Chapter 9

Investigating the Perception of Structural Boundaries in Steve Reich’s City Life

Nerdinga Letulė, Duncan L. Snape, and Marc R. Thompson

**Abstract**

We all grow up in an informal musical learning environment where the most common types of music conform to established rules with regard to structure. By the time we reach adulthood these rules are embedded in our cognitive processes when experiencing music. Many studies have investigated the perception of structure in these traditional musical styles, but what happens when the music we listen to does not conform to our conditioning? An experiment was conducted in order to investigate the perception of structural boundaries in Steve Reich’s ‘City Life’, a prominent piece of contemporary art music with non-traditional structure, amongst two groups of participants - those with formal musical training and those without. 20 participants, 10 with formal musical training and 10 with informal musical training, were asked to identify and rate the salience of structural boundaries in three short excerpts from ‘City Life’. Participant responses were compared to a musicological analysis of the score and a computational analysis of the audio for each excerpt. The study found that when structural boundaries were more salient, so participants’ responses exhibited more consensus and less variation. The participants with informal musical training showed more overall consensus of response on the temporal point and salience of the structural boundary, and were more likely to choose the temporal point identified by score analysis.

**Introduction**

Researchers have argued that the scale of our daily exposure to music represents an ‘informal learning environment’ (ILE), from which we develop both familiarity with music and individual preferences (Batt-Rawden & De Nora, 2005). When we consider this idea it is not to be wondered at that the differences between non-musicians (informal learners) and musicians (formal learners) have been found to be relatively small in a variety of tasks related to the perception of music (Bigand & Poulin-Charonnat, 2006). Each individual's informal training begins before birth (Granier-Deferre & Busnel, 2013) and there is evidence that 'there is an initial predisposition of the human brain for music processing that is triggered by the extensive exposition to musical stimuli in everyday life.' (Bigand & Poulin-Charonnat, 2006, p.120).

The perception of music relies on the segmentation of the auditory stream into smaller units, melodic phrases and motifs. On the surface this segmentation would seem to be quite subjective and arbitrary, and yet in the compositional styles of many musical eras and genres clear structural rules, that listeners, both informally and formally trained, can become familiar with, exist. The sonata and rondo forms from the art music of the late 18th and 19th centuries (Cadwallader & Gagne, 2010), and the variations around the structure of ABABCAB in popular song (Cooper & Foote, 2003), are some examples of these. It is reasonable, therefore, to suppose that, due to repeated exposure to music in the ILE that obeys these structural rules, any individual might become familiar with common structural forms from a relatively early age without formal training. If we seek to understand how the structural perception of music occurs in individuals it follows that we need to present participants with musical stimuli that circumvent their familiarity and require the full engagement of cognitive processes.

 Studies focusing on the investigation of structure in music have employed a variety of methods in analysing both stimuli and results. The most common of these come from two quite distinct areas of research - musicological analysis and music information retrieval (MIR). The musicological approach is based on the extraction and interpretation of musical features from score notation (Cadwallader & Gagne, 2010; Beach, 2012). Examples of this include Schenkerian analysis and the Generative Theory of Tonal Music (Lerdahl & Jackendorf, 1983). Computational systems perform music feature analysis on digital musical data, either audio or MIDI. One of the most commonly applied methods in the assessment of structure is that of self-similarity. The acoustic similarity between any two instants of an audio recording is displayed as a two-dimensional comparative figure (Foote, 1999). The output of the self-similarity matrix can be further processed by finding repeated patterns and points of novelty (Cooper & Foote, 2003).

**Research Aims**

 The following study aimed to explore the perception of structural boundaries amongst both informally and formally trained listeners, in a piece of music that did not conform to commonly-experienced structural rules. It also sought to compare the results of a musicological analysis to those of a computational analysis with regard to which better represented the results of listener perception. With these aims in place, an experiment was conducted to investigate the following questions:

* Is there consensus as to the salience and temporal point of the structural boundary?
* Do participants with and without formal musical training perceive structural boundaries differently?
* Which analysis method (musicological or computational) more closely approximates listener perception?

**Previous Studies**

 There have been a number of previous studies that have compared the perception of musical structure of informally trained (IT) and formally trained (FT) participants, although these labels are only applied in this study. Some of these have used stimuli from within the ILE, others from outside it.

