Structural boundary perception in popular music

Michael J. Bruderer[1], **Martin McKinney**[2], **Armin Kohlrausch**[1,2]
[1]Technische Universiteit Eindhoven, Postbus 513, 5600 Eindhoven, The Netherlands
[2]Philips Research Laboratories, Prof. Holstlaan 4 (WO-02), 5656 AA Eindhoven, The Netherlands
m.j.bruderer@tm.tue.nl, {armin.kohlrausch, martin.mckinney}@philips.com

Abstract

The automatic extraction of musical structure from audio is an important aspect for many music information retrieval (MIR) systems. The criteria on which structural elements in music are defined in MIR systems is often not clearly stated but typically stem from (music) theoretical or signal-based properties. In many cases, however, perceptual-based criteria are the most relevant and systems need to be trained on or modeled after the perception of structural elements in music. Here, we investigate the perception of structural boundaries to Western popular music and examine the musical cues responsible for their perception. We make links to music theoretical descriptions of structural boundaries and to computational methods for extracting structure. The methods and data presented here are useful for developing and training systems for the automatic extraction of musical structure as it is perceived by listeners.

Keywords: Music cognition, music structure, music perception, music segmentation.

1. Introduction

Automatic music structure analysis is an important component of music information retrieval. The ability to automatically identify or extract various structural elements from musical audio would be a boon for automatic metadata generators and would benefit music consumers, retailers and libraries. Several recent studies, employing different algorithms, have shown that one can extract elements of structure from music audio with modest success. The algorithms rely on signal based features, such as MFCC, chroma, linear prediction coefficients or spectral-based contrasts [1, 2]. These features are used in various ways: as input to neural nets to predict the likelihood of a musical boundary [2]; as the basis for a similarity matrix from which repeated patterns are identified [3, 4]; or as input to algorithms for "saliency" to identify the most representative excerpt of a musical piece [5]. A more recent study employs a hierarchical approach in an effort to incorporate musical knowledge into the system [6]. This system first analyzes the rhythm and uses it to extract music-theoretical features including chord information, singing voice detection, and song structural elements. The structural elements are based on repetition as in the other studies, where repetition here refers to melody or chords. While all of these algorithms can, to some degree, extract or identify elements of the musical structure, it is unclear how well those elements relate to *perceived* elements of structure or to those defined clearly by rules of music theory.

In developing algorithms for the extraction of musical structural boundaries, defined as perceptual points in time between consecutive segments, we can turn to music theoretical studies on structure to learn which cues are important. Lerdahl and Jackendoff [7], in their seminal book on structure in Western classical music, propose, among others, the following cues for segmenting monophonic melodies: Pauses; longer notes in between short notes; changes in register, dynamics, articulation, and length; and repetition (parallelism). Using these cues they propose rules on how the experienced listener segments music. If several cues occur together at the same time, they are added and a stronger boundary is perceived. One missing element of this model is that there is no quantification of the salience of the different cues. It is also unclear how well their rules relate to the *perception* of structure in music.

Deliège [8] tested some of the rules from Lerdahl and Jackendoff for their perceptual relevance and in doing so gave an order of importance of the different rules. She found that the most important rules were register change, attack-point (a long note in between two short notes), and rest. She also proposed an additional important rule: timbre change.

A recent study by Frankland and Cohen [9] quantified some of the rules of Lerdahl and Jackendoff [7] and tested the quantified rules for their perceptual validity. They found that only two of these rules were used by subjects for segmenting monophonic melodies: attack-point and rest.

Another quantitative model based on music theory uses the change in intervals to segment monophonic melodies [10]. An interval is defined as a difference in pitch, in intensity, and in the IOI (inter-onset-intervals). Preliminary results on four pieces show reasonable performance but this model still needs further testing.

All the models above were mainly conceived and tested on monophonic Western classical music. A more prevalent form of music, however, is Western popular music. This study extends the perceptual validation of the models for popular music.

For many applications it is desirable to have a system that can automatically segment music similar to the way in which humans do. While the models based on musicology described above can serve as a starting point, a model for the perception of music structure or a set of perceptually-based ground-truth data is required. We present here a method for investigating the perception of structural boundaries in music and for assigning perceptual relevance to various segmenting cues based on musicology. We show results and analysis for six songs from Western popular music. This method and data can be used as a training ground for systems to automatically extract structural boundaries from music.

