Facilitation and Coherence Between the Dynamic and Retrospective Perception of Segmentation in Computer-Generated Music

FREYA BAILES
Sonic Communications Research Group, University of Canberra

ROGER T. DEAN
Sonic Communications Research Group, University of Canberra

ABSTRACT: We examined the impact of listening context (sound duration and prior presentation) on the human perception of segmentation in sequences of computer music. This research extends previous work by the authors (Bailes & Dean, 2005), which concluded that context-dependent effects such as the asymmetrical detection of an increase in timbre compared to a decrease of the same magnitude have a significant bearing on the cognition of sound structure. The current study replicated this effect, and demonstrated that listeners (N = 14) are coherent in their detection of segmentation between real-time and retrospective tasks. In addition, response lag was reduced from a first hearing to a second hearing, and following long (7 s) rather than short (1 or 3 s) segments. These findings point to the role of short-term memory in dynamic structural perception of computer music.

Submitted 2006 November 24; accepted 2006 December 19.

KEYWORDS: *music perception, segmentation, cognitive discrimination, computer music*

THE perception and cognition of music involves the apprehension of the temporal variation of sonic structure. Previous research has focussed largely on the perception of structures of pitch and duration, neglecting the dynamic processes involved in perceiving non-tonal, non-metric computer composition. This music typically features timbral and textural variation; dimensions which are less easily understood within a discrete relational structure than pitch (scale) and duration (rhythm). In recent work (Bailes & Dean, 2005), we studied the perception of segments of sound in computer-generated sequences that are not based on the relationship of discrete pitches or metric patterns, but concern the interplay of timbre and texture. We showed that while listeners can perceive segmentation efficiently, the ordering of different segments might alter their perception. Specifically, we observed an asymmetrical perception in which listeners more readily detected a change in segment when the sound texture increased (addition of partials), but not when it decreased by the same magnitude (subtraction of partials). A similar asymmetry has been previously observed in the guise of auditory looming (Neuhoff, 2001) in which listeners are more sensitive to an approaching auditory source than to a retreating one. On the level of attention in musical listening, Huron (Huron, 1990a, 1990b, 1991) found that various classical music composers stagger the onset of multiple voices or prolong the increase in dynamics as compared to the duration of the equivalent offset or decrease in texture. This he dubs 'the ramp archetype'. In view of these contextual asymmetries in perception and musical structure, we concluded that in considering the cognition of sound structure, the nature of the sound context plays a significant role.

Repetition may also influence the perception of structure, though this has not been tested using non-tonal, non-metric computer-generated music. Repeated listening to a sound sequence might be expected to facilitate response in a segmentation task as compared to an initial 'real-time' response, improving confidence, memory for the juxtaposed sounds, representation for the timing of the segment change, and motor coordination to execute the task. Moreover, the length of sound segment may impact on the speed of response. First, the longer the period of time before a change in sound, the greater the stability

of that sound in the listener's short-term memory, and consequently the greater the certainty that this sound has subsequently changed. Snyder (2000) describes short-term auditory memory as extending between 3 s and 5 s (but this description does not take into account the potential impact on memory of static and homogeneous sounds versus time-variant sounds). Second, an experiment participant instructed to respond to a change in sound will be in a state of increasing readiness to act as time passes.

The purpose of this study is to examine the impact of the contextual factors of segment length and real-time versus retrospective listening on the perception of segmentation in computer-generated sound sequences.

METHOD

In previous research (Bailes & Dean, 2005), we asked participants to judge retrospectively whether sound sequences were one long segment, or comprised two separate consecutive segments. In the *segmentation procedure* (Deliège et al., 1996), listeners indicate segmentations by a key press when perceiving a change in sound. The current study combines both techniques to contrast retrospective and real-time perceptions of segmentation.

Participants

Participants (N = 14) were recruited through the 1st year psychology undergraduate credit system, at the University of Canberra. Half of these participants reported having studied music beyond school, and four reported still occasionally playing an instrument. Five men and nine women took part, with a median age of 19.5 years (ages 18-25). Participants reported normal hearing, and listening to a median of 14 hours of music per week (range 3-42 hours).