 Starting with those studies that worked within the ILE, Deliège (1987) used material from recordings of Bach, Mozart, Beethoven and Stravinsky and found very little difference in segmentation patterns between IT and FT participants. The difference that was observed was that FT participants tended to place fewer boundaries and group items into longer segments. Krumhansl (1996) used Mozart’s Piano Sonata K. 282 and found that the groups showed broad consensus on structural changes. Bruderer, McKinney & Kohlrausch (2009) used a number of popular songs by Billie Holliday, Paul McCartney, and Van Morrison amongst others. They found that there was considerable variety in the number and location of boundaries identified between participants. This was significantly linked to the salience of changes in the stimuli, but their results showed no significant effect of formal training. The results of these studies are not unexpected, given that we presuppose both groups of participants to have had similar levels of exposure, through the ILE, to the underlying rules of structure governing the music used as stimuli.

 Deliège & Ahmadi (1990), on the other hand, presented their participants with music from outside the ILE, from post-tonal composer Luciano Berio. Nevertheless, they still found that there was general agreement between the responses of both groups and those of two experts. Addessi & Caterina asked their participants to segment the music of Bruno Maderna (2002) and Kurtág (2005), and in both cases found very little evidence of an influence of formal training on results. These results are, perhaps, a little more surprising. In a situation where participants are presented with material that does not follow expected rules we might expect a FT participant to have more preparation for the successful identification of a structural boundary than an IT participant. However, it can equally be argued that the extra preparation a FT participant receives might relate only to the repertoire of one instrument, style or era.

 Finally, with regard to the question of comparison of different analysis methods, we were able to find relatively few studies. The one considered most relevant to the current study was Leman et al. (2005), where manual (score) and acoustical (computed) structural cues were used to predict the musical affect of participants. Their results showed that manual cues were more successful than acoustical.

**Experimental Methodology**

 Particular consideration was given in the design of the following to both external and ecological validity. It was reasoned that in order to achieve the capture of those cognitive processes involved in the everyday perception of music, we should undertake to capture them under circumstances approximating everyday conditions. Data would be gathered in as naturalistic a setting as possible, following the recommendations of Brewer (2000). This reasoning had a considerable impact on the design of experimental instruments, the location and approach of data gathering, and choice of stimuli.

 In both experiments a portable experimental setup consisting of a laptop and high quality closed-back headphones was used. Participants were recruited opportunistically and data was gathered in public areas where it was expected participants would be likely to be listening to music naturally. The combined excerpt player and data gather tool, created in Max MSP (v6.0) and used in both reaction time test and perception experiment, was designed to look and function like a contemporary media or MP3 player, so that participants would feel comfortable and familiar with the interface. A pilot study with exit survey (N = 6) of the perception experiment was also conducted in order to improve both the instructions and presentation to participants.

 *City Life* was decided on as the source of stimuli in order to present participants with musical excerpts that did not follow structural rules that were expected from the ILE. Where the approach differed from other studies was in the desire to present participants with unfamiliar structural boundaries, whilst maintaining familiar harmonic content. If we consider the works of Maderna, Berio or Kurtág featured in those other studies, all three present a dense and relatively inaccessible post-tonal approach, combining rhythmic irregularity, stark instrumentation and strictly atonal or dissonant harmonics. *City Life*, by contrast, is both melodic and harmonic in a more traditional way. It represented a controlled departure from the ILE along desirable lines.

**Reaction Time Test**

 The aim of the reaction time test was to establish a window of reaction time that could be linked to specific musical events in the stimuli in the perception experiment. Reaction times were captured using the experimental setup already discussed. A random sample of opportunistically recruited participants (N = 50) completed the test. The stimulus was a 10 s audio file, in which the notes of a major third interval, created using an artificial sound source, were introduced sequentially. Participants were asked to press a button when they perceived the onset of the second note of the dyad. They were given the opportunity to press as many times as they wished to position their response as near to the note onset as they felt was appropriate.

 Responses varied from 223 ms to 997 ms, with a mean of 488 ms. Based on these results it was decided to allow a 1000 ms window of response after the onset of an event in the stimuli, into which responses could be clustered.

