Multimodal Structure Segmentation and Analysis of Music Using Audio and Textual Information

Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Ching Lin, and Homer H. Chen
Department of Electrical Engineering
National Taiwan University, Taipei 10617, Taiwan
{mikejdionline, affige, vagante}@gmail.com, homer@cc.ee.ntu.edu.tw

Abstract— **In this paper, we present a multimodal approach to structure segmentation of music with applications to audio content analysis and music information retrieval. In particular, since lyrics contain rich information about the semantic structure of a song, our approach incorporates lyrics to overcome the existing difficulties associated with large acoustic variation in music. We further design a constrained clustering algorithm for music segmentation and evaluate its performance on commercial recordings. Experimental results show that our method can effectively detect the boundaries and the types of semantic structure of music segments.**

Index Terms— **Music, segmentation, lyrics, music information retrieval**

I. Introduction

Understanding the structure of music (e.g. intro, verse, chorus, bridge, and outro) is important as it allows us to divide a song into semantically meaningful segments, within which musical characteristics are relatively consistent. Structure segmentation can serve as a front end processor for music content analysis [1] since it enables a local description of each disparate section rather than a coarse, global representation of the whole song. Thus, a user can input a favorite section of a song as a query to find similar music pieces. Structure segmentation can also be directly applied to music summarization [4] and thumbnailing [8], by which a user can quickly grasp the key section without listening to the whole song. In view of the fast growth of digital music collection and media playback on portable devices, such applications are indispensable.

Generally speaking, structure segmentation consists of two stages: an *audio segmentation* stage that divides audio into segments and a *semantic labeling* stage that labels each segment with a structure type. Although much work has been done in finding chorus or repeated parts in music [2]–[4], full-song audio segmentation remains challenging [6] unless some strong assumptions are made about the form of music being processed [5]. The difficulty exists because the number of segments is hard to infer from the audio signal, and audio features are not necessarily consistent even within segments of the same structure type due to common acoustic variations such as transposition, ornamentation, or improvisation. As the accuracy of audio segmentation is still low, rare attempts have been made to achieve semantic labeling.

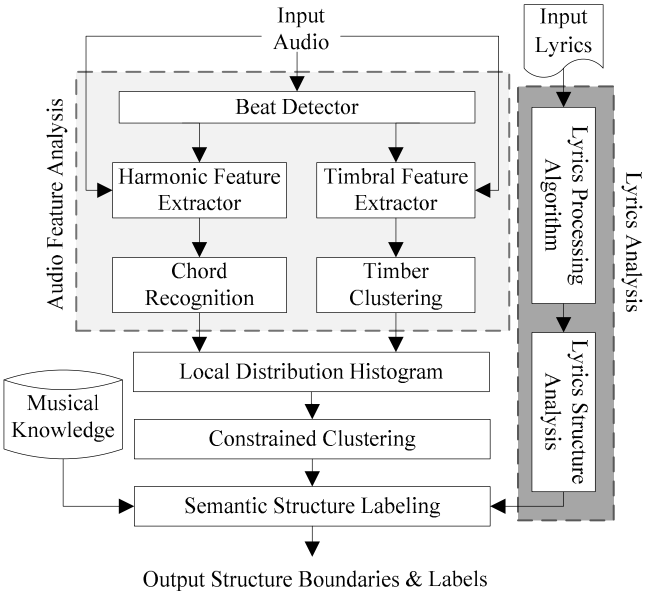
Input Audio
Input Lyrics
Audio Feature Analysis
Beat Detector
Harmonic Feature Extractor
Timbral Feature Extractor
Chord Recognition
Timber Clustering
Lyrics Processing Algorithm
Lyrics Structure Analysis
Lyrics Analysis
Local Distribution Histogram
Musical Knowledge
Constrained Clustering
Semantic Structure Labeling
Output Structure Boundaries & Labels

Fig. 1. The system framework of structure segmentation.