Stimuli

As in our previous experiments (see Bailes & Dean, 2005, 2007), a range of computer-generated sound segments was created. Sounds were selected that varied little within a segment, and which, when juxtaposed, would range from more obvious to subtle segmentation. A description of the algorithms used to generate the sounds is presented in Table 1.

Table 1. Brief description of the types of sound segment used as stimuli, including the patch used for their generation

Segment type	Patch	Description
At	Atau's 'Relooper Redux' (comes with *MAX/MSP* software[a])	Short chunks of a speech sample at different speeds
60Hz	'60Hz: Embrace the inner ground loop' (*MAX/MSP*[a])	Multiple sinusoids based on the harmonic series root 60Hz
FB	'Forbidden Planet' filtering patch (*MAX/MSP*[a])	Noise input filtered by notches
LwH	*MAX/MSP*[a] patch written by Roger T. Dean	Entirely synthetic noise and sine components
PiS	N/A - overlaid samples	Sounds derived from a sample of a fizzing noise

[a]*MAX/MSP (4.2-4.5)* (Cycling '74, San Francisco, CA 94103, USA)

For two-segment stimuli, the point of segment change was designed to occur at various intervals, so that a listener could not predict it. Three sets of stimuli were devised. The first set of two-segment items had the form AB, BA as lengths such as x seconds: y seconds, y seconds: x seconds. The second set had the form AB, BA as lengths such as x seconds: y seconds, x seconds: y seconds. There were 6 AB stimuli and their reverse (6 x BA) in each set, so that item/content order was consistently controlled. Sound content differed in the two sets. A third set comprised 12 one-segment files (4-12 s in length), using sounds from sets 1 and 2.

Segment lengths were chosen to distribute total and constituent lengths as evenly as possible across the pool of stimuli. We ranged the segment length from very short (1 s) to long (7 s) with intermediate lengths (3 s and 5 s) to see whether the supposed limits of short-term memory would impact on segment detection using these different lengths (see Snyder, 2000).

Sound segments were generated in *MAX/MSP (4.2-4.5)* (Cycling '74, San Francisco, CA 94103, USA) and recorded direct to AIFF by *Audio HiJack Pro (2.2)* (Rogue Amoeba, Cranbury, NJ, 08512, USA) (44.1 kHz sampling rate, 16 bit throughout). All files were normalized to -1dB in *ProTools* (v.7; Digidesign, a Division of Avid Technology, Daly City, CA 94014, USA).

Changes between different sound segments were sudden rather than gradual, although the interface was adjusted by ear: any detectable clicks were removed using a cross-fade in *ProTools* of no more than 200 ms. The 36 stimuli were assembled in *Soundedit* as mono files.

Among the more subtly differentiated segmentations were stimuli that merely comprised a difference in filtering between segment A and segment B. For example, stimuli using the 'Forbidden Planet' algorithm (see Table 1) systematically applied controlled levels of filtering to the same noise file. These are summarised in Table 2.

Table 2. The 'Forbidden Planet' filtering algorithm was applied to the same noise-based sound, and this table summarises the varieties of filter

Segment name	Description
FB01	through filter
FB03(-3)	small notch high (with subsequent coarse pitch shift down 3 semitones)
FB04	small notch middle
FB06	spot notch low
FB07	spot notch high
FB09-FB10	cumulative series of notches letting through less and less bass
FB13	low pass
FB14	high pass
FB15	brick wall low pass
FB17	time variant moving of the filtering window with cursor low
FB18(-3)	time variant moving of the filtering window with cursor low-mid (with subsequent coarse pitch shift down 3 semitones)

Many of the individual sound segments are time-variant, and consequently the inter-segment relationships are not easy to quantify. As with more traditional forms of musical material, qualitative terms are used to describe most of the 36 stimuli, and the contrast between one sound segment and another.

Procedure

Participants were tested individually, hearing items in a random order over *AKG K271 Studio* headphones, seated at an *iBook 900MHz PowerPC G3*. They were presented with written instructions first, and then instructions were presented on the computer screen with the playback and recording mechanism of *Psyscope* (Cohen et al., 1993).

