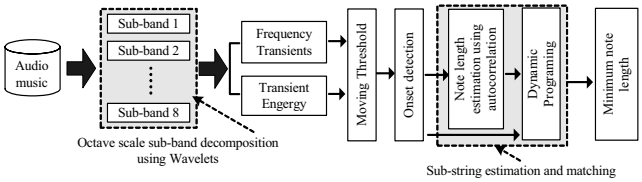

Figure 1: Rhythm tracking and extraction

Table 1: The frequency ranges of the sub-bands

Sub-band No	01		02	03	04	05	06	07	08
Octave scale	~ B1	C2 ~ B2	C3 ~ B3	C4 ~ B4	C5 ~ B5	C6 ~ B6	C7 ~ B7	C8 ~ B8	Higher Octaves
Freq-range (Hz)	0 ~ 64	64~128	128~256	256~512	512~1024	1024~2048	2048~4096	4096~8192	(8192 ~ 22050)

In order to detect hard and soft onsets, we take the weighted summation of onsets, detected in each sub-band as described in Equation (1). *On(t)* is the sum of onsets detected in all eight sub-bands Sb_i *(t)* at time 't' in the music signal. In our experiments, it is noticed that hard onsets are found in sub-band 01 and 02 usually from bass drums, bass guitar and bass notes of piano. The timing of snares and side drums are highlighted in sub-band 07 and 08. These onsets can indicate the bar timing. The soft onsets are specially found in sub-band 03 to 06. Thus the weight matrix $w = \{0.6, 0.9, 0.7, 0.9, 0.7, 0.5, 0.8, 0.6\}$ is empirically found to be the best set for calculating soft and hard onsets to extract the inter-beat time lengths.

$$On(t) = \sum_{i=1}^{8} w(i).Sb_i(t) \quad (1)$$

The initial inter-beat length is estimated by taking the autocorrelation over the detected onsets. We employ dynamic programming approach to check for patterns of equally spaced strong and weak beats among the detected onsets, *On(t),* and compute both inter-beat length and the smallest note length τ. This smallest note length τ (i.e. eighth or sixteenth notes) are played in the bars to align the melody with the rhythm of the lyrics and fill the gap between lyrics. Thus segmenting the music into the smallest note length frames instead of conventional fixed length segmentation in speech processing is important to detect the vocal/instrumental boundaries and the chord changes accurately. In addition, the length of music phrase can be estimated using the smallest note length. We will describe it in detail in section 4.

3. MUSIC STRUCTURE ANALYSIS

After rhythm structure has been extracted from the song, we should perform music structure analysis. The song structure generally comprises of Intro, Verse, Chorus, Bridge and Outro. These sections are built upon the melody-based similarity regions and content-based similarity regions. Melody-based similarity regions are defined as having similar pitch contours constructed from the chord patterns. Content-based similarity regions are defined as the regions which have both similar vocal content and melody. Corresponding to the music structure, the Chorus sections and Verse sections in a song are considered to be the content-based similarity regions and melody-based similarity regions respectively.

Our music structure analysis procedure can be summarized in the following steps:

1) Segment the song according to the smallest note length that we detected using the method proposed in the previous section.
2) In order to detect the melody-based similarity regions, the chords of each segment are detected and sub chord patterns are matching with the whole song using the Dynamic Programming method. The melody-based similarity regions have the similar chords patterns..
3) In order to detect the content-based similarity regions, the vocal content of the melody-based similarity regions should be further analyzed. First, we apply a machine learning method to find the boundary between vocal and instrumental music. Then, features sensitive to vocal contents are extracted and similarities are measured between the similar melody-based similarity regions based on these features, and the regions with high similarity can be define as content-based similarity regions.
4) Based on detected melody-based and content-based similarity regions, music structure can be identified. For example, content-based similarity regions can be formed as chorus, melody-based similarity regions with vocal music(excluding chorus) can be formed as verses. Some other regions with pure instrumental music can be formed as Into, Bridge, Outro based on the time order of appearance in the song.

The detailed description of music structure analysis scheme can be found in [10].

4. MUSIC SUMMARY GENERATION

The aim of music summarization is to extract the most common and salient sections of a given music, and generally these sections are readily recognized or remembered by the listeners. In today's popular music, the Chorus is the strongest and most repeated part of the

song [7]. It definitely should be contained in the final summary.