2. Method

2.1. Material

From a pool of twenty songs we chose six to cover a range of popular music styles and to have time-distributed boundary cues based on musicology. We used MIDI files corresponding to the audio tracks and the models of Cambouropoulos [10] and Frankland and Cohen [9] to analyze the salience and timing of various segmentation cues. We also marked three other cues: the ending of harmonic cycles and the introduction and ending of instrumental voices. If the various cues all occurred at the same time (high temporal correlation), it would be difficult to assign relative perceptual salience to specific cues. Thus, if a song had low temporal correlation between these different cues we deemed it suitable as a candidate for the experiment.

The six chosen songs were: the song “Heart to hurt” by Kousuke Morimoto taken from the RWC database [11], “Moondance” by Van Morrison, “Live and let die” by Paul McCartney, “And when I die” by Blood Sweat and Tears, “Body and soul” performed by Billy Holiday (vocal) and “Body and soul” by Coleman Hawkins (instrumental). The songs had a duration of between 3 and 5 minutes.

2.2. Procedure

The experiment consisted of two parts; first subjects listened to music and pressed a key whenever they encountered a phrase or segment boundary; second, subjects were asked to rate the salience of a selected number of boundaries.

In the first part subjects listened to the song four times, the first time to familiarize, and three times to record the key presses, without any symbolic representation of the song. The data from the first part were analyzed and the boundaries where 90% of the subjects tapped within a time-window of 2.4-s (empirically calculated as being the optimal windowsize) were taken for the second part. Additionally, two to three medium and weak boundaries per song were selected close to 80% and 40% agreement respectively, yielding 98 boundaries in total (14-21 per song).

In the second part of the experiment subjects rated the salience of the selected boundaries on a scale from 0 to 6. Subjects were presented a horizontal line representing the timespan of the song. Vertical lines indicated the boundaries and a moving cursor indicated the momentary playing position. Subjects also gave a free description of the cues – what in the music contributed to the perception of the boundary. In this part of the experiment, subjects could listen to the whole song as well as parts of it as many times as they liked. The order of the songs was randomized for both parts of the experiment.

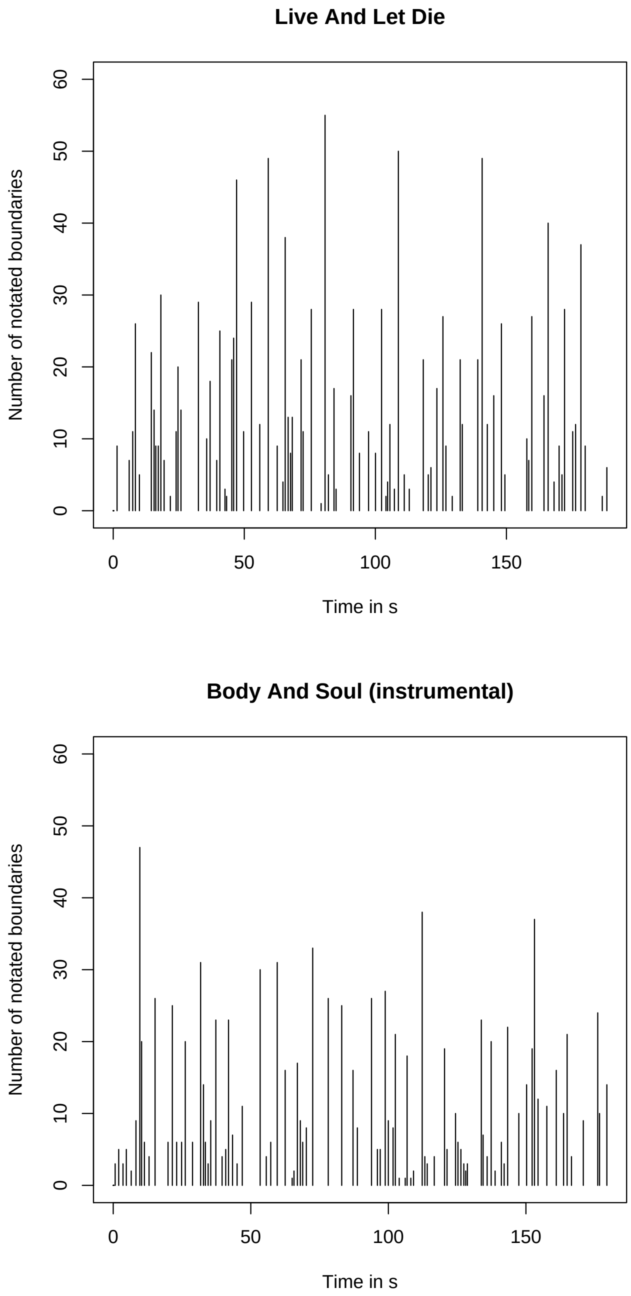
2.3. Subjects

Eighteen subjects (14 male, 4 female) participated in the first part of the experiment. From these, fifteen also participated in the second part. Musical experience differed widely among subjects, with a mean of 6.8 years (SD of 7.0) for the first part and a mean of 6.7 years (SD of 7.3) for the second part, and ranging from 0 to 21 years of musical training. The average subject age was 26.5 years, ranging from 21 to 37 years.

