The Perception of Grouping Boundaries in Music

Marcus T. Pearce

Department of Computing, Goldsmiths College,
University of London, New Cross, London SE14 6NB
m.pearce@gold.ac.uk

May 1, 2008

1 Introduction

Some of the most fundamental questions in the study of music perception concern the manner in which the human perceptual system groups musical elements together.

In this paper, we examine the grouping of musical elements into contiguous segments that occur sequentially in time or, to put it another way, the identification of boundaries between the final element of one segment and the first element of the subsequent one. This way of structuring a musical surface is usually referred to as *grouping* or *segmentation* and is to be distinguished from the grouping of musical elements that occur simultaneously in time, a process usually referred to as *streaming* (Bregman, 1990). In musical terms, the kinds of group we shall consider might correspond to motifs, phrases, sections and other aspects of musical form. Just as speech is perceptually segmented into words which subsequently provide the building blocks for the perception of phrases and complete utterances (Brent, 1999), motifs or phrases in music are identified by listeners, stored in memory and made available for inclusion in higher-level structural groups (Lerdahl & Jackendoff, 1983; Peretz, 1989; Tan *et al.*, 1981). The low-level organisation of the musical surface into groups allows the use of these primitive perceptual units in more complex structural processing and may alleviate processing and memory demands.

We restrict ourselves primarily to research on symbolic representations of musical structure that take discrete note-like events as their musical surface (Jackendoff, 1987). Research on segmentation from sub-symbolic or acoustic representations of music (e.g., Abdallah *et al.*, 2006; Gjerdingen, 1999; Todd, 1994; Šerman & Griffith, 2004) are not discussed. Furthermore, the present work emphasises melody (although not exclusively) reflecting the predominant trends in theoretical and computational treatments of perceived grouping structure in music.

Grouping structure is generally agreed to be logically independent of metrical structure (Lerdahl & Jackendoff, 1983) and some evidence for a separation between the psychological processing of the two kinds of structure has been found in cognitive neuropsychological (Liegeoise-Chauvel *et al.*, 1998; Peretz, 1990) and neuroimaging research (Brochard *et al.*, 2000). In practice, however, metrical and grouping structure are often intimately related and both are likely to serve as inputs to the processing of more complex musical structures (Lerdahl & Jackendoff, 1983). Nonetheless, most theoretical, empirical and computational research has considered the perception of grouping structure independently of metrical structure (Stoffer, 1985 and Temperley, 2001 being notable exceptions).

The review is structured as follows. Theoretical accounts, empirical studies and computational models of the perception of grouping structure in music are reviewed and discussed in §2, §3 and §4 respectively. The review concludes in §5 with a general discussion of the research summarised herein.

2 Theoretical Descriptions

2.1 A Generative Theory of Tonal Music

The *Generative Theory of Tonal Music* (GTTM) of Lerdahl & Jackendoff (1983) is probably the best known effort to develop a comprehensive method for the structural description of tonal music, which is heavily inspired in many respects by Chomskian grammars. It is, for example, founded on the assumption that a piece of music can be partitioned into hierarchically organised segments which may be derived through the recursive application of the same rules at different levels of the hierarchy. Specifically, the theory is intended to yield a hierarchical, structural description of any piece of Western tonal music which corresponds to the final cognitive state of an experienced listener to that composition.

According to GTTM, a listener unconsciously infers four types of hierarchical structure in a musical surface: first, *grouping structure* which corresponds to the identification of segment boundaries (on the basis Gestalt-like principles of event proximity and similarity); second, *metrical structure* which corresponds to the pattern of periodically recurring strong and weak beats; fourth, *time-span reduction* which represents the relative structural importance of pitch events within contextually established rhythmic units; and finally, *prolongational reduction* reflecting patterns of tension and relaxation amongst pitch events at various levels of structure. According to the theory, grouping and metrical structure are largely derived directly from the musical surface and these structures are used in generating a time-span reduction which is, in turn, used in generating a prolongational reduction. Each of the four domains of organisation is subject to *well-formedness rules* that specify which hierarchical structures are permissible and which themselves may be modified in limited ways by *transformational rules*. While these rules are abstract in that they define only formal possibilities, *preference rules* select which well-formed or transformed structures actually apply to particular aspects of the musical surface. Time-span and prolongational reduction additionally depend on tonal-harmonic *stability conditions* which are internal schemata induced from previously heard musical surfaces.

When individual preference rules reinforce one another, the analysis is stable and the passage is regarded as stereotypical whilst conflicting preference rules lead to an unstable analysis causing the passage to be perceived as ambiguous and vague. In this way, according to GTTM, the listener unconsciously attempts to arrive at the most stable overall structural description of the musical surface.

