

SEMANTIC SEGMENTATION OF MUSIC AUDIO CONTENTS

Bee Suan Ong & Perfecto Herrera
Institut Universitari de l'Audiovisual,
Universitat Pompeu Fabra
Ocata 1-3, 08003 Barcelona, Spain
{beesuan, pherrera}@iua.upf.es
<http://www.iua.upf.es>

ABSTRACT

This paper proposes a novel approach to detect structural changes in music audio signals and provides a way to separate the different “sections” of a piece, i.e. “intro”, “verse” or “chorus”. Herein, we divide the segment boundaries detection task into a two-phase process with each having different functionalities. In order to obtain appropriate structural boundaries, we propose a combination of low-level descriptors to be extracted from music audio signals. A database of 54 audio files (The Beatles’ songs) is used for evaluation on a mainstream popular music collection. The experiment results show that our approach has achieved 71% of accuracy and 79% of reliability in identifying structural boundaries in music audio signals. These measures indicate that the performance of our method improves the results reported in the still scarce literature that includes quantitative analyses.

1. INTRODUCTION

Music structure varies widely from composer to composer and from piece to piece. Transformation, repetition, elaboration and simplification of music materials help to create the unique identity of music. Hence, it is believed that structural description provides a powerful way of interacting with audio content (i.e. browsing, summarizing, retrieving and identifying). Seeing this uniqueness of music structure, it is interesting to ask the question: is it possible to detect non-trivial/significant structural changes (i.e. intro->verse, verse->chorus, chorus->bridge, etc.) in music audio signals? This paper presents a novel, two-phased approach to this problem based on audio content analysis and similarity computation. In order to obtain appropriate musical content descriptions to detect structural changes, we propose a combination set of low-level descriptors to be extracted from music audio signals. In this paper, we address the problem of finding acceptable structural boundaries, without prior knowledge about musical structure. There is a second related problem consisting on assigning labels to the found segments. This will be reported in the first coming publication.

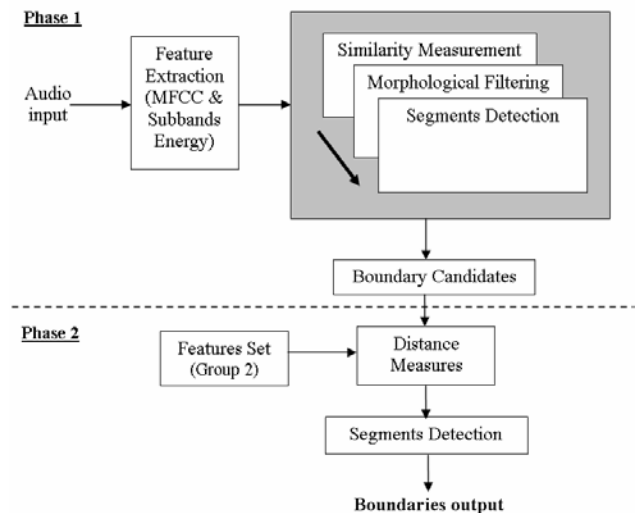


Figure 1. Overview framework of our approach.

This paper is organized as follows: Section 2 presents our method for detecting structure changes in music audio. Section 3 shows and discusses the experimental results of our approach. Finally, Section 4 gives concluding remarks and future research plans.

2. APPROACH

Audio segmentation facilitates partitioning audio streams into short regions. It seems an indispensable process in certain content-based applications, such as audio notation, audio summarization, audio content analysis, etc. Because of this reason, research in this area has receiving an increasing attention in recent years. A number of different approaches have been proposed [1][3][6][7][8].