3. Results and Analysis

All the notated boundaries were collapsed into one vector and quantized to one millisecond resolution. In order to estimate a density function of notated boundaries, the quantized vector was convolved with a Gaussian window. The peaks in the convolved vector indicated a boundary detected by several subjects. At each peak all notated boundaries within a 2.4-s window were summed and this sum was taken to represent the salience rating of the boundary, here called the number of notated boundaries. A few times subjects pressed a key more then once within a 2.4-s window and thus some boundaries exceeded the theoretical maximum of 54. Two examples of the distribution of the notated boundaries within a 2.4-s window are shown in Figure 1. The distribution for “Live and Let Die” (top panel) shows several boundaries indicated by almost all subjects, as well as many boundaries indicated by fewer subjects. The distribution for “Body and Soul” shows much less agreement across subjects, with only a single boundary getting more than 40 indications.

Figure 2 shows a result from the second part of the experiment: the relation between the salience rating and the number of indicated boundaries. The ratings of the boundaries given by the subjects in the second part of the experiment were significantly correlated with the number of subjects that pressed a key at the boundary within a 2.4-s window ($R = 0.88$, $p < 0.001$). In a previous study with a similar method [12] it had been assumed that the number of notated boundaries was an indication of the strength of a boundary but it had never been shown. It is possible that the salience rating of a boundary for a given subject is not related to the number of subjects that perceive the boundary at a particular


Live And Let Die

Body And Soul (instrumental)

Figure 1. Two examples of the distribution of the notated boundaries within 2.4-s windows. The figure above shows a high consistency over the strongest boundaries, the figure below shows less consistency.

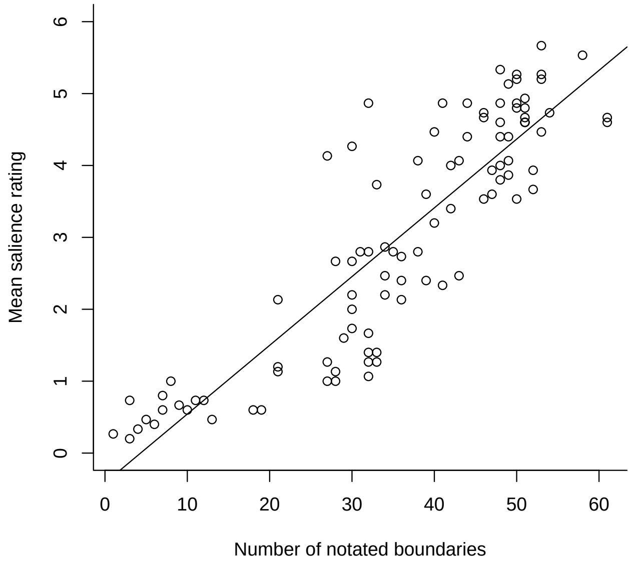


Figure 2. The correlation between the number of notated boundaries and the respective rating of the boundary. The line represents the linear regression. The correlation between the two is 0.884 (p $<$ 0.001).

point in time. Also, some boundaries may be perceptually more diffuse over time than others. However, here we found a very high correlation between the number of subjects that indicate a boundary within 2.4-s window and the respective salience rating.

In order to see which cues contributed to the subjects' perception of boundaries, we analyzed their descriptions of contributing cues. The descriptions were classified into twelve different classes: level change, change in timbre (drums, voice, other), tonality (harmonic progression, melody change), rhythm (a change in the strength of the rhythm, tempo change, rhythm change), and structural descriptions (global structure, repetition, break). Using these classes the descriptions of each of the boundaries were then classified. The terms mentioned most often were 'global structure' and 'change in timbre: other', both being mentioned more than 300 times, followed by 'change in level', 'repetition', and 'break/pause', all three mentioned about 160 times.

We then computed the *mean-term-rating*, i.e., the mean salience value association with a specific term, to get an indication of the relative importance of each term type. Using this measure the strongest cues seem to be the change in the strength of the rhythm and change in drums (timbre). It must be mentioned, however, that these cues were only mentioned a few times (5 and 32 times), thus it is questionable if they are the overall most salient cues. When looking at the cues that were mentioned more then a hundred times the strongest cue is a change in rhythm.