Of particular interest here is the grouping component of the theory which determine the segmentation of a musical surface into a hierarchically organised structure of embedded groups through the identification of perceived segment boundaries. The Grouping Well-formedness Rules ensure that a piece is assigned a single hierarchical grouping structure such that each group consists of events which are contiguous in time and any group containing a smaller group is exhaustively partitioned into subgroups (see London, 1997, for a critique of these and other aspects of the theory). In general, overlapping groups are not well-formed although limited provision is made for accommodating single notes being the last in one group and the first in the following group. The seven Grouping Preference Rules (GPRs) of the theory are reproduced in Table 1. Of these, only GPRs 2 and 3 actually specify possible boundary locations whilst the

GPR 1		Avoid analyses with very small groups – the smaller the less preferable.
GPR 2	Proximity	Consider a sequence of four notes n_1, n_2, n_3 and n_4. All else being equal, the transition n_2-n_3 may be heard as a group boundary if:
	a. Slur/Rest	the interval of time from the end of n_2 to the beginning of n_3 is greater than that from the end of n_1 to the beginning of n_2 and that from the end of n_3 to the beginning of n_4;
	b. Attack-point	the interval of time between the attack points of n_2 and n_3 is greater than that between n_1 and n_2 and that between n_3 and n_4.
GPR 3	Change	Consider a sequence of four notes n_1, n_2, n_3 and n_4. All else being equal, the transition n_2-n_3 may be heard as a group boundary if:
	a. Register	the transition n_2 to n_3 involves a greater intervallic distance than both n_1 to n_2 and n_3 to n_4;
	b. Dynamics	the transition n_2 to n_3 involves a change in dynamics and n_1 to n_2 and n_3 to n_4 do not;
	c. Articulation	the transition n_2 to n_3 involves a change in articulation and n_1 to n_2 and n_3 to n_4 do not;
	d. Length	n_2 and n_3 are of different lengths, and both pairs n_1, n_2 and n_3, n_4 do not differ in length.
GPR 4	Intensification	Where the effects of GPRs 2 and 3 are relatively more pronounced, a larger level group boundary may be placed.
GPR 5	Symmetry	Prefer grouping analyses that most closely approach the ideal subdivision of groups into two parts of equal length.
GPR 6	Parallelism	Where two or more segments of the music can be construed as parallel, they preferably form parallel parts of groups.
GPR 7	Time-span and Prolongational Stability	Prefer a grouping structure that results in more stable time-span and/or prolongation reductions.

Table 1: The seven GPRs of GTTM (Lerdahl & Jackendoff, 1983).

rest (GPRs 1 and 4-7) determine which boundaries are retained at higher levels of the grouping hierarchy. As a consequence, GPRs 2 and 3 have received the most attention in experimental studies of the theory (see §3.4) and computational models based on it (although Temperley (2001) and Cambouropoulos (2006) also consider parallelism as it appears in GPR 6; see §4).

2.2 The Implication-Realisation Theory

Like GTTM, the *Implication-Realisation* (IR) theory (Narmour, 1990, 1992) is a detailed attempt to draw up a comprehensive theoretical account of music cognition. This theory, which is applied exclusively to melody, views the experience of listening to music as an dynamic process in which sounding structures create implications for forthcoming structures which may or may not be realised. The process of implication is held to be dependent both on the musical experience of the listener (the *top-down system*) and a set of hard-wired, innate and universal cognitive principles (the *bottom-up system*). The IR theory has been the subject of several detailed reviews published in the psychological and musicological literature (Cross, 1995; Krumhansl, 1995; Thompson, 1996).

According to the theory, the principal determinant of grouping boundaries is *melodic closure*, which is created by structures that create little or no implication. In the bottom-up system, sequences of melodic intervals vary in the degree of *closure* that they convey according to the degree to which they exhibit the following characteristics:

1. an interval is followed by a rest;
2. the second tone of an interval has greater duration than the first;
3. the second tone occurs in a stronger metrical position than the first;
4. the second tone is more stable (less dissonant) in the established key or mode than the first;
5. three successive tones create a large interval followed by a smaller interval;
6. registral direction changes between the two intervals described by three successive tones.

Narmour (1990) provides rules for evaluating the influence of each condition on the closure conveyed by a melodic fragment.

Closure signifies the termination of ongoing melodic structure and the melodic groups either side of the boundary thus created can share different amounts of structure depending on the degree of closure conveyed. Furthermore, structural tones marked by strong closure at one level can *transform* to a higher level, itself amenable to analysis as a musical surface in its own right, thus allowing for the emergence of hierarchical levels of structural description of a melody.

3 Empirical Studies

The goal in this section is both to summarise the various experimental procedures which have been developed to examine the perception of melodic phrase structure in music and to review the results of such empirical behavioural studies especially insofar as they reflect on the assumptions of the theoretical accounts summarised in §2.

3.1 Basic Evidence for Phrase Boundary Perception

Some basic evidence that listeners perceptually organise melodies into structural groups has been obtained using a click localisation paradigm in which musically trained participants listen to a musical passage in one ear and are asked to mark on a printed score the position of a click presented to the other ear. Research using this paradigm has demonstrated that listeners exhibit migration of perceived click location towards phrase boundaries defined by the notated grouping of the notes (Gregory, 1978) and by the melodic phrase structure in terms of rhythmic accent and implied harmony (Sloboda & Gregory, 1980) as well as metric structure and contour changes (Stoffer, 1985, experiment 1). In addition, Stoffer (1985, experiment 1) found that clicks were more likely to migrate to positions which interrupted fewer hierarchically organised segment boundaries above the level of the bar. This finding was interpreted in terms of a top-down process of click localisation in a hierarchically organised representation of segment boundaries leading to less efficient access to lower-level representational units than to higher-level ones.