However, since chorus appears several times at different places in a song, we should choose one of the appearances as our selection. The selection of the chorus can be arbitrary. In our proposed method, we choose the chorus which appears in earlier part of the song. Another important issue we should consider in final music summary generation is the length of summary. Normally, the chorus lasts less than the required length of the summary which is about 30 seconds. Therefore, the preceding music phrases or the succeeding music phrases should be integrated into the selected chorus to satisfy the length requirement for the summary. According to music theory [11] , one music phrase is usually four bars in length. Therefore, the rhythm information is useful for aligning musical phrases such that the generated summary has a smooth melody. For example, with the assumption that time-signature of an input song is 4/4, if the smallest note length we detected is τ, the length of the music phrase can be calculated according to the different note level of this smallest note. Table 2 lists the different music phrase lengths calculation schemes correspond to three different note levels to which the smallest note commonly belongs in popular songs.

Table 2: The music phrase length of the different note levels.

	Bar Length	Music phrase length
Quarter note level	4*τ	4*4*τ (16τ)
Eighth note level	8*τ	4*8*τ (32τ)
Sixteenth note level	16*τ	4*16*τ (64τ)

* The smallest note length detected is τ

Figure 2 illustrates the process for generating music summary based on the structural analysis. The summary is created with the chorus, which is melodically stronger than the verse and the music phrases are included anterior or posterior to selected chorus to get the desired length of the final summary.

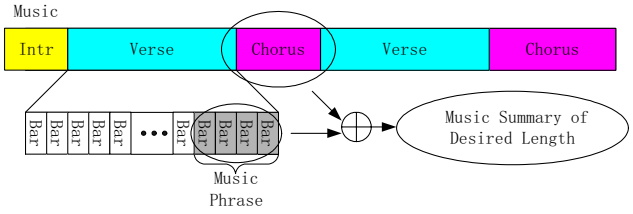

Figure 2: Music summarization using music structure analysis

5. EXPERIMENTS AND EVALUATION

In order to measure how well our music summary grasps the music themes, we should compare the summary generated by our proposed method with the ideal music summary. However, it is very difficult to get such a perfect summary. We assume the music summary generated manually by music experts as the ideal one. Figure 3 shows the comparison of summary for the song "Top of the world" (by Carpenter) selected by our proposed method, manually by music experts from EMI Singapore and by previous clustering method [3], separately.

We can see from the figure that summary selected using our proposed method has a high overlap ratio with the summary selected by music experts, and both of them contain the strongest and most repeated part of the song-chorus. While for summary created using previous clustering method, as we mentioned before, since it can't guarantee summaries beginning and ending at meaningful music phrase boundaries, some discontinuous sections and incomplete music phrases are included in the summary, which is not desirable to the listeners.

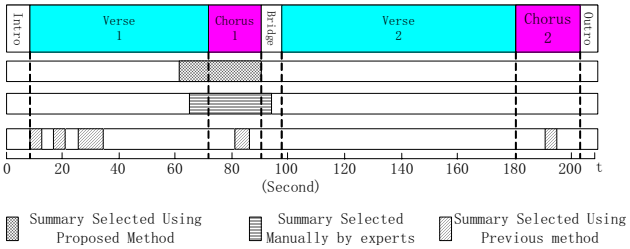

Figure 3: Experiment results on "Top of the world"

Since there is no ground truth available to evaluate the quality of a music summary, we employed a subjective study [12] to evaluate the performance of our music summarization approach. There are various attributes that are considered in an ideal summary and the summary must be evaluated based on these attributes.

a. *Clarity*: This pertains to the clearness and comprehensibility of the music summary. A good music summary should capture the gist of the original music.
b. *Conciseness*: This pertains to the terseness of the music summary. A good music summary should not contain much redundancy.
c. *Coherence*: This pertains to the consistency and natural drift of the segments in the music summary.
d. *Overall quality*: This pertains to the general perception or reaction of the users to the summaries.

Four genres of music were used in the test. They are Country, Classical, Rock and Jazz. Each genre contains five music samples. The aim of providing different music of different genres is to determine the effectiveness of the proposed method in creating summary of different genres. The length of music testing samples is from 2m52s to 3m33s. The length of the summary for each sample is 30s.