Participants were instructed that they would hear a passage of computer-generated sound, and that their task would be to decide whether the sound changed so there were two segments of different sound, or whether the sound stayed the same so there was only one segment of sound. If they thought the sound changed from one segment to another, they were to indicate the point of change by pressing the space bar as soon as possible after it. If they heard only one segment, there was no need to press the space bar.

After the first listening, participants were asked to make a categorical statement as to whether they had heard one or two segments of sound, by pressing '1' or '2'. Then they had a second chance to listen and, if appropriate, to indicate when in the passage they heard any change in sound segment. Again, if they thought the sound changed from one segment to another, they were to indicate the point of change by pressing the space bar as soon as possible after it. If they heard only one segment, there was no need to press the space bar. Participants were encouraged to answer differently at different phases of the task if necessary, as it was explained that their final response, during the second hearing, should best reflect their overall judgement. Three practice trials with on-screen feedback preceded the main session. 36 sequences were presented in a random order (12 two-segment files and their reverse, plus 12 one-segment files). Participants initiated trials by key press, so the interval between items was self-regulated. Sessions took around 30 minutes, including filling out a questionnaire at the end that elicited background demographic and music training information (including familiarity with musical genres).

RESULTS

'Errors' were counted for each of the 36 items at both hearings and for the categorical judgement task, where 'error' is defined as a 'one-segment' judgment for an algorithmically segmented file, or 'two-segments' for an algorithmically non-segmented file (note that the use of the word 'error' does not mean that listeners were wrong in their judgement, but denotes a discrepancy with the algorithmic composition). Participants made a total of 9% error in judging whether sequences consisted of one- or two-segments. A chi-square test to compare the proportion of errors against chance performance (i.e. 50% error) was used on all items. The results indicate that participants detected segmentation significantly better than chance for 31 of the 36 items. Such a high level of accuracy is approaching ceiling performance on the task overall. Nevertheless, it was of interest to determine whether participants improved in detecting segmentation from the first listening to the categorical judgement to the second hearing. A one-tail paired t-test between errors for the first listening and the subsequent categorical judgement revealed no significant difference [$t_{13} = -1.4$, p = n.s.]. In addition, errors were not statistically different between first and second hearings [$t_{13} = 0$, p = n.s.].

Accuracy in detecting segmentation did not change from the initial real-time listening to the retrospective tasks. However, it was of interest to see whether participants reduced the lag in indicating the moment of segment change between first and second hearings. Response time (RT) to indicate a real-time change in segment was measured in milliseconds for two-segment items, from the algorithmically defined change in segment to the point of key press. A one-tail paired t-test revealed that mean RT per item was significantly closer to the algorithmic point of segmentation in the second hearing than in the first [$t_{23} = 8.14$; $p < 0.001$].

The items for which participants were no more accurate than chance in detecting segmentation are in line with findings from our previous experiments (Bailes & Dean, 2005). Namely, an asymmetry in detecting segmentation was found for a couple of items when the order of segment A and B was reversed. For instance, participants were no better than chance in detecting a change in segment in FB9FB10 (Use the following link to download the Audio 1 sound file:

https://kb.osu.edu/dspace/bitstream/1811/28854/1/EMR000026a_bailes01.ogg .) [X^2 = 2.74; df = 2; p = n.s.], but did detect segmentation in FB10FB9 (Use the following link to download the Audio 2 sound file: https://kb.osu.edu/dspace/bitstream/1811/28854/2/EMR000026a_bailes02.ogg .) [X^2 = 17.29; df = 2; p < 0.001].

Participants also failed to detect segmentation in FB15FB13, FB13FB15, FB01FB04 and FB04FB01. On closer examination, participants made more errors detecting segmentation in FB15FB13 (seven) than FB13FB15 (two). In both FB9FB10 and FB15FB13, the change from the first segment to the second represents a diminution in timbre intensity, with partials filtered out (see Table 2). As Figure 1 shows, FB09FB10 comprises two segments which change at the 3 second point by the removal (filtering) of a band of high frequencies distributed around 10750Hz. Using the speech and sound analysis software *Praat 4.2* (freely available from www.praat.org <http://www.praat.org>), on Macintosh OSX, we measured the spectral power at 10750 Hz before and after the transition. Quite stable values occur in each segment, and the transition can be summarised by the values at 2.5 and 3.5 seconds. These were 0.03331 and 0.00035 Pa^2/Hz (Pascals-squared/Hz) respectively, a numerical difference of almost 2 orders of magnitude.