Stoffer (1985, experiment 2) adapted the click localisation paradigm in order to examine the perception of phrase structure for both musically trained and untrained listeners. In the resulting click detection paradigm, musically trained and untrained participants were asked to listen to a musical passage presented in one ear and to press a key as soon as they detected a click presented in the opposite ear.[1] The stimuli were 12 bar German folk melodies which fell into two categories: those with simple phrase structure which could be parsed according to a heuristic of binary subdivision and those with complex phrase structure which could not. The experiment was run in four sessions with an interval of one week between each session and reaction times (RT) on the click detection task were collected as an indication of the perceived prominence of the click location in a hierarchy of segment boundaries. In general, the results confirmed those of the click localisation experiments with shorter RTs to clicks located at higher levels in the hierarchy of segment boundaries. The most notable finding was that the untrained (but not the trained) participants exhibited a significant training effect between the first and final sessions for the stimuli with complex phrase structure only. The RT data suggested that in the first session, they were attempting to segment the melody according to a principle of binary subdivision while in the final session, their perceptual segmentation of these melodies approached that of the trained listeners which itself corresponded closely in both sessions to the actual phrase structure.

3.2 The Influence of Harmonic and Rhythmic Structure

Tan *et al.* (1981) conducted a study of harmonic influences on perceptual organisation in which listeners were presented with isochronous two-phrase melodies and asked to indicate whether a two-note sequence (the probe) had occurred in the melody. The participants included both musicians and non-musicians and the melodies varied according to whether the first phrase ended with a full cadence or a semicadence. The critical probes were taken from one of three locations in the melody: first, ending the first phrase; second, straddling the phrase boundary; and third, beginning the second phrase. As predicted by Tan *et al.*, the results demonstrated that probes in the second position were more difficult to recognise than those in other positions. This effect was found to be much stronger for the musicians than for the non-musicians. Furthermore, the results showed other effects of musical training: while for non-musicians, the effect of probe position was no different for full cadences and semicadences, the musicians not only showed the strongest probe position effect in the full cadence condition but also strikingly better performance on last-phrase probes than on first-phrase probes in this condition (but not the semicadence condition).

On the basis of these results, Tan *et al.* (1981) argue that harmonic structure influences perceptual organisation of melodies in ways analogous to the influence of clause relations on the perceptual organisation of sentences and that this influence is critically dependent on musical experience. They suggest that training may improve the ability to encode a melody in terms of its abstract harmonic properties. The presence of a full cadence (rather than a semicadence) facilitates this encoding leading to a more marked probe position effect; furthermore, to the extent that the first phrase has been fully encoded, recognition performance for first-phrase

[1] In both experiments reported by Stoffer (1985), the participants were asked to perform a recognition task immediately after detecting the click to encourage them to focus primarily on the structure of the melody, and to avoid response biases resulting from a focus of attention on the ear to which the click was presented.

probe positions would be expected to be inferior since a representation of the precise sequence of notes is no longer available.

Using a similar experimental paradigm, Dowling (1973) examined the influence of rhythmic organisation on the perception of phrase structure in music. In a short-term recognition-memory task, participants listened to tone sequences consisting of four groups of five tones where each group was distinguished by a relatively lengthened final tone. Recognition performance for test phrases which crossed rhythmic groups was significantly worse than for those which corresponded to a rhythmic group in the trial stimulus. There was also a recency effect such that later items were more easily recognised.

Deutsch (1980) has examined the influence of hierarchical structure on the perception of music using a melody recall task in which trained musicians were asked to listen to a sequence of tones and subsequently recall the sequence in music notation. The stimuli varied both in terms of the degree of recursively generated repetition of intervallic patterns (structure vs unstructured) and in terms of their temporal segmentation (no segmentation, segmented in accordance with intervallic pattern structure and segmented in violation of intervallic pattern structure). An analysis of variance conducted on the percentage of tones correctly recalled revealed a significant interaction between structure and temporal segmentation. As predicted, performance was good for the structured sequence, and considerably better than for any of the unstructured sequences, except when temporally segmented in conflict with sequence structure. In a subsequent experiment (Deutsch, 1980, experiment 2), these findings were found to be independent of the number of tones forming the temporal and intervallic groups. Furthermore, for sequences structured in terms of intervallic patterns into groups of four tones, performance with temporal groups of two tones was almost as good as that for temporal groups of four tones.