In light of the problems mentioned above, we propose a new framework utilizing both audio and lyrics information for structure segmentation. The incorporation of lyrics enables technical advancement in several aspects. First, it is relatively easier to infer the number of segments from the lyrics structure and thereby avoid over-segmentation and under-segmentation that often occurs in previous audio-based methods. Second, as the lyrics of the same structure type are mostly similar despite of acoustic variations, the computation of segment similarity becomes more reliable. Third, accurate semantic labeling is possible with the high-level semantic cues provided by the lyrics.

Fig. 1 shows the proposed system framework for multimodal structure segmentation. Audio features are leveraged to capture the local statistics of music frames, which are then clustered into several audio segments. To ensure temporal continuity and to obviate the risk of unreasonable segment lengths, we propose a constrained clustering algorithm that considers the information of segment length and neighboring labels by using the concept of energy minimization. Meanwhile, the lyrics is analyzed by a dynamic

This work is supported by a grant from Chunghwa Telecom under the contract RAC970362 and a grant from the National Science Council of Taiwan under the contract NSC 97-2221-E-002-111-MY3.

programming algorithm and incorporated to the semantic labeling process. To our best knowledge, this work is the first attempt that leverages lyrics for music structure segmentation.

The paper is organized as follows. Section II reviews previous work on music structure segmentation. The details of the proposed algorithm are elaborated in Sections III and IV. Section V shows the experimental results, and Section VI concludes the paper.

II. Previous Work

Recent research on structure analysis and segmentation of music can roughly be divided into two categories: similarity-matrix-based approach and clustering-based approach. In the former approach [2]–[4], [14], pair-wise similarity between two audio frames is computed by calculating the distance between the two associated feature vectors. Repetitive parts of music are then detected from the resulting similarity matrix. The latter approach [6], [8], [12] assumes that different structure types differ in the distributions of audio features and formulates audio segmentation as a clustering problem. For example, the method described in [6] uses Markov random fields for clustering and introduces a neighboring constraint to ensure temporal continuity of audio segments. Though some promising results have been reported, many existing works assume the form of music, such as the number of verse and chorus segments [5], is known a priori. Moreover, the problem of semantic structure labeling is usually left unaddressed.

III. Audio Segmentation

To enhance the accuracy of audio segmentation, we propose a constrained clustering algorithm to cluster local statistics related to the timbral and harmonic aspects of music. The use of constrained clustering improves the accuracy of segment boundary detection by imposing some local and global constraints. We describe each system component in details.

A. Beat Detection

After converting an input audio to mono channel and 22,050 Hz sampling rate, we partition it into basic processing units according to beat times detected by a beat tracking system called BeatRoot [9]. Audio features are then extracted from each beat interval, within which the music characteristics are likely to be more uniform and salient in contrast to those from frame-by-frame basis [1].

B. Audio Feature Representation

We adopt timbral features and harmonic features in this work because it has been reported [12] that combining them generally improves the performance of structure segmentation. For timbral feature, we extract the audio spectrum envelope, a power spectrum with frequency domain divided into logarithmically spaced subbands between 62.5Hz and 16 kHz to mimic human audition [10]. The resulting timbral feature vectors of a song are then clustered into 80 timbre types used in [6] with k-means clustering. Because a timbre type roughly corresponds to some combination of instruments [8] and different structure types of a song often consist of different combinations of instruments, the timbre type has been shown useful for structure segmentation [6].

1. **Initialize** the K centroids to be the histograms of K beat intervals evenly distributed in the song

2. **Assign** a label y_i^* to each beat interval such that

$$y_i^* = \arg\min_{y_i}(E_{data}(y_i) + E_{smooth}(y_i)) \qquad (1)$$

where

$$E_{data}(y_i) = d(x_i, c_{y_i}) \qquad (2)$$

$$E_{smooth}(y_i) = w\exp(f(y_i)) \qquad (3)$$

$$f(y_i) = \begin{cases} (T_{\min} - n(y_i))/z_{\min} & \text{if } n(y_i) < T_{\min} \\ (n(y_i) - T_{\max})/z_{\max} & \text{if } n(y_i) > T_{\max} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

3. **Update** c_k to be the mean of all x_i that belong to segment k

4. **Repeat** Step 2 and 3 until the assignment no longer changes or a predetermined number of iterations is reached

Fig. 2. The constrained clustering algorithm

While most existing works use pitch class profile (PCP) or chromagram to describe harmonic content, we adopt the higher-level chord sequence as it represents the harmonic progression and tonal structure of music effectively [1]. We use the automatic chord recognition algorithm proposed in [11] to recognize chords and assign each beat interval with one of the 24 major/minor triads, which can be considered as 24 chord types of harmonic combinations.

