

Music Thumbnailing via Structural Analysis

Wei Chai

MIT Media Laboratory
Cambridge, MA 02139 USA
+1 617 2530112

chaiwei@media.mit.edu

Barry Vercoe

MIT Media Laboratory
Cambridge, MA 02139 USA
+1 617 2530112

bv@media.mit.edu

ABSTRACT

Music thumbnailing (or music summarization) aims at finding the most representative part of a song, which can be used for web browsing, web searching and music recommendation. Three strategies are proposed in this paper for automatically generating the thumbnails of music. All the strategies are based on the results of music structural analysis, which identifies the recurrent structure of musical signals. Instead of being evaluated subjectively, the generated thumbnails are evaluated by several criteria, mainly based on previous human experiments on music thumbnailing and the properties of thumbnails used for commercial web sites. Additionally, the performance of the structural analysis is demonstrated visually using figures for qualitative evaluation, and by three novel structural similarity metrics for quantitative evaluation. The preliminary results obtained using a corpus of Beatles' songs demonstrate the promise of our method and suggest that different thumbnailing strategies might be proper for different applications.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Abstracting methods and indexing methods; H.5.5 [Sound and Music Computing]: Signal analysis, synthesis, and processing.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Structural analysis, pattern matching, music segmentation, music thumbnailing, music information retrieval.

1. INTRODUCTION

Music thumbnailing (or music summarization) aims at finding the most representative part of a song, which can be used for web browsing, web searching and music recommendation. Although what makes part of a song memorable or distinguishable is still unclear, previous research typically assumes it to be the most repeated section. Some research on music thumbnailing deals

with symbolic musical data (e.g., MIDI files and scores) [9]. There have also been studies on thumbnailing of musical signals. Logan and Chu [11] attempted to use a clustering technique or Hidden Markov Models to find key phrases of songs. Mel Cepstral features were used to characterize each song. Bartsch and Wakefield [1] used the similarity matrix proposed by Foote [6][7] and chroma-based features for music thumbnailing. A variation of the similarity matrix was also proposed for music thumbnailing [12].

This paper presents an approach of automatically generating thumbnails of musical signals via structural analysis, which identifies the recurrent structure of musical pieces from acoustic signals. Specifically, an improved version of the algorithm based on [3][4] will output structural information, including both the form (e.g., AABABA) and the boundaries indicating the beginning and the end of each section. It is assumed that no prior knowledge about musical forms or the length of each section is provided, and the restatements of a section may have variations. Once the recurrent structure is obtained, various strategies can be employed to generate the thumbnails of music based on particular assumptions for different applications.

Instead of being evaluated subjectively, the generated thumbnails are evaluated by several criteria, mainly based on previous human experiments on music thumbnailing and the properties of thumbnails used for commercial web sites. Furthermore, comparing to previous research, three novel structural similarity metrics are also proposed in this paper to quantitatively evaluate the performance of the structural analysis algorithm, in addition to the qualitative evaluation presented by figures.

The remainder of this paper is organized as follows. Section 2 presents the structural analysis approach. Section 3 illustrates the strategies for music thumbnailing using the results of structural analysis. Section 4 presents the evaluation methods and experimental results. Section 5 gives conclusions and proposes future work.

2. STRUCTURAL ANALYSIS

There has been some recent research on music structural analysis. Dannenberg and Hu [5] presented a method to automatically detect the recurrent structure of musical signals. Although the promise of the method was demonstrated in several examples, there was no quantitative evaluation of the method in their paper. On the other hand, previous research on music thumbnailing also shed light on structural analysis, although typically the methods for music thumbnailing aim at a simpler task of finding the most repeated phrases of a piece instead of finding the recurrent structure of the whole piece.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'03, November 2-8, 2003, Berkeley, California, USA.

Copyright 2003 ACM 1-58113-722-2/03/0011...\$5.00.

Chai [3][4] illustrates a structural analysis method, which follows five steps:

- 1) **Feature extraction:** Segment the signal into frames and compute the feature of each frame;
- 2) **Pattern matching:** Segment the feature vector sequence into overlapped segments of fixed length and compute the recurrent property of each segment using dynamic programming;
- 3) **Repetition detection:** Detect the repetitions of each segment by finding the local minima in the dynamic programming result;
- 4) **Segment merging:** Merge consecutive segments that have the same recurrent property into sections and generate pairs of similar sections.
- 5) **Structure labeling:** Segment and label the recurrent structure.