Deutsch (1980) interpreted these results as supporting a theory of music perception in which music is represented in terms of a hierarchical structure generated by the recursive application of transformational operators on tone sequences (Deutsch & Feroe, 1981). However, the probability of recall errors was much higher for the first tone in a temporal group than for tones within temporal groups (Deutsch, 1980, experiment 1) and an analysis of the proportion of groups correctly recalled in their entirety indicated better performance for temporal groups than intervallic groups suggesting that temporal proximity exerted a greater influence than the repetition of intervallic patterns (Deutsch, 1980, experiment 2). In subsequent research, however, Boltz & Jones (1986) compared recall performance for melodies exhibiting symmetrical hierarchical compositional structure, asymmetrical linear compositional structure and no compositional structure at all. The results demonstrated that differences in recall performance between the linear and hierarchical stimulus conditions were strictly a result of the number and timing of contour changes within the melody rather than the ease with which a melody could be hierarchically structured through the recursive application of rules.

In subsequent research, Boltz (1991) conducted a melody recall experiment in which musically trained participants were asked to listen to unfamiliar folk melodies and then recall them using music notation. The melodies varied in two dimensions: first, according to whether phrase endings (determined by metric and intervallic structure) were marked by tonic triad members; and second, according to whether temporal accents coincided with the melodic phrase structure. The performance metric was the percentage of recalled notes forming the correct absolute interval with the preceding note. The results demonstrated that performance decreased when temporal accents conflicted with phrase boundaries and that the marking of phrase boundaries by tonic triad members resulted in significantly better performance but only when these boundaries were also marked by temporal accents. In a subsequent analysis, Boltz (1991) classified the errors into three classes: those due to missing notes; those due to incorrect interval size but correct contour; and those due to incorrect interval size and incorrect contour. For melodies with coherent temporal accents, a significantly large proportion of the errors were contour preserving and this was not the case for melodies with incoherent temporal accents. Finally, Boltz (1991) conducted an error analysis specifically at phrase endings which demonstrated that for coherent melodies with phrases marked by tonic triad members, a large proportion of the recalled notes were tonic triad members (correct or incorrect). For incoherent melodies with phrases marked by tonic triad members, however, the phrase-final notes were most frequently recalled as non-tonic triad members or missing notes.

3.3 The Relationship Between Pitch and Rhythmic Structure

Other research has examined in more detail the manner in which pitch and temporal structure influence the perception of phrase structure. Palmer & Krumhansl (1987a) developed two paradigms in which musical segments are rated according to how good or complete a musical phrase they make. In the first paradigm, a musical excerpt is manipulated in order to generate three conditions: the *pitch* condition preserved the original pitch pattern but all tones have the same duration; the *temporal* condition preserves original rhythmic pattern but all tones have the same pitch; and the *combined* condition retains both the original pitch and rhythmic patterns. In this paradigm, different trials are generated with stimuli of different length by truncating a variable number of events from the end of the excerpt. Palmer & Krumhansl identify two possible criticisms of this paradigm: first, it might be argued that the stimuli in the pitch and temporal conditions do not resemble real music; and second, the final sounded events may contribute too heavily to the phrase judgements since, within conditions, the beginning of each trial is the same and only the end of each trial differs. In order to address these concerns, Palmer & Krumhansl describe a second paradigm in which the musical excerpt is again manipulated to generate three conditions: the *pitch shift* condition preserves the original rhythmic pattern but shifts the pitch pattern; the *temporal shift* condition preserves the original pitch pattern but shifts the rhythmic pattern; and the *combined shift* condition shifts both the pitch and rhythmic patterns by the same amount and in the same direction. Distinct trials differ in terms of the amount by which the material is shifted.

Palmer & Krumhansl (1987a, experiments 1 and 2) used these paradigms to collect phrase judgements from musically trained participants using a Bach fugue subject as the original musical passage. In both experiments, the phrase judgements for the pitch and temporal conditions both correlated significantly with the combined condition but not with each other. Furthermore, in subsequent multiple regression analyses, a linear combination of judgements in the pitch and temporal conditions yielded a significantly better fit than either of the simple correlations in both experiments. Palmer & Krumhansl (1987a) found no significant relationships between musical experience or training on the component weights of the multiple regression solution in either experiment. Since more complex regression models with interaction terms failed to provide a better fit than the simple additive model, Palmer & Krumhansl (1987a) argued on the basis of these results that pitch and temporal information are represented independently in the perception of musical phrases.

It is possible that the characteristics of the fugue excerpt studied by Palmer & Krumhansl (1987a) may have encouraged perceptual separation of pitch from temporal information. In order to test this hypothesis, Palmer & Krumhansl (1987b) repeated the two experiments using a classical harmonic passage consisting of the opening theme from the A major piano sonata by Mozart (K. 331). In all of their experiments, the linear combination of phrase judgements from pitch and temporal conditions accurately predicted those from the combined condition. The correlations between phrase judgements in pitch and temporal conditions were insignificant or only weakly significant in all cases except for listeners familiar with the excerpt in the first experimental paradigm. Palmer & Krumhansl (1987b) argued that these high correlations were a result of imagery for the missing component which was not possible in the shifting patterns paradigm. Finally, no significant differences were found when the excerpts were performed with expressive timing suggesting that the observed independence of pitch and temporal information was not a result of the lack of timing deviations which are typically related to pitch structure.