C. Local Distribution Histogram

After feature extraction, at each beat interval we count the timbre type and chord type distributions of neighboring beat intervals over a histogram window of length W. The resulting histogram represents the local characteristic of the music signal and captures specific timbral/harmonic patterns over a longer timescale [6]. The effect of W on the performance of segmentation is discussed in Section V.

D. Segmentation by Constrained Clustering

Using the local distribution histograms as observations, we propose a constrained clustering algorithm for assigning each beat interval to an optimal structure type. While conventional k-means algorithm minimizes total intra-cluster variance simply by assigning points to the nearest centroid, we introduce additional constraints based on neighboring and global information.

A brief description of the algorithm is provided in Fig. 2. Given observations $X = \{x_1, \ldots, x_N\}$, where N is the total number of beat intervals, we use the constrained clustering to divide X into K clusters (K is empirically set to 8) and regard the cluster labels of each beat interval $Y = \{y_1, \ldots, y_N\}$ as the corresponding structure types. The assignment of cluster labels is performed by minimizing Eq. (1), where the first term on the right hand side measures the L2-distance between an

Lyrics	N/A	I used to think that I could not go on/ And life was nothing but an awful song/ …	If I can see it / then I can do it / If I just believe it / there's nothing to it	I believe I can fly/ I believe I can touch the sky/ …	See I was on the verge of breaking down/ Sometimes silence can seem so loud/ …	If I can see it / then I can do it / If I just believe it / there's nothing to it	I believe I can fly/ I believe I can touch the sky/ …	Hey, cause I believe in you, oh	If I can see it / then I can do it / If I just believe it / there's nothing to it	I believe I can fly/ I believe I can touch the sky/ …	N/A
Label	intro	verse 1	verse 2	chorus	verse 1	verse 2	chorus	bridge	verse 2 (transposed)	chorus (transposed)	outro

Fig. 3. Lyrics and the semantic structure of the song "I believe I can fly."

observed histograms x_i and a cluster centroid, and the second term (a smoothness term) introduces a penalty function $f(.)$ to make the labeling of neighboring beat intervals smooth. As shown in Eq. (4), the smoothness terms is influenced by a factor $n(y_i)$ that denotes the number of consecutive beat intervals that have the same label up to the current beat interval. A penalty is imposed if $n(y_i)$ is smaller than a minimum length threshold $T_{\min}$, or larger than a maximum length threshold $T_{\max}$, while $z_{\min}$ and $z_{\max}$ are two constants that adjust the increasing rate of the penalty. The intuition behind the penalty function is that structure type does not change frequently in music. Since most songs can be partitioned into only 8 to 12 segments, no segments should be extremely short. Likewise, as a segment gets longer, it is more probable that another structure type is about to begin. The weight w in Eq. (1) is introduced to balance the influence of the two terms. Empirically we set the parameters via a separate validation set.

IV. Semantic Structure Analysis

Since a paragraph of the lyrics most likely corresponds to an audio segment of a song, we can map the structure type labels to the audio segments once the structure of lyrics is determined (Fig. 3). The system components of the semantic structure analysis shown in Fig. 1 are described in this section.