In this paper, we propose a novel approach for detecting structural changes in audio signals. We first divide the process in two phases (see diagram in figure 1). Each phase is given a different task: Phase 1 focuses in detecting boundaries, which may contain structural changes from the audio signal; Phase 2 focuses in refining detected boundaries obtained from phase 1 by aggregating contiguous segments while keeping those which really mark structural changes in music audio. Our proposed method consists of 9 steps as follows:

Phase 1

- (1) Segment the input signal into overlapped frames of fixed length and compute audio descriptors for each frame;
- (2) Compute between-frames cosine distance to obtain several similarity matrices [3] for each one of the used features (see section 2.1);
- (3) Apply a morphological filter (see section 2.2) to similarity matrices for enhancing the intelligibility of the visualization;
- (4) Compute novelty measures by applying kernel correlation [3] along the diagonal of the post-processed similarity matrices;
- (5) Detect segments by finding the first 40 highest local maxima from novelty measure plot;
- (6) Combine the detected peaks to yield boundary candidates of segment changes of music audio;

Phase 2

- (7) Assign frames according to detected segments obtained from phase 1 and compute the average for all the used features (see table 2) in each segment;
- (8) Compute between-segments' distances using the mean value of each features in each segment;
- (9) Select significant segments based on a distance measure.

The following sections explain each step in detail.

2.1. Feature Extraction

We use a combination of low-level descriptors extracted from music audio signals [4][9]. The algorithm first segments input signal into overlapped frames (4096-sample window length) with the hop size of 512 samples, then followed by extracting feature descriptions for each of these frames. The proposed features are:

MFCC, also called Mel-Frequency Cepstral Coefficients, a compact representation of an audio spectrum that takes into account the non-linear human perceptual of pitch, as described by the Mel scale.

Spectral Centroid: A representation of the balancing point of the spectral power distribution within a frame.

Sub-bands energy: A measure of power spectrum in each sub-band. We divide the power spectrum into 9 non-overlapping frequency bands as described in [5].

Zero Crossings: A time-domain measure that gives an approximation of the signal's noisiness.

Spectral Rolloff: A measure of frequency, which is below 95 percentile of the power spectral distribution. It is a measure of the "skewness" of the spectral.

RMS energy: A measure of loudness of the sound frame.

Phase 1	Phase 2
MFCC Sub-bands Energy	Zero Crossings rate, Spectral Centroid, Spectral Flatness, Spectral Rolloff, Spectral Flux, RMS, Low Bass Energy, High-medium Energy

Table 2. The list of audio descriptors for Phase 1 and Phase 2.

Spectral Flux: The 2-norm of the frame-to-frame spectral magnitude difference vector. It measures spectral difference, thus it characterizes the shape changes of the spectrum.

Spectral Flatness: A measure of the flatness properties of spectrum within a number frequency bands. High deviation from a flat shape might indicate the presence of tonal component.

High-medium energy: A ratio of spectrum content within the frequency range of 1.6 kHz and 4 kHz to the total content. This frequency range comprises all the important harmonics, especially for sung music.

Low Bass energy: A ratio of low frequency component (up to 90 Hz) to the total spectrum content. This frequency range includes the greatest perceptible changes in "bass response".

The computed descriptors are grouped into two sets to be used in different phases of segment detection process. Table 2 above shows the grouping of the audio descriptors.

2.2. Phase 1 - Rough Segmentation

After computing feature vectors for each frame, we group every 10 frames (116ms) and calculate the mean value for every feature. In this phase of segment detection process, we only work with MFCC and subband energies. We treat those features in separately in order to combine both results in the final stage of detection process in phase 1. In order to find the structural changes in the audio data, we measure the distance between each feature vectors and their neighbouring vectors using cosine angle distance [3].

To improve the intelligibility of the segment information in the distance representations, we exploit one of the most widely used filtering techniques in image processing field to post-process the computed similarity matrix. We apply a morphological filter [2] to the similarity matrix to get rid of low value points while keeping the rest of the matrix intact. This is to facilitate the enhancement of the segment boundaries. Following

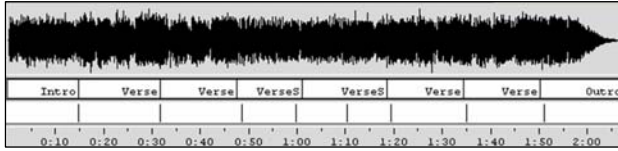


Figure 2. Manually labelled segment boundaries (top) and detected segment boundaries by our proposed algorithm (middle) with time position information (below) for SongID-35 entitled “Words of Love”. The label VerseS means an instrumental solo playing the verse. Labels are not yet assigned by the algorithm.