The independence between pitch and rhythmic structure observed by Palmer & Krumhansl (1987a,b) stands at odds with other research which has demonstrated an interactive influence of these two dimensions of musical structure on behaviour. Experimental variations of rhythmic structure have been found to facilitate performance for some pitch patterns but not others in studies of recognition memory for melodies (Jones *et al.*, 1982), melody recall (Boltz, 1991; Boltz & Jones, 1986; Deutsch, 1980), the generation of expectations (Boltz, 1989b, 1993; Schmuckler & Boltz, 1994) and judgements of melodic completeness (Boltz, 1989a). For example, Boltz (1989a) found that melodies ending on an interval from the leading tone to a tonic triad member were judged as more complete when the rhythmic context accented that interval and that this occurred only when a tritone appeared in the final few bars to clarify the tonality.

In an effort to resolve this apparent paradox, Boltz (1999) hypothesised that the extent to which pitch and temporal information are processed independently might vary as a function of the degree of concordance between these dimensions in the stimulus, the familiarity of the listener with the stimulus and the extent to which their attention is focused on one or the other dimension. These hypotheses were examined in a task in which non-musicians were familiarised with a set of melodies and subsequently asked to listen to each melody again and then reproduce the total duration of the melody via a button press. The melodies were of two kinds: for those the *coherent* condition, temporal and melodic accents coincided whilst for those in the *incoherent* condition, the two types of accent conflicted. Attention was manipulated by dividing the participants in to four groups each of which received different instructions during the familiarisation phase: the first group was given no instructions whilst the remaining groups were respectively instructed to attend to the duration of each melody, the pitch information in each melody, and both of these factors. The performance metric used was the mean percent absolute error in reproduced duration. The results demonstrated that performance did not vary across the four attention conditions for coherent melodies. For incoherent melodies, however, performance was comparable to the coherent condition when participants were attending to duration and considerably worse in the other three conditions (though less so when given no instructions).

In a subsequent experiment, Boltz (1999, experiment 2) repeated the experiment with an additional attending condition in which participants were instructed to attend to rhythmic information. Furthermore, performance was assessed on two additional memory tasks: first, a simple rhythm recall task; and second, a pitch recognition task. The results were analogous to those in the first experiment. Performance was generally good for coherent melodies and did not vary across attending conditions. For incoherent melodies, however, performance on the duration reproduction task and rhythm recall task was significantly better in the duration and rhythm conditions and declined in the two conditions in which participants attended to pitch information. In contrast, performance on the pitch recognition task for incoherent melodies was significantly higher in the pitch attending condition than in any of the other four conditions. In a final experiment, Boltz (1999, experiment 3) repeated the second experiment but varied the degree of familiarisation given before the main phase of the experiment. The results demonstrated that with small numbers of learning trials, performance was comparable for coherent and incoherent melodies on all three memory tasks. Memory for any given structural dimension was quite accurate when it was selectively attended to but declined when directed elsewhere or divided between other dimensions. With increasing numbers of learning trials, performance on incoherent melodies continued to show this trend whilst performance on coherent melodies showed less and less variation between attending conditions until it disappeared completely.

3.4 Testing the Predictions of GTTM

Palmer & Krumhansl (1987a,b) also found that the phrase judgements obtained in their experiments generally correlated well with the tonal stability (Krumhansl & Kessler, 1982) as well as the metric strength and time-span reduction (Lerdahl & Jackendoff, 1983) of the final tone of the stimulus. In these cases, the time-span reductions were based on detailed hand-crafted analyses (by one of the authors of GTTM) of the musical passages examined in the experiments. This strategy was also employed by Large *et al.* (1995) who found that the patterns of structural importance predicted by the time-span reductions and metrical hierarchies for children's melodies provided an excellent fit to cognitive representations of structural importance as assessed by empirical data on the events retained by trained pianists across improvised variations on these melodies.

Clarke & Krumhansl (1990) also indirectly examined the predictions of GTTM in a series of experiments examining large-scale segmentation in two pieces for piano composed by Stockhausen (experiments 1-3) and Mozart (experiments 4-6). In experiments 1 and 4, musically trained listeners were familiarised with the piece and then asked to indicate the location of large-scale segment boundaries by pressing a foot pedal while listening to the piece again. In experiments 2 and 5, different groups of musically trained participants were familiarised the piece before listening to the segments extracted from the piece in isolation and indicating on a horizontal line, representing the time-span of the piece, the start and end points of each segment. The participants in experiments 1 and 5 were additionally asked to describe the features of the music that helped formed each boundary. These features included: silences or pauses (cf GPR 2a); discontinuities in dynamics, register, texture, rhythm and tempo (cf GPR 3); changes in pitch content, contour or between vertical and horizontal organisation; changes of metre; and the repetition of material (cf GPR 6). Many of these features relate quite directly to the GPRs providing qualitative support for these rules as a model of segmentation strategies in music perception (or at least of those strategies used by trained musicians which are available for conscious introspection).