**Perception Experiment: Method**

 Two groups of 10 opportunistically recruited participants (N = 20) took part: those with informal training (IT), and those with formal training (FT). FT participants were recruited from among the students and faculty of the Music Department at the University of Jyväskylä, and IT participants were a random sample recruited in the main library of the University. Three 20 s excerpts were selected from *City Life* as stimuli, with the goal of presenting structural boundaries with decreasing salience.

 Excerpts were evaluated and selected based on a method of score analysis devised with the aim of making identified structural changes more easily comparable to audio analysis. This analysis was linked to feature change. Four defined features– Dynamics, Harmony, Tempo, and Timbre – which had been linked to specific computational analysis functions, were used, plus one undefined feature – Other. Other was used for any other notated cue of structural boundary (e.g. rhythmic change, pause, textural change, repetition). Each feature was considered to have equal importance. These features were analysed in the score in order to determine the point of the greatest number of simultaneous feature changes in each excerpt, which, it was reasoned, would signify the point of a structural boundary. Using this method the most significant point of change (MSP) and the salience of change (SoC) rating, equal to the number of features that changed at that point, were identified.

 Audio analysis was performed on the excerpts using MIR Toolbox (v1.4) functions in Matlab (R2012a). Peaks were highlighted on a novelty curve extracted from a self-similarity matrix, which was based on a time-decomposed spectrogram. This series of functions accounted in some measure for four of the features used in the musicological approach – Dynamics, Harmony, Timbre and Other (Rhythm). Because novelty curves are measured against a coefficient value of relative novelty, a novelty peak, the point of greatest relative novelty (PGN) in each excerpt, was also identified.

 Participants were asked to listen to each excerpt and decide if there was a significant change. If they could identify more than one change then they were instructed to choose the one they considered the strongest. Participants were asked to press a button on the keyboard when this change occurred. Again, they were able to press the button as many times as they wanted. They were then asked to rate the salience of the change on a Likert-type scale from 1 to 5, where 5 was the greatest salience. Participants were able to leave all options blank and move to the next clip if they felt that there was no change in an excerpt. Excerpts were presented to each participant in random order.

(Insert Excerpts 1, 2, 3)

**Perception Experiment: Results**

**Excerpt 1**

 Score analysis revealed that the MSP in this excerpt occurred on the first beat of bar 25. This point was at 9325 ms in the audio. The boundary salience was rated four, as at this point four features – Harmony, Dynamics, Timbre, Other (Rhythm) – changed simultaneously. Computational analysis showed that the PGN was at 6875 ms.

 Figure 1 shows the temporal location of the responses of both groups of participants, plotted against salience ratings. Responses were linked to events using the 1000 ms window established in Experiment 1.

 The majority of participant responses could be linked to three events. Event A is the PGN. It is the onset of the chord at bar 23 in the score, which is preceded by a quarter note rest. Event B is the MSP. At this point there is an ambiguity (Wagner, 1995): a chord from the previous bar is held for an eighth note into bar 25, overlapping the MSP. This creates uncertainty as to whether the new section starts with the MSP (Bi) or the end of the chord (Bii). We allowed a further window of 300ms (the length of the overlap), as well as the established 1000 ms, to account for this uncertainty. Event C is the introduction of a recorded sample ‘Check it out’, at bar 30 (15875 ms). Two participant responses could not be linked to events or to any other responses.

(Insert Fig 1)

**Excerpt 2**

Analysis of the score revealed that the MSP in this excerpt occurred on the second beat of bar 499. This was at 12275 ms in the audio. The boundary salience was rated three, as at this point three features (Harmony, Timbre, Other (Texture) changed simultaneously. Computational analysis showed that the PGN was at 8125 ms.

The majority of participant responses could be linked to two events: A and D (Fig. 3). Event A is the onset of the note on the first beat of bar 496 in the score, played on clarinet and violin (5775 ms) (ten responses). Event D is the onset of the note on the first beat of bar 502 in the score, played on clarinet and violin (17425 ms) (five responses). No participants responded to the PGN (Event B on the graph) or the MSP (Event C). Five participants responded in other ways. Four participant responses could not be linked to events in the score or to any other responses, and one further participant (represented at point (0,0) on the graph) identified that there was no change in the excerpt.