Footo’s approaches [3], we then apply a kernel correlation, with the width of 10, along the diagonal of the post-processed similarity matrix to measure the audio novelty. Based on the novelty measure, the first 40 highest local maxima are selected for each individual features. We then aggregate all selected local maxima to yield candidates of segment boundaries for further processing.

2.3. Phase 2 - Segment Boundaries Refinement

In the second phase of the detection process, we recombine (join) some of the previously extracted segments into larger units. Here, we consider the within-segment values for all the attributes used in phase 2 (see table 2) and compute the segment averages for each of them. Hence, each detected segment now comprises only a set of feature vectors representing the mean value of the attribute in that segment. It has to be mentioned that our used attributes present a large range of values. Apparently, attributes whose values are larger than the other would have more influence in determining the similarity of any two sequences. Hence, in order to avoid such effect and to have an equal importance weight among the used attributes, we normalize all attributes so that its feature values are within the range of 0 and 1. We then compute (dis)similarity between each segment and its neighboring segments by measuring the Euclidean distance between their feature vectors. Similarly to the previous steps in computing novelty measures from the similarity representations, we apply a kernel correlation, along the diagonal of (dis)similarity representation of segments to yield the novelty measures, N , between each segment and its next sequential segment. Finally we select the significant segment boundaries from the computed novelty measures, $N = \{n_s | s = 1, 2, \dots, l\}$ (where l is the number of segment boundaries candidates) based on the following steps:

1. Select all the peaks that lie above a predefined threshold, P_t , based on their computed novelty measures, N_s , and organize them into a group, which is represented as $P = \{p_i | i = 1, 2, \dots, M\}$ (M is the number of selected peaks). Whereas those peaks that lie below the predefined threshold, P_t , are organized into another group denoted by

$$E = \{e_j | j = 1, 2, \dots, N\} \quad (N \text{ is the number of unselected peaks}).$$

2. Organize all peaks in E in ascending order according to their distance measures.
3. Select the highest peak in E for further evaluation.
4. Based on temporal information, if the evaluated peak is located at least 4 sec apart from any peaks in P , insert it in group P and reorganize all peaks in group P in ascending order based on the segment index number; otherwise delete it from E . This is based on the assumption that each section in music (ex. verse, chorus, etc.) should at least hold 4 sec (1 bar for songs with quadruple meter with 60 bpm tempo) in length before moving to the next section.
5. Go to step 3.

The whole iterative peak selection process ends when there is no more peak in E . Finally, segment boundaries in P are considered as significant segment boundaries that mark structural changes in music audio signals.

3. RESULTS

3.1. Data Set

In our experiment, we use the 54 songs from first four CD’s of The Beatles (1962 – 1965) as a test set. Each song is sampled at 44.1 kHz, 16-bit mono. For algorithm evaluation, we have generated a ground truth by manually labelling all the sections (i.e. intro, verse, chorus, bridge, verse, outro, etc.) of all the songs in the test set, according to the information provided by Allan W. Pollack’s “Notes On” Series website on song analyses of Beatles’ twelve recording project¹. A music composer supervised the labelling process and results.

3.2. Recall and Precision Measures

To quantitatively evaluate the detected segments from the proposed algorithm against the ground truth, we calculate the precision and recall and F-measure measures of the test. In evaluating the identified segment, we allow a tolerance deviation of ± 3 seconds (approximately 1 bar for a song of quadruple meter with 80 bpm in tempo) from the manually labelled boundaries. Precision and recall are mainly used to evaluate the accuracy and reliability of the proposed algorithm, whereas F-measure is mainly used to measure overall effectiveness of detection by combining recall and precision with an equal weight.