In other research, the predictions of GTTM (Lerdahl & Jackendoff, 1983) about the perception of segment boundaries in music has been examined in more specific detail. Deliège (1987), for example, conducted two experiments in which musicians and non-musicians were asked to indicate perceived segmentation boundaries in a range of musical excerpts. In the first experiment, 32 homophonic excerpts were taken from instrumental or orchestral works in the Western art repertoire ranging from the Baroque period to the early twentieth century and participants responded by reference to a line of dots matching the sounds in the upper voice of the excerpt. The segmentation task was performed after listening to the entire excerpt which could be reheard on request. Three between-subjects conditions were created by presenting each excerpt with no preceding context, with the actual context preceding the excerpt or with a preceding context consisting of another excerpt from the same piece or one from a related piece. The segmentation data were analysed with respect to GPRs 2 (proximity) and 3 (change) shown in Table 1. A comparison of the data for the musicians and non-musicians showed that the former responded significantly more often in accordance with these GPRs than the former.

Furthermore, the musicians responded significantly more often in accordance with the GPRs following the true context than in the other two context conditions whilst the responses of the non-musicians showed no such effect of context. A more detailed analysis demonstrated that there was no effect of training on responses in accordance with GPRs 2b (Attack-point), 3b (Dynamics change) and an additional rule Timbre change (based on a parenthetical remark in GTTM) but that the musicians responded in accordance with GPRs 2a (Slur/Rest), 3a (Register change), 3c (Articulation change) and 3d (Length change) significantly more often than the non-musicians. Both groups responded most often in accordance with GPR 2b (Attack-point) and least often in accordance with GPR 3c (Articulation). Deliège suggests that only the musicians were sensitive to rules pertaining to the performance of each piece by way of an account for the observed effects of musical training.

In a second experiment, Deliège (1987) asked musically trained and untrained participants to identify a single segment boundary in melodic sequences which could be segmented according to just one of two competing GPRs at adjacent locations. An additional rule was also examined which predicts segmentation at points of change of melodic contour since a large proportion of the incorrect responses of the non-musicians in the first experiment were justified in these terms. The location of the segment boundary varied between trials. A large proportion of the segmentations produced by both groups of participants were in accordance with GPRs (although significantly more so in the case of the musically trained group). A more detailed analysis indicated that both groups tended to prefer the second of the possible segmentation points for conflicts which appeared early in the sequence and the first for those that appeared later. Regarding conflicts between specific rules, the preferred segmentations were most often in accordance with Timbre Change followed by GPRs 3b (Dynamics change) and 1a (Slur/Rest) for the musicians, and GPRs 3a (Register Change) and 1a (Slur/Rest) for the non-musicians. The lowest proportions of segmentations were made in accordance with the additional contour rule and GPR 3d (Length Change) for both groups.

Peretz (1989) reports three experiments intended to examine the claim that the low-level grouping boundaries identified by the GPRs determine the boundaries represented at later stages of processing and higher levels of representation (Lerdahl & Jackendoff, 1983). All three experiments examined the perception of segment boundaries by musicians and non-musicians in French folk melodies each containing one boundary determined by either: a pitch skip (GPR 3a, Register Change); a length change (GPR 3d, Length change); or a length change combined with parallelism (GPR 6) in terms of individual tone durations and total duration. The first experiment used an online explicit segmentation task in which participants marked the location of the phrase boundary on a line of dots while listening to the melody (cf Deliège, 1987). The second experiment used an offline probe recognition task for probes of three tones which either crossed the phrase boundary in the melody or fell either side of it (cf Dowling, 1973; Tan *et al.*, 1981). Peretz hypothesised that both tasks should yield similar results if segment boundaries identified in online listening are maintained in memory after a melody has been processed. The result of Experiment 1 demonstrated that both musicians and non-musicians responded significantly more often in accordance with GPR 3d (Length change) and GPR 3d in combination with GPR 6 (Parallelism) than they did with GPR 3a (Register Change). Only non-musicians gave a greater proportion of conforming responses in the latter condition although the failure to find a difference for the musicians may have been due to a ceiling effect (the musicians responded significantly more in accordance with the predicted boundaries than the non-musicians). The results of Experiment 2, on the other hand, showed no effect of probe location (within or across), boundary type or training on the probe detection task.

Peretz (1989) suggested that the discrepancy between the results obtained using the two

GPR	Description	n	Boundary Strength
2a	Rest		absolute magnitude of rest (semibreve = 1.0)
2b	Attack-point	length	$\begin{cases} 1.0 - \frac{n_1+n_3}{2\times n_2} & \text{if } n_2 > n_3 \wedge n_2 > n_1 \\ \perp & \text{otherwise} \end{cases}$
3d	Length change	length	$1.0 - \begin{cases} n_1/n_3 & \text{if } n_3 \geq n_1 \\ n_3/n_1 & \text{if } n_3 < n_1 \end{cases}$
3a	Register change	pitch height	$\begin{cases} 1.0 - \frac{\lvert n_1-n_2\rvert+\lvert n_3-n_4\rvert}{2\times\lvert n_2-n_3\rvert} & \begin{array}{l}\text{if } n_2 \neq n_3 \wedge \\ \lvert n_2-n_3\rvert > \lvert n_1-n_2\rvert \wedge \\ \lvert n_2-n_3\rvert > \lvert n_3-n_4\rvert\end{array} \\ \perp & \text{otherwise} \end{cases}$