A. *Lyrics Processing*

Lyrics are readily available on the web and can be retrieved efficiently by a simple crawler or an automatic retrieval algorithm [7]. After acquisition, we measure the similarity between each pair of paragraphs by the longest common subsequence (LCS), a sequence of matched words whose orderings is unaltered. LCS is a suitable similarity measure because lyrics often partially repeat themselves in a song. Furthermore, this measure is language-independent as it only considers whether two words are the same or not. The recursive formulation of the dynamic programming algorithm that computes LCS is:

$$c[i,j]=\begin{cases}0 & \text{if } i=0 \text{ or } j=0\\ c[i-1,j-1]+1 & \text{if } i,\, j>0 \text{ and } a_i=b_j\\ \max(c[i,j-1],c[i-1,j]) & \text{if } i,\, j>0 \text{ and } a_i\neq b_j\end{cases}\qquad(5)$$

where $c[i, j]$ denotes the length of LCS of two input sequences A and B up to the i-th and j-th position, $A_i=(a_1, a_2, \ldots, a_i)$ and $B_j=(b_1, b_2, \ldots, b_j)$, $i<|A|$, $j<|B|$. Finally, we obtain $c[|A|, |B|]$, which is normalized by sequence length $\min(|A|,|B|)$ to a value between 0 (totally different) and 1 (exactly the same). If the normalized LCS length of two paragraphs exceeds a threshold τ, we label the two paragraphs with the same structure type. The parameter τ is empirically set to 0.6 in our work.

B. *Lyrics Structure Analysis*

The set of structure types we consider are intro, verse, chorus, bridge, and outro, the five major types of traditional popular music [5], [13]. According to basic definition and characteristic of each type of song structure [13], the strategy for determining lyrics structure is as follows:

- The most repeated paragraphs according to the LCS calculation are labeled as choruses.
- The shortest paragraph with no repetition, if exists, is labeled as bridge.
- The remaining paragraphs are labeled as verses.

C. *Semantic Structure Labeling of Audio Segments*

For the audio segments, we first directly label the first and the last audio segments as intro and outro, respectively, since they are instrumental parts with no lyrics. If the number of audio segments other than intro and outro exceeds the number of lyrics sections, the shortest segment is chosen and then merged with the neighboring segment whose local distribution histogram is more similar in terms of the L2-distance. This process is repeated until the number of audio segments is equal to that of lyrics sections, and then we are ready to completely map the lyrics structure labels to the audio segments.

V. Experimental Results

A. *Evaluation on Boundary Detection*

Due to the lack of a large common database for comparison between different methods, we evaluate our system on a dataset of 13 manually annotated popular songs of various genres and artists from Chinese and Western albums, which is available on our website[1]. For segment boundary detection, the performance evaluation is made on the precision and recall rate, and the f-value is computed by

$$\text{f-value}=2\cdot\text{precision}\cdot\text{recall}/(\text{precision}+\text{recall})\qquad(6)$$

which simultaneously considers the precision and recall rate. An estimated boundary is considered correct if it falls within 3 seconds from the ground-truth, as suggested in [6], [12].

As shown in Table I, the constrained clustering outperforms the traditional k-means clustering by 6.4% in terms of f-value. In particular, the use of constraints enhances the precision of segment boundary detection. The comparison

[1] http://sites.google.com/site/hengtzecheng/projects/iscas09

TABLE I
THE ACCURACY OF SEGMENT BOUNDARY DETECTION

Method	Precision	Recall	f-value
k-means	0.410	0.458	0.433
Constrained	0.488	0.510	0.497

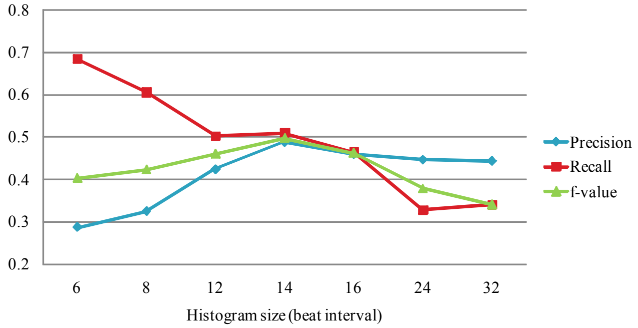

Fig. 4. The effect of histogram size on boundary detection accuracy.

of accuracy with different histogram size settings is shown in Fig. 4, where smaller histogram sizes yield lower precision yet a higher recall rate, and vice versa. The f-value reaches its maximum at somewhere between 14 to 16 beat intervals, a reasonable number because it corresponds to the common length of a music phrase, which is roughly the basic pattern of a music section. We set the histogram size to 14 in the following evaluations.