The mean-term-rating is also dependent on the song. For the two songs “Live and Let Die” and “And When I Die”, change in tempo has a high mean-term-rating, but for the other songs this cue is less relevant.

The cue associated with the maximum mean-term-rating for each song is: change in strength of rhythm (6.0) for “Heart to hurt”; progression (5.8) for “Moondance”; repetition (5.4) for “Body and soul (vocal)”; repetition (5.10) for “And when I die”; drums (6.0) for “Body and Soul (instrumental)”; no clear strong mean-term-rating for “Live and Let die”.

4. Discussion

The experiment shows that there are structural boundaries in music that are perceived by almost all listeners. A novel finding of this study is the strong correlation between the number of times a boundary was indicated by subjects and the reported salience rating of that boundary. The high correlation shows that a ground-truth database of boundary salience can be generated in two manners: Either subjects segment a piece of music by pressing a key and then all notated boundaries within a certain time-window are summed up, or a set of boundaries are given to subjects with the task of rating the salience of each boundary.

When comparing the different cues with which subjects described a boundary, it can be seen that a change in timbre was often mentioned, which is in agreement with the study of Deliège [8]. Repetition and change in dynamics, rules used by Lerdahl and Jackendoff [7], are also often found in the description of the boundaries. Frankland and Cohen [9] found that an important cue is rest, which corresponds to our often-mentioned break/pause class. These results extend the findings of previous studies and show that there are cues that contribute to the perception of musical structure across music styles.

Algorithms that segment music are often binary in nature: a boundary either exists or does not. Our study shows that perceptually, there is a wide range of salience across different boundaries. Thus, the *perception* of boundaries is not binary. Algorithms for automatic segmentation that intend to extract perceptually relevant structural elements should account for this range of salience in structural boundaries.

5. Summary and Conclusions

We have described an experimental method for examining the perception of structural boundaries in music. We have shown experimental data on how subjects segment Western popular pieces and what cues they used in describing the segment boundaries. These results are important for music information retrieval because there exist few perceptual studies that explore the perception of structural boundaries, especially for popular music. For automatic segmentation algorithms, however, it is crucial to have a perceptually relevant ground truth, if the algorithm should segment the same way humans do.

Our analysis shows that for Popular music there are several cues that strongly contribute to the perception of structural boundaries in music: changes in timbre, changes in level, repetition, and breaks/pauses. These results, however, are based only on the subjects' description of the boundaries and it is possible that there are other contributing factors. Our next step is to examine correlations between music theoretical boundaries and strengths (based on analysis of corresponding MIDI data) and the perceptual data collected here.

References

[1] G. Peeters, A. L. Burthe, and X. Rodet, “Toward automatic music audio summary generation from signal analysis.” in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 2002.

[2] N. Hu and R. B. Dannenberg, “A bootstrap method for training an accurate audio segmenter.” in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005, pp. 223–229.

[3] R. B. Dannenberg and N. Hu, “Discovering musical structure in audio recordings,” in *Music and Artificial Intelligence, Second International Conference, ICMAI 2002*, ser. Lecture Notes in Computer Science, vol. 2445. Springer, Sept 2002, pp. 43–57.

[4] J. Foote, “Visualizing music and audio using self-similarity,” in *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. New York, NY, USA: ACM Press, 1999, pp. 77–80.

[5] L. Lu and H.-J. Zhang, “Automated extraction of music snippets,” in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 2003, pp. 140–147.

[6] N. C. Maddage, “Automatic structure detection for popular music,” *IEEE MultiMedia*, vol. 13, no. 1, pp. 65–77, 2006.

[7] F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music*. Cambridge, Mass.: MIT Press, 1983.

[8] I. Deliège, “Grouping conditions in listening to music,” *Music Perception*, vol. 4, no. 4, pp. 325–360, 1987.

[9] B. W. Frankland and A. J. Cohen, “Parsing of melody: Quantification and testing of the local grouping rules of “Lerdahl and Jackendoff's a Generative Theory of Tonal Music”,” *Music Perception*, vol. 21, no. 4, pp. 499–543, 2004.

[10] E. Cambouropoulos, “Towards a general computational theory of musical structure,” Ph.D. dissertation, University of Edinburgh, faculty of music and department of artificial intelligence, 1998.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical, and jazz music databases,” in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, October 2002, pp. 287–288.

[12] E. F. Clarke and C. L. Krumhansl, “Perceiving musical time,” *Music Perception*, vol. 7, no. 3, pp. 213–252, 1990.