Table 2: The quantification of GPRs 2 and 3 by Frankland & Cohen (2004). See Table 1 for qualitative descriptions of the GPRs giving the meaning of the variables n_1, n_2, n_3 and n_4.

experimental tasks may be a result of a loss of critical information regarding the segmentation on the representations that are used in the offline probe recognition task. In order to test this hypothesis, Experiment 3 used an inverted version of the probe recognition task in which participants listened to a probe followed by the melody and were asked to indicate as quickly and accurately as possible whether the probe occurred in the melody. The analysis of the results indicated no effects of training but did yield a significant interaction between probe location and boundary type. Within probes were identified more accurately than across probes but only for boundaries formed according to GPR 3d (Length change) or by GPR 3d in combination with GPR 6 (Parallelism). There was no difference between these two boundary types suggesting that parallelism exerted little additional influence on probe recognition performance. Since Experiments 2 and 3 are directly comparable, differing only in that the former is offline and the latter online, Peretz argued that the representations yielded by cognitive processes based on GPR 3d which are operative during online segmentation of a melody may not persist in memory after the melody has been processed.

Peretz also made some observations regarding implicit and explicit perceptual segmentation. Since no effects of training were found in experiments 2 (implicit and online) and 3 (explicit but online), the effects of training found for the explicit offline judgements of experiment 1 (and in Deliège, 1987) may be a result of musicians being more skilled than non-musicians at explicitly examining their musical percept.

Frankland & Cohen (2004) argue that research on the GPRs has suffered from the fact that they have never been individually quantified. In the absence of a precise operational definition, it is hard to compare applications of: (i) the same rule at different locations; (ii) different rules at the same location; and (iii) different rules at different locations. Without quantification, for example, it is impossible to say whether the different utilities observed by Deliège (1987) for each of the rules actually resulted from the selection of stimuli exhibiting strong forms of one rule and weak versions of another. Similarly, the stimuli used by Peretz (1989) to exhibit GPR 3d (Length Change) could also have represented GPR 2b (Attack-point). In order to address these issues, Frankland & Cohen quantified GPRs 2a, 2b, 3a and 3d as shown in Table 2. Since a slur is a property of the interstimulus interval (ISI) whilst a rest is an absence of sound following a note, they argued that these two components of GPR 2a should be separated and, in fact, only quantified the rest aspect. Since GPRs 2a (Rest), 2b (Attack-point) and 3d (Length) concern perceived duration, they were based on linearly scaled time in accordance with results in psychoacoustic research (Allan, 1979). Finally, a natural result of the individual quantifications is that they can be combined using multiple regression to quantify the implication contained in GPR 4 (Intensification) that co-occurrences of two or more aspects of GPRs 2 and 3 lead to stronger boundaries.

Frankland & Cohen (2004) examined the predictions of the quantified GPRs in two experiments using an online task in which participants varying widely in their musical training indicated perceived boundaries in tonal melodies using a manual key press (cf Clarke & Krumhansl, 1990). In general, the participants were quite consistent across repetitions, being more consistent with familiar melodies and with increased repetition, and listeners tended to parse the melodies in a similar manner (an effect of training was only found for a melody taken from the classical repertoire).[2] The results indicated that GPR 2b (Attack-point) produced consistently strong correlations with the empirical boundary profiles while GPR 2a (Rest) was also significant in the one case where it applied. No empirical support was found for GPRs 3a (Register Change) and 3d (Length change). There were several instances of perceived boundaries were not predicted by any of the GPRs. Due to the construction of the stimuli, these *misses* could not have been accounted for by GPRs 2a (Slur), 3b (Dynamics change) or 3c (Articulation change). In fact, most corresponded to boundaries perceived after the third of a group of two notes of equal duration followed by two notes of equal (but longer) duration. This situation is covered neither by Attack-point, since $n_2 = n_3$, and Length change predicts a boundary after second tone.[3] Frankland & Cohen suggest that Attack-point be redefined such that it predicts a boundary after an event that is relatively longer than its two predecessors (thereby partially subsuming the function of length change).

3.5 Infant's Perception of Phrase Boundaries

Krumhansl & Jusczyk (1990) set out to examine the extent to which infants segment musical passages and on what basis they do so. Their stimulus materials consisted of the initial sections of 16 Mozart minuets each of which appeared in two versions: in *natural* versions, the musical passage was presented with two-beat pauses between structural phrases in the music; in *unnatural* versions, the pauses appeared in the same absolute temporal location but the musical passage was shifted such that the pauses appeared in the middle of phrases. In a visual preference experiment, 6- and $4\frac{1}{2}$-month-old infants oriented significantly longer on average to the natural versions than the unnatural versions. While this suggests that these infants preferred the natural versions, it remained possible that they had responded on the basis of differences between conditions in the beginnings of phrases. However, Jusczyk & Krumhansl (1993, experiment 1) replicated the result with $4\frac{1}{2}$-month-old infants and unnatural versions of the minuets in which the musical passage was unaltered and the pauses were shifted such that they occurred within phrases. Furthermore, Jusczyk & Krumhansl (1993, experiment 2) found no difference in orientation times to the original minuets (played without pauses) and the natural versions, demonstrating that infants treat the musical passages with pauses inserted at phrase boundaries much as they do unaltered versions of the same passages.