B. *Evaluation on Semantic Structure Labeling*

As shown in Table II, high precision and recall rate are achieved in intro and outro detection, mainly due to the effectiveness of the adopted timbral feature that discerns the difference between non-vocal and vocal parts. Our system also performs well on chorus detection, even for the 7 songs having transposed choruses in our dataset. This shows the robustness brought by the use of lyrics. The accuracy of verse detection is slightly worse than that of chorus detection, possibly because the boundaries and the repeated music patterns of verses are less clear, and some particularly short verses might be erroneously classified as bridges.

The low accuracy of bridge detection stems from its irregularity nature. Some bridges are purely instrumental, while some are vocal parts with lyrics. Though the problem is currently hard to solve, in most cases bridges are relatively short segments in music, which do not have large influence on the overall structure analysis. On the contrary, the promising results on verse and chorus detection are of great value in music information retrieval, since the two structure types are associated with the main theme of songs and are relatively important for common listeners.

In addition to precision and recall rate of semantic labeling for each structure type, we also calculate the overall labeling accuracy, which is measured on a frame-by-frame basis. Only an exact match between estimated label and ground-truth label is counted as a correct labeling. The overall labeling accuracy is 52.0%, which is as competitive as the state-of-the-art system that has also made attempt on semantic labeling [14].

TABLE II
THE ACCURACY OF SEMANTIC STRUCTURE LABELING

Structure Type	Average Time %	Precision	Recall	f-value
Intro	8.6%	0.720	0.820	0.767
Verse	27.9%	0.461	0.615	0.527
Chorus	43.3%	0.633	0.593	0.612
Bridge	10.8%	0.033	0.057	0.042
Outro	9.4%	0.873	0.628	0.731

VI. CONCLUSION

In this paper, we have presented a multimodal approach that contributes to music structure segmentation and analysis. For audio segmentation, the proposed constrained clustering algorithm improves the accuracy of boundary detection by introducing constraints on neighboring and global information. For semantic labeling, we derive the semantic structure of songs by lyrics processing to achieve a robust structure labeling. The consistency between the resulting segmentation and the lyrics structure, along with the promising results of semantic labeling, make our system particularly attractive for content-based and user-oriented music information retrieval.

REFERENCES

[1] M. Casey et al., "Content-based music information retrieval: current directions and future challenges," in *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[2] J. Foote, "Visualizing music and audio using self-similarity," in *Proc. ACM MM*, pp. 77–80, 1999.

[3] J. Paulus and A. Klapuri, "Music structure analysis by finding repeated parts," in *Proc. ACM Audio and Music Computing for Multimedia Workshop*, pp. 59–68, 2006.

[4] W. Chai, "Semantic segmentation and summarization of music," *IEEE Signal Process. Magazine*, pp. 124–132, 2006.

[5] N. C. Maddage et al., "Content-based music structure analysis with applications to music semantics understanding", in *Proc. ACM MM*, pp. 112–119, 2004.

[6] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE TASLP*, vol. 16, no. 2, pp. 318–326, 2008.

[7] G. Geleijnse and J. Korst, "Efficient lyrics extraction from the web," in *Proc. ISMIR*, pp. 371–372, 2006

[8] M. Levy et al., "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *Proc. ICASSP*, pp. 1433–1436, 2006.

[9] S. Dixon, "Evaluation of the audio beat tracking system BeatRoot," *Journal of New Music Research*, pp. 39–50, 2007.

[10] M. Casey, "General sound classification and similarity in MPEG-7," *Organised Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[11] H.-T. Cheng et al., "Automatic chord recognition for music classification and retrieval," in *Proc. IEEE International Conf. Multimedia and Expo.*, pp. 1505–1508, 2008.

[12] M. Levy, K. Noland, and M. Sandler, "A comparison of timbral and harmonic music segmentation algorithms", in *Proc. ICASSP*, pp. 1433–1436, 2007

[13] J. P. G. Mahedero et al., "Natural language processing of lyrics," in *Proc. ACM MM*, pp. 475–478, 2005.

[14] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and an integrated musicological model," in *Proc. ISMIR*, pp. 369–374, 2008