In this paper, several improvements have been made based on this algorithm.

- 1) In the feature extraction step, one more representation, the chroma representation, is investigated besides the pitch representation and the spectral representation investigated in [3][4]. The chroma representation combines octave-related frequency components to construct a twelve-dimensional feature vector for each frame [1][5].
- 2) In the repetition detection step, one more rule is used - if two consecutive minima are so close that the matching parts corresponding to the two minima overlap, only the deeper minimum is kept.
- 3) In the structure labeling step, the strategy for dealing with conflicts (meaning a later labeled section has an overlap with previously labeled sections) is revised to make the result more robust. The rule is that the previous labeled sections will always remain intact and the current section will be truncated. Only the longest truncated unoverlapped part, if it is long enough, will be labeled as a new section. The shifted version of the section will be truncated accordingly, even if there is no conflict, to resemble the structure of its original version.

All the parameter configurations were tuned empirically based on the experimental corpus, which is described in Section 4. Please also note, pattern matching is the most time consuming step in the algorithm. Its time complexity is $O(n^2)$ (n is the number of frames), which is the same as methods based on the similarity matrix [1][6][7][12].

3. MUSIC THUMBNAILING VIA STRUCTURAL ANALYSIS

The problem of music thumbnailing aims at finding the most representative part of a song. It would be helpful if the song has been segmented into semantically meaningful sections before summarization, because, although what makes a part of a song most memorable is not clear, this part often appears at particular locations within the structure, e.g., the beginning or the end part of each section. For example, among the 26 Beatles' songs in our

experiment (see Section 4), 6 songs have the song titles in the first sentence of a section; 9 songs have them in the last sentence of a section; and 10 songs have them in both the first and the last sentences of a section. For many pop/rock songs, titles are contained in the "hook" sentences. This information is very useful for music thumbnailing: once we have the structure of a song, we can have different strategies for music thumbnailing, for different applications. Please note that the thumbnailing result using either of the following strategies highly depends on the accuracy of the structural analysis results, though some methods need high accuracy of generated section boundaries while others do not.

3.1 Section-beginning Strategy

The first strategy assumes that the most repeated part of the music is also the most representative part and the beginning of a section is typically essential. Thus, this strategy chooses the beginning the most repeated section as the thumbnail of the music. The algorithm first finds the most repeated sections based on the structural analysis result, takes the first section among these and truncates the beginning (20 seconds in our experiment) of it as the thumbnail. This strategy requires both high boundary accuracy and high structural accuracy.

3.2 Section-transition Strategy

We have also investigated the music thumbnails at Amazon (assumed to be marked by human experts) and found that the transition parts (end of section A and beginning of section B) tend to be used as the thumbnails. The assumption is that the transition part can give a good overview of both sections and is more likely to capture the hook of the song, though it typically will not give a thumbnail right at the beginning of a sentence. Based on the structural analysis result, the algorithm finds the transition from section A to section B such that the sum of the repeated times of A and those of B is maximized; and then it truncates the end of section A (of half the length limitation; 10 seconds in our experiment), the bridge and the beginning of section B. The boundary accuracy is not very important for this strategy.

3.3 Multiple-phrase Strategy

We can also combine the above two strategies and choose two segments (instead of one continuous segment) coming from the beginnings of the most and the second most repeated sections. Each segment is of half the total length limitation. This strategy needs even higher accuracy of both structure and boundary.

4. EXPERIMENT AND EVALUATION

This section presents the experimental results and evaluations of both structural analysis and thumbnailing.

4.1 Data Set

The experimental corpus consists of the 26 Beatles' songs in the two CDs *The Beatles* (1962-1966). All these songs have clear recurrent structures and leading vocals. The data were mixed to 8-bit mono channel and downsampled to 11kHz.

4.2 Metrics of Structural Similarity

To qualitatively evaluate the results, figures as shown in Figure 1 are used to compare the structure obtained from the algorithm to the ideal structure obtained by manually labeling the recurrent structure. This paper also proposes three metrics of structural

similarity to quantitatively evaluate the result. The values of all the metrics need to be as small as possible, ideally equal to zero.



Figure 1: Comparison of the computed structure using the FFT representation (above) and the ideal structure (below) of *Eight Days a Week*. Sections in the same color indicate restatements of the section. Sections in the lightest grey correspond to the sections with no repetition.