(Insert Fig 2)

**Excerpt 3**

The MSP for this excerpt was rated as No Change. Analysis of the score revealed that this excerpt comprised a series of repetitions (Ockleford, 2005). The initial event occurred on the first beat of bar 709 in the score (125 ms). This was repeated four times; at bar 715 (4925 ms), 722 (10325 ms), 728 (15125 ms), and 733 (19025 ms). This created a number of two (Harmony, Timbre or Harmony, Other (Rhythm)) feature changes with no point emerging as more salient than another. Audio analysis showed that the PGN was at 15125 ms (bar 728).

The largest group of participant responses were those who identified that there was no change in the excerpt (represented at point (0,0) on the graph (Fig. 5)). The greater part of the remaining responses could be linked to three events. Event A is the introduction of a recorded sample ‘stand by stand by’ at bar 716 (5725 ms). Event B is the onset of the chord on the first beat of bar 722. At this point there is a variation in the repetition – the start of the sustained chord is staggered over two beats, each marked with a separate drum hit. Event C is the PGN. Four participant responses could not be linked to events or to other responses.

(Insert Fig 3)

**Overall Results**

On the question of consensus it was found that there was a perfect negative correlation between the salience rating of structural boundaries in the excerpts and the number of distinct responses of participants as to the temporal location and salience of the structural boundary (r = -1 p = .00).

There were clear differences between the responses of the two participant groups. IT participants showed significantly more consensus on the point of the structural boundary (M = 2.67, SD = 1.53) than FT participants (M = 5.67, SD = 0.58), t (4) = -3.18, p < .05, and showed more consensus on the salience of the structural boundary (M = 3.00, SD = 1.73) than FT participants (M = 3.67, SD = 0.57), t (4) = -0.63, p = .56. Whilst only the former finding was significant, the latter observably followed the same general trend, with the mean number of distinct responses for IT participants lower than that of FT participants.

Using chi-square tests of goodness-of-fit it was determined that IT participants were significantly less likely to choose the PGN when deciding on the temporal location of the structural boundary, x² (2, N = 30) = 15, p = .00. They were also more likely than FT participants to choose the MSP as the temporal location of the structural boundary. Whilst a further chi-square test of goodness-of-fit did not reveal a significant result, x² (1, N = 22) = 2.22, p = .13, it was again observed that 15 (half of overall) IT responses were linked to the MSP, compared to 7 (less than one quarter of) FT responses.

FT participants were significantly more likely to choose neither the MSP nor the PGN, x² (2, N = 30) = 15.8, p = .00, when temporally locating the structural boundary. They were more likely than IT participants to choose the salience rating identified by score analysis, x² (1, N = 26) = 0.04, p = .84, although the difference on the last point was very small.

A final chi-square test of goodness-of-fit revealed that, overall, participants were significantly less likely to choose the PGN when temporally locating the structural boundary, x² (2, N = 60) = 25.9, p = .00.

**Discussion**

The findings of this study broadly fall into two parts. The expected was embodied in the finding that if structural boundaries were more salient, so participants’ responses exhibited more consensus and less variation, in line with the findings of Bruderer, McKinney & Kohlrausch (2009).

The unexpected was the amount and nature of the variation between the participant groups. In the early stages of this experiment it was expected that there would either be no difference between groups, as found by previous studies, or that FT participants might perform better at the task. Instead the study found the exact opposite. IT participants showed more consensus in both aspects of the experiment, and were more likely to choose the structural boundary identified by score analysis than their FT counterparts. These findings would support those of Bigand & Poulin-Charronnat (2006) that one does not need formal training to become an ‘experienced listener’.

Questions remain, however, as to why the variation between participant groups was so pronounced. Further research would need to use a larger and wider sample in order to fully account for this. There is the possibility that the variation is due to some identified methodological limitations. The version of musicological analysis employed in this study was necessarily limited in order to define a simple method of equating musicological and computational analysis. The selection of only five features, four defined and one undefined, would inevitably impact on the level of analysis possible. The term structural boundary was replaced with most significant point of feature change in the instructions to participants, on the assumption that the two were commensurate, in line with the reasoning of Addessi & Caterina (2002). This was not always the case, as exemplified by the results of Excerpt 2, where no participants perceived a boundary at either point generated by the analysis methods. Further research is required to test the robustness of these approaches in evaluating other pieces of music.