20 subjects with music experience are invited. Before the tests, the subjects could listen to each testing sample as many times as needed till he/she grasped the theme of the sample. Then the subjects listened to summaries generated from test samples and rated the summaries in four categories (Clarity, Conciseness, Coherence and Overall quality) on a scale of 1-5, corresponding to worst and best respectively. The average grade of summaries in each genre from all subjects is the final grade of this genre. In order to make comparison, we also asked subjects to rate summaries generated using our previously proposed clustering method [3] and summaries generated manually by experts. In addition, to remove the potential biased evaluation results, we presented the music summaries created by different methods in a random order, and the participants do not know which technique had been used to generate each summary before they rate the summaries.

Table 3: Results of User Evaluation.

Genre	Clarity			Conciseness		
	I	II	III	I	II	III
Country	4.5	4.7	4.3	4.1	4.2	3.7
Classic	4.0	4.2	3.7	4.0	4.3	3.8
Rock	4.6	4.5	4.2	4.5	4.5	4.1
Jazz	4.3	4.1	3.6	4.2	4.7	3.6

Genre	Coherence			Overall quality		
	I	II	III	I	II	III
Country	4.8	4.7	3.5	4.0	4.2	3.3
Classic	4.5	4.6	3.7	4.4	4.5	3.5
Rock	4.6	4.9	3.2	4.5	4.7	3.2
Jazz	4.7	4.7	3.3	4.2	4.3	3.0

I: Music summarized using our proposed method
II. Music summarized manually by music experts
III: Music summarized using our previous method

From the evaluation results in Table 3, it can be seen that the summarization using our proposed method is comparable with the summarization conducted manually by music experts for all genres of music testing samples, and has better performance than the previous clustering method.

6. CONCLUSIONS AND FUTURE WORK

We have presented a novel approach for music summarization based on music structure analysis. The experimental result and the evaluation by a subjective study have shown that the proposed summarization approach has a good performance.

In the future, we need to apply our music structure analysis based summarization to larger database containing other music genres which don't have strong music structure, such as Hip-Hop, Reggae etc. These weak music structure genres need a more sophisticated music structure analysis. In addition, we need to explore more relevant music knowledge to reveal the music structure.

7. REFERENCES

[1] Lu L, and Zhang H , Automated Extraction of Music Snippets, In *Proc. ACM International Conference on Multimedia* ,Berkeley, CA, USA, pp.140-147,2003.

[2] Logan B and Chu S , Music Summarization Using Key Phrases, In *Proc. IEEE International Conference on Audio ,Speech and Signal Processing*, Istanbul ,Turkey, vol.2 ,pp.II749 - II752, 2000.

[3] Xu C, Zhu Y and Tian Q, Automatic music summarization based on temporal, spectral and cepstral features, In *Proc. IEEE International Conference on Multimedia and Explore*, Lausanne, Switzerland, pp. 117-120, 2002.

[4] Chai W and Vercoe B, Music Thumbnailing via Structural Analysis, In *Proc. ACM international conference on Multimedia ,* Berkeley, CA, USA, pp.223-226,2003.

[5] Bartsch M A and Wakefield G H, To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing, In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics(WASPAA),* New Paltz, New York ,pp. 15 – 18 ,2001.

[6] Cooper M and Foote J , Automatic Music Summarization via Similarity Analysis, In *Proc. International Conference on Music Information Retrieval*, Paris, France, pp. 81-85, 2002.

[7] Ten Minute Master No 18: Song Structure. *MUSIC TECH magazine.* www.musictechmag.co.uk (Oct. 2003), 62 – 63.

[8] Scheirer, E. D. Tempo and Beat Analysis of Acoustic Musical Signals. *Journal of the Acoustical Society of America.* January 1998, Vol 103, No 1, 588 - 601.

[9] Duxburg. C, Sandler. M., and Davies. M. A Hybrid Approach to Musical Note Onset Detection. In *Proc. International Conference on DAFx.* 2002.

[10] Maddage C. N, Xu.C, Kankanhalli M.S , Shao X, Content-based Music Structure Analysis with the Applications to Music Semantic Understanding, *In ACM Multimedia Conference*, New York,2004.

[11] *Rudiments and Theory of Music.* The associated board of the royal schools of music, 14 Bedford Square, London, WC1B 3JG, 1949.

[12] Chin J.P., Diehl A.V. and Norman L.K., Development of an instrument measuring user satisfaction of the human-computer interface, In *Proceedings of SIGCHI'88,* pp.213-218 , New York, 1988.