Metric 1 (structural inaccuracy, TI) is defined as the edit distance between the strings representing different forms. For the example in Figure 1, the distance between the ideal structure AABABA and the computed structure AABAABA is 1, indicating one insertion. Here how the algorithm labels each section is not important as long as the recurrent relation is the same; thus, this ideal structure is deemed as equivalent (0-distance) to structure BBABAB, or structure AACACA.

Metric 2 (labeling inaccuracy, LI) is mainly used to evaluate the consistency of labels of the ideal structure and those of the computed structure. It is defined as

$$LI = (1 - r) / s \quad (1)$$

where r is the ratio of the length of parts where both structures have the same labeling to the whole length, and s is the number of the recurrent sections in the ideal structure.

Metric 3 (segmentation inaccuracy, SI) is mainly used to evaluate how accurate the boundaries of each section are. It is defined as

$$SI = 1 - \frac{3 \cdot prc \cdot rcl}{prc + rcl + 1} \quad (2)$$

where

$$prc = \frac{\# \text{correctly found boundaries}}{\# \text{computed boundaries}} \quad (3)$$

$$rcl = \frac{\# \text{correctly found boundaries}}{\# \text{ideal boundaries}} \quad (4)$$

Equations (2), (3) and (4) are similar to the segmentation metrics proposed in [10], while the denominator is guaranteed not equal to 0. A computed boundary t_1 is regarded as correct if and only if there exists an ideal boundary t_2 such that $|t_1 - t_2| < \Delta t$, where Δt is the tolerance interval and empirically set to be 40 frames (about 1.8s) in our experiment.

Finally, we can combine all the above three metrics to evaluate a structural analysis result:

$$E = (TI + 1) \cdot LI \cdot SI \quad (5)$$

4.3 Structural Analysis Results

Figure 2 shows the structural analysis results based on equation 5 using all the three representations. The performance of each song can be easily seen from this figure. For example, the fifth song using the pitch representation, the tenth song using the FFT

representation, and the sixteenth song using the chroma representation all have good performances (low values of E), whose computed structures are shown in Figure 3, 1 and 4. Figure 2 also shows that not every song in the corpus has good performance. For example, the seventeenth song's performance is bad for all the three representations.

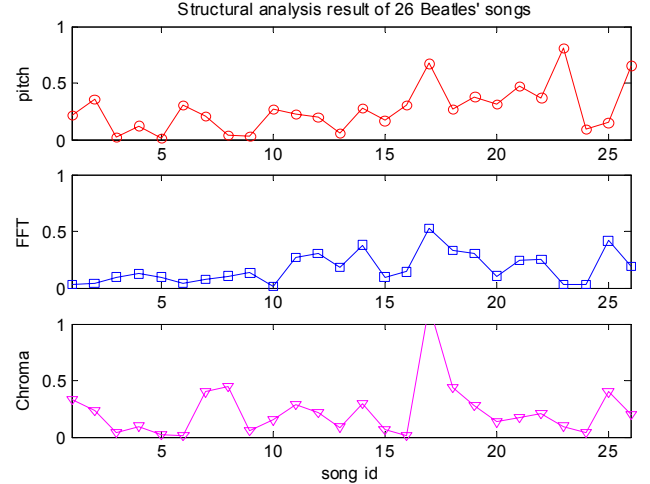


Figure 2: Structural analysis result of the 26 Beatles' songs (top: pitch representation; middle: FFT representation; bottom: chroma representation).



Figure 3: Comparison of the computed structure using the pitch representation (above) and the ideal structure (below) of *I want to hold your hand*.



Figure 4: Comparison of the computed structure using the FFT representation (above) and the ideal structure (below) of *We can work it out*.

4.4 Thumbnailing Results

In our experiment, twenty-second thumbnails were generated using all the three strategies and the three representations. To evaluate these thumbnails, four criteria are considered based on the human experiments presented by Logan and Chu [11]. These criteria include:

- 1) The percentage of generated thumbnails that contain a *vocal* portion;
- 2) The percentage of generated thumbnails that contain the song's *title*;
- 3) The percentage of generated thumbnails that start at the beginning of a section (*B. S.*);
- 4) The percentage of generated thumbnails that start at the beginning of a phrase (*B. P.*).

In addition, for the section-transition strategy, we also count the percentage of generated thumbnails that contain a transition of two different sections.

Table 1, Table 2 and Table 3 show the performances of all the three strategies. In Table 2, only the 22 songs in our corpus that have different recurrent sections were counted for the fifth column. In Table 3, the first, third and fourth columns were counted using each 10-second thumbnail.