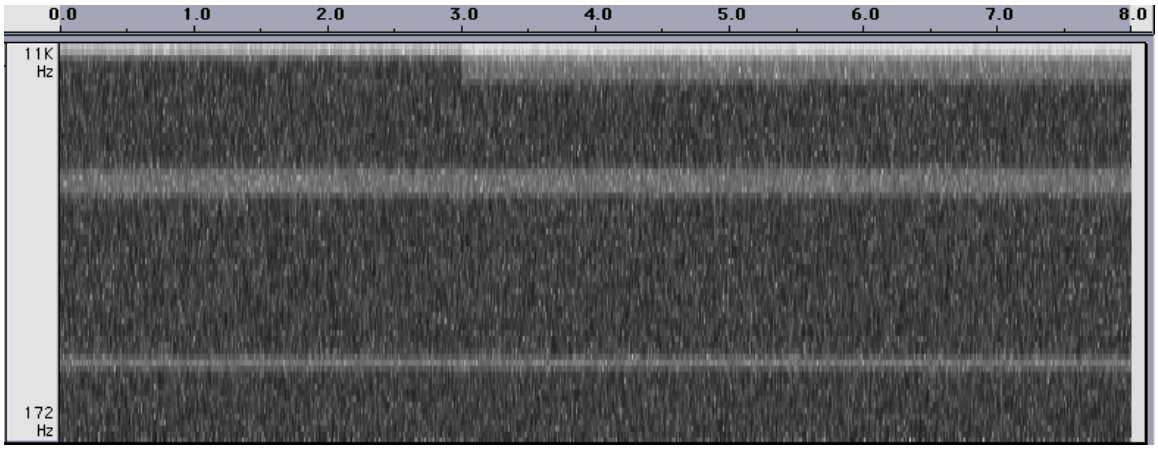

Fig. 1. FB09FB10. The vertical axis in this sonogram is frequency, density of greyscale is intensity, and the time axis (in s) is horizontal [1]

In FB10FB9 and FB13FB15, the change in segment conversely represents an increase in timbral texture, with the addition of partials. Thus the ramp archetype (or auditory looming) is apparent, in that an increase in sound is perceived while a decrease of the same magnitude and extent is not.

One anomaly is the finding that participants were no more accurate than chance in indicating that FB10FB10 (Use the following link to download the Audio 3 sound file: https://kb.osu.edu/dspace/bitstream/1811/28854/3/EMR000026a_bailes03.ogg .) is a one-segment item [X^2 = 2.57; df = 1, p = n.s.]. A closer examination of the pattern of error for this item suggests that participants indicated 'one segment' for both the first hearing and the categorical task, and then changed their response during the second hearing. The reason for this shift in response strategy between hearings is not obvious.

To determine whether initial segment length affected response lag, mean RT were analysed using a repeated measures ANOVA and a within-subjects factor of segment length. This analysis was conducted separately for first and second hearings. No overall effect of segment length was found for either the first [$F_{3, 11}$ = 1.29; p = n.s.] or second [$F_{3, 11}$ = 3.07; p = n.s.] hearings. However, planned comparisons showed that response was significantly faster for initial segments of 7 s (mean RT of 211 ms) than for 1 s (282 ms) or 3 s (287 ms) [p < 0.05] in the second hearing.

Two-segment items were classified according to their relative segment durations as short-long or long-short sequences. RT were analysed using a repeated measures ANOVA, with two within-subjects factors of relative duration and item/content order. Only data from set 1 were used in this analysis, comprising the items controlled for both relative duration and item/content order. For data from the first hearing, a significant effect of relative duration was observed [$F_{1, 5}$ = 8.63; p < 0.05], with short-long sequences eliciting a greater lag than long-short (577 ms and 425 ms respectively).