3.3. Experimental Results

Our proposed structural changes detection approach has achieved accuracy higher than 71% and a reliability of 79% using the ground truth set. The overall F-measure

¹ The Twelve Recording Projects of the Beatles webpage: http://www.icce.rug.nl/~soundscapes/DATABASES/AWP/awp-beatles_projects.html

reached 75%. In another words, with 10 detected segment boundaries, 7 of them are correctly detected compared to ground truth data. Whilst about 2 out of 10 manually labelled boundaries are missed by our automatic boundaries detector. The distribution of precision scores has a standard deviation of 0.11 and the range of precision values spans across 0.41-0.94. From our results, we observe that the best performance in the case of SongID-35 with its recall and precision score of 100% and 94% respectively. Whereas the worst performance is observed in the case of SongID-47, which only reaches the recall rate of 38% and precision rate of 56%. Figure 2 illustrates the detected segment boundaries by our proposed algorithm with manually labelled segment boundaries for SongID-35. It is worthwhile to pay attention to the fact that precision and recall rate are particularly low for those songs which comprise of smooth transition between sections. It seems that our descriptors are not sensitive enough to mark these changes. On the other hand, songs with coarse transition between sections usually achieve a better rate on these measures. Using some other disregarded descriptors perhaps may be able to cope with this matter.

With our test data set, we have compared our approach with previous method described in [9]. It was reported that with an allowed tolerance deviation of 3.7 sec (higher than ours), the author reported less than 60% for both recall and precision measures respectively whereas we achieved over 70% for these both measures. However, it should be noted that the generality of our music database is quite limited. So far, we have not yet tested our approach on different music genres (i.e., instrumental music, techno or jazz).

4. SUMMARY

This paper presented a new approach for detecting structural changes in music audio using a two-phased procedure and different descriptors for each phase. A combination set of audio descriptors has also been shown useful in detecting music structure changes. Evaluation results have shown the validity and the performance of our proposed approach. For ongoing research to further improve the detection algorithm, more attention will be given to the following factors:

- Integration of some previously disregarded lower-level feature attributes (i.e. Pitch class profile, etc.);
- Making use of higher-level analysis techniques (i.e. beat detection, phrase detection, etc.) to achieve better segment truncation.
- Automatic labelling of sections according to their structural title (i.e. intro, verse, bridge, etc.).
- Testing the performance using an annotated database comprising different music genres different from "60's pop music".

5. ACKNOWLEDGEMENTS

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents; <http://www.semanticaudio.org/>) The authors would like to thank members of SIMAC and AUDIOCLAS projects at the Music Technology Group in the UPF for their useful comments and discussions.

6. REFERENCES

- [1] Aucouturier, J.-J., and Sandler, M. "Segmentation of Musical Signals Using Hidden Markov Models", *Proceedings of the Audio Engineering Society 110th Convention*, Amsterdam, Netherlands, May 2001.
- [2] Burgeth, B., Welk, M., Feddern, C., and Weickert, J. "Morphological operations on matrix-valued images", *The 8th European Conference on Computer Vision*, Prague, Czech, May 2004, pages 155-167.
- [3] Foote, J. "Automatic Audio Segmentation using a Measure of Audio Novelty", *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, USA, 2000, pages 452-455.
- [4] Foote, J. "Visualizing Music and Audio using Self-Similarity", *Proceedings of ACM Multimedia Conference*, Orlando, Finland, 1999, pages 77-80.
- [5] Maddage, N.C., Xu, C., Kankanhalli, M.S., and Shao, X. "Content-based Music Structure Analysis with the Applications to Music Semantic Understanding", *ACM Multimedia Conference*, 2004, New York, Oct 2004, pages 112-119.
- [6] Ong, B., and Herrera, P. "Computing Structural Descriptions of Music through the Identification of Representative Excerpts from Audio Files", *Proceedings of 25th International AES Conference London*, UK, June 2004.
- [7] Peeters, G., La Burthe, A., and Rodet, X.. "Toward Automatic Music Audio Summary Generation from Signal Analysis", *Proceedings of ISMIR 2002, 3rd International Conference on Music Information Retrieval*, Paris, Oct 2002, pages 94-100.
- [8] Tzanetakis, G., and Cook, P. "Multifeature Audio Segmentation for Browsing and Annotation", *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct 1999, pages 103-106.
- [9] Wei, C. *Structural Analysis of Musical Signals for Indexing, Segmentation and Thumbnailing*. Paper for the Major Area of the PhD General Exam, March 2003.