Finally, it was necessary to use computational feature extraction in order to acquire temporal points that reaction windows could be linked to. This proved problematic since the computational system used relied heavily on single feature analysis and human perception (and musicological analysis) does not. The series of functions used provided the nearest approximation, in our estimation, to a perceptual approach, basing computations on analysis of a spectrogram which could account for a wider range of features, particularly when time-decomposed. The limitation of such a system was that, as identified by Leman et al. (2005), the extracted points still needed to be related to musical features using musicological analysis. Also, as was evident in the differences between the MSP, the PGN, and the relative number of responses linked to each, our findings further supported those of Leman at al. in that “manual structural cues worked better than acoustical structural cues.” (2005, p.59).

**Summary and Conclusions**

We have presented experimental data on how participants segmented excerpts and linked that data to musicological and audio analysis. Our results are applicable to musicology, music psychology and music information retrieval, because they show some of the ways in which perception of non-traditionally structured music differs between participant groups, and from the results of theoretical approaches.

We found that when structural boundaries were more salient participants’ responses exhibited more consensus and less variation. Informally trained participants showed more overall consensus and were more likely to choose the temporal location of the structural boundary identified by score analysis. Participants overall were less likely to choose the temporal location of the structural boundary identified by computational analysis.

**Bibliography**

Addessi, A. R., & Caterina, R. (2002). Perception of the macroform in the Quartetto per Archi in due tempi (1955) by Bruno Maderna. In Second International Conference: Understanding and Creating Music. Caserta, Seconda Università degli Studi di Napoli.

Addessi, A. R., & Caterina, R. (2005). Analysis and perception in post-tonal music: an example from Kurtág’s String Quartet Op. 1. Psychology of Music, 33 (1), 94-116.

Batt-Rawden, K., & Denora, T. (2005). Music and informal learning in everyday life. Music Education Research, 7(3), 289-304.

Beach, D. (2012). Advances Schenkerian Analysis: Perspectives on Phrase Rhythm, Motive, and Form. New York: Routledge.

Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. Cognition, 100(1), 100-130.

Bruderer, M. J., Mckinney, M. F., & Kohlrausch A. (2009). The perception of structural boundaries in melody lines of Western popular music. Musicae Scientiae, 8(2), 273-313.

Cadwallader, A. & Gagne, D. (1998). Analysis of Tonal Music: A Schenkerian Approach. Oxford: Oxford University Press.

Cooper, M., & Foote, J. (2003). Summarising Popular Music via Structural Similarity Analysis. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 127-130.

Deliège, I. (1987). Grouping Conditions in Listening to Music: An Approach to Lerdahl & Jackendoff’s Grouping Preference Rules. Music Perception, Vol. 4, No. 4, 325-360.

Deliège, I., Ahmadi, A. E. (1990). Mechanisms of Cue Extraction in Musical Groupings: A Study of Perception on Sequenza VI for Viola Solo by Luciano Berio. Psychology of Music, 18, 18-44.

Foote, J. (1999). Visualising Music and Audio Using Self-Similarity. In Proceedings of ACM Multimedia ’99, Orlando, FL: ACM Press, 77-80.

Granier-Deferre, C., & Busnel, M. C. (2013). Human Fetal Auditory Perception.Fetal Development: A Psychobiological Perspective, 239.

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. Music Perception, 13, 401-432.

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of musical affect using a combination of acoustic structural cues. Journal of New Music Research, 34(1), 39-67.

Lerdahl, F., Jackendoff, R. (1983). A Generative Theory of Tonal Music. Cambridge, Mass.: MIT Press.

Ockelford A. (2005). Repetition in Music: Theoretical and Metatheoretical Perspectives. Aldershot: Ashgate Pub. Ltd.

Wagner, N. (1995). No Crossing Branches? The Overlapping Technique in Schenkerian Analysis. Theory and Practice Vol. 20, 149-175.