CONCLUSIONS

The results extend our previous findings of efficient perception of segmentation in digital sound. Previous data concerned segmentation of 14 s sound sequences, in which the segmentation point, when present, was at the midpoint of the sequence. In this work we have used a variety of temporally asymmetric segmentation constructs, showing that efficient perception remains, and even with very short segments (1 s). More importantly, the findings from this experiment collectively reflect the need to consider the dynamic perception of music and the role of temporal context on the cognition of sonic structure. First, differences between 'on-line' or real-time response and retrospective judgements were negligible with respect to accuracy, demonstrating coherence between the two modes of perception (this outcome is in spite of the emphasis on changing response during the successive tasks such that the final reaction would best reflect the listener's judgement). However, a repetition priming effect occurred as participants significantly reduced their response lag from the first hearing to the second. A question for future research is how many listenings are required for participants to optimally locate the segmentation point.

The context-dependence of structural perception is highlighted by the different lag following different length segments. During the first hearing, it seems that having heard the initial segment for longer facilitated the speed of response to the change in sound. During the second hearing, response was significantly faster for segments of 7 s than for 1 s or 3 s. Perhaps a more stable representation of the sound against which to compare the change is responsible for this finding, with the longer segment exceeding the temporal extent of short-term memory (Snyder, 2000). It is also plausible that participants were better primed to respond physically after a greater overall delay.

Finally, our observed replication of the ramp archetype (Bailes & Dean, 2005) in which participants are better able to hear an increase than a decrease in intensity reiterates the importance of considering the immediate sound context in the perception of structure, and thus dynamic, context-dependent cognition [2].

NOTES

[1] The sonogram is a screen captured from an Audacity spectrogram.

[2] The research reported in this paper is supported in part by an *Australian Research Council* 'Discovery' grant (DP0453179) held by Roger Dean.

REFERENCES

Bailes, F., & Dean, R. T. (2005). Structural judgements in the perception of computer-generated music. In *Proceedings of the 2nd International Conference of Asia Pacific Society for the Cognitive Science of Music.* Seoul, Korea, pp. 155-160.

Bailes, F., & Dean, R. T. (2007). Listener detection of segmentation in computer-generated sound: An experimental study. In press, *Journal of New Music Research.*

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments and Computers*, Vol. 25, pp. 257-271.

Deliège, I., Mélen, M., Stammers, D., & Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, Vol. 14, No. 2, pp. 117-160.

Huron, D. (1990a). Crescendo/diminuendo asymmetries in Beethoven's piano sonatas. *Music Perception*, Vol. 7, No. 3, pp. 395-402.

Huron, D. (1990b). Increment/decrement asymmetries in polyphonic sonorities. *Music Perception*, Vol. 7, No. 3, pp. 385-394.

Huron, D. (1991). The ramp archetype: A score-based study of musical dynamics in 14 piano composers. *Psychology of Music*, Vol. 19, No. 1, pp. 33-45.

Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, Vol. 132, pp. 87-110.

Snyder, B. (2000). *Music and memory: An introduction*. Cambridge, MA: The MIT Press.

APPENDIX

Audio 1: FB09FB10 was identified as a two-segment file at chance levels of accuracy only. The audio examples presented with this paper are compressed as .ogg files, with a constant bit rate of 320kbps. They were made in *Audacity* (open source software available from http://audacity.sourceforge.net/) exploiting the LAME plug-in. Original uncompressed audio files are available from the authors. (Use the following link to download the audio file for this example: https://kb.osu.edu/dspace/bitstream/1811/28854/1/EMR000026a_bailes01.ogg .)

Audio 2: FB10FB09 was correctly identified as a two-segment sequence at levels significantly above chance performance. (Use the following link to download the audio file for this example: https://kb.osu.edu/dspace/bitstream/1811/28854/2/EMR000026a_bailes02.ogg .)

Audio 3: Listeners tended to judge this one segment file, FB10FB10, as two consecutive segments. (Use the following link to download the audio file for this example: https://kb.osu.edu/dspace/bitstream/1811/28854/3/EMR000026a_bailes03.ogg .)