Table 1: Thumbnailing results using Section-beginning strategy (20 seconds for each song).

	Vocal	Title	B. S.	B. P.
Pitch	100%	69%	54%	65%
FFT	100%	58%	69%	73%
Chroma	100%	58%	65%	65%

Table 2: Thumbnailing results using Section-transition strategy (20 seconds for each song).

	Vocal	Title	B. S.	B. P.	Transition
Pitch	100%	77%	27%	42%	64%
FFT	100%	77%	35%	50%	73%
Chroma	100%	77%	27%	38%	68%

Table 3: Thumbnailing results using Multiple-phrase strategy (20 seconds for each song).

	Vocal	Title	B. S.	B. P.
Pitch	96%	77%	36%	48%
FFT	96%	73%	39%	58%
Chroma	96%	73%	40%	49%

Comparing the results of the three thumbnailing strategies, it is clearly shown that, using the section-transition strategy, the percentage of thumbnails starting from the beginning of a section or a phrase significantly decreases, while it is more likely to contain titles in the thumbnails, which means thumbnailing using this strategy might capture the “hook” of music better. Even the multiple-phrase strategy can improve the probability of catching the title comparing to the section-beginning strategy; however, since this strategy requires very high structural and boundary accuracies, its overall performance is not as good as the other two. It is possible though that it can achieve good performance if we can improve our structural analysis accuracy in the future.

5. CONCLUSIONS AND FUTURE WORK

This paper presents three strategies for music thumbnailing based on the structural analysis result. Preliminary results were evaluated both qualitatively and quantitatively, which demonstrate the promise of the proposed method. To improve the accuracy, more representations need to be investigated. This will also help generalize the method to other music genres.

Although only three strategies for thumbnailing based on structural analysis are proposed in this paper, there can be other variations as well. Furthermore, we should choose appropriate strategies for different applications. For example, the section-beginning strategy might be good for indexing of query-based applications, because it is more likely that the user will query from the beginning of a section or a phrase. The section-transition strategy might be good for music recommendation, where it is more important to contain the “hook” sentence in the thumbnail. The multiple-phrase strategy would be good for long thumbnails.

Finally, the solution to music thumbnailing and structural analysis depends on the study of human perception of music, for example, what makes part of music sounds like a complete phrase and what makes it memorable or distinguishable. Human experiments are always necessary for exploring such questions.

6. REFERENCES

- [1] M.A. Bartsch and G.H. Wakefield, “To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing,” *In Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.
- [2] W.P. Birmingham, R.B. Dannenberg, G.H. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand, “MUSART: Music Retrieval via Aural Queries,” *In Proc. International Symposium on Music Information Retrieval*, Bloomington, IN, 2001.
- [3] W. Chai, “Structural Analysis of Musical Signals via Pattern Matching,” *In Proc. ICASSP*, 2003.
- [4] W. Chai and B.L. Vercoe, “Structural Analysis of Musical Signals for Indexing and Thumbnailing,” *In Proc. Joint Conference on Digital Libraries*, 2003.
- [5] R.B. Dannenberg and N. Hu, “Pattern Discovery Techniques for Music Audio,” *In Proc. International Conference on Music Information Retrieval*, October 2002.
- [6] J. Foote, “Visualizing Music and Audio using Self-Similarity,” *In Proc. ACM Multimedia Conference*, 1999.
- [7] J. Foote, “Automatic Audio Segmentation using a Measure of Audio Novelty,” *In Proc. of IEEE International Conference on Multimedia and Expo*, 2000.
- [8] J. Foote, “ARTHUR: Retrieving Orchestral Music by Long-Term Structure,” *In Proc. International Symposium on Music Information Retrieval*, October 2000.
- [9] J.L. Hsu, C.C. Liu, and L.P. Chen, “Discovering Nontrivial Repeating Patterns in Music Data,” *IEEE Transactions on Multimedia*, Vol. 3, No. 3, pp. 311-325, September 2001.
- [10] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, “Strategies for Automatic Segmentation Audio Data,” *In Proc. ICASSP*, 2000.
- [11] B. Logan and S. Chu, “Music Summarization using Key Phrases,” *In Proc. ICASSP*, 2000.
- [12] G. Peeters, A. L. Burthe and X. Rodet, “Toward Automatic Music Audio Summary Generation from Signal Analysis,” *In Proc. International Conference on Music Information Retrieval*, October 2002.