Krumhansl & Jusczyk (1990) also identified a number of structural differences between the natural and unnatural versions of the minuets and examined the relationship between quanti-

[2] Frankland & Cohen (2004) conducted additional experiments in which participants were subsequently asked to perform either a click-detection task (cf Stoffer, 1985) or a probe recognition task (cf Dowling, 1973; Peretz, 1989; Tan *et al.*, 1981) using the boundaries identified by each participant in the first stage of the study. The results on both tasks indicated that the boundaries identified had some validity. In contrast to the results of Peretz (1989), however, no major differences appeared between the tasks.

[3] Deliège (1987, experiment 1) noted a similar situation with respect to Length change for one of her stimuli.

tative variables based on these differences and orientation time. In particular, they found that phrase endings in the natural versions (but not in the unnatural versions) tended to be characterised by falling pitch patterns and a lengthening of note durations. Furthermore, Krumhansl & Jusczyk observed a significant negative correlation between orientation time and the average pitch height of the penultimate and final tones before the pauses in each stimulus as well as a significant positive correlation between orientation time and the average duration of the penultimate and final tones before the pauses in each stimulus. These results were replicated by Jusczyk & Krumhansl (1993, experiment 1) and Jusczyk & Krumhansl (1993, experiment 3) demonstrated that this was not simply a result of discontinuities across phrase boundaries caused by the falling pitch and lengthened durations before the pause; Four$\frac{1}{2}$-month-old infants oriented significantly longer to the natural versions than to reversed natural versions suggesting that the direction of changes in these dimensions between the pauses is more important than the simple presence of a discontinuity. Further evidence to support this conclusion was obtained by Jusczyk & Krumhansl (1993, experiment 4) who found no difference in orientation times to reversed naturally segmented versions, which exhibited discontinuities in pitch and duration before pauses (but not phrase-final descending pitch intervals and lengthening durations), and reversed unnaturally segmented versions, which did not. Finally, Jusczyk & Krumhansl (1993, experiment 5), found no significant difference in orientation times to the original versions and the reversed original versions, suggesting that the previous findings could not be accounted for by a general preference for the forward versions relative to the reversed versions of the minuets. In these experiments, there was also some evidence for a relationship between orientation times and the proportion of harmonic intervals forming an octave directly preceding a pause in each sample (Jusczyk & Krumhansl, 1993; Krumhansl & Jusczyk, 1990).

3.6 Statistical Learning of Tone Sequences

Saffran *et al.* (1999) report three experiments which examined the role of statistical learning in the segmentation of pure tone sequences using a grammar learning task. The stimuli were constructed from 11 pure tones of the same octave which were combined into groups of three to produce *tone words*. These were then concatenated together randomly, with no pauses or other acoustic markers between words, to create the tone sequences used in the experiments. Since the goal was to examine learning, the tones themselves and the tone words were carefully chosen to avoid creating tonal contexts. Crucially, the only consistent cue to the boundaries between tone words was the first-order transition probability between tones such that the transition probabilities within tone words were high and those between tone words were low. In the first experiment, adult non-musicians listened three times to the same tone sequence and subsequently undertook a two-alternative forced-choice task in which they were presented with a word from the artificial language and a non-word constructed from the same alphabet and asked to identify the most familiar of the two. The results demonstrated that performance was significantly greater than chance and that performance was better for tone words with higher transitional probabilities.

In the second experiment, Saffran *et al.* extended the findings of experiment 1 using a more difficult task involving a forced-choice between a tone word and a part word which is a legitimate tone word with the first or last tone changed to make it a non-word. Although performance was lower than in the first experiment (due to the additional demands of the task), it was still significantly above chance. In a final experiment, Saffran *et al.* (1999) studied the performance of 8-month-old infants on a similar task in which visual orientation times were recorded as an indication of familiarity. The results demonstrated significantly longer orientation times to part-words than to words. In all three experiments, the participants were assigned to two groups each of which learned a distinct set of tone words with non-words (or part-words) corresponding to the tone words of the other group. Since there was no significant effect of group, the results cannot be an artefact of uncontrolled differences between the words and non-words.

It is possible, however, that the participants in these experiments were responding on the basis of regularities in pitch interval rather than absolute pitch. Although Saffran *et al.* (1999) cannot completely rule out this possibility, the distribution of intervals in the two tone languages were similar. Furthermore, in the second experiment, there was no difference in the proportion of incorrect responses on trials with and without part-words containing novel intervals (intervals which did not occur within valid words) and there was no evidence that performance was consistently better for tone words containing more frequent intervals. In subsequent research, Saffran & Griepentrog (2001) attempted to distinguish absolute and relative pitch processing using two sets of stimuli: in the first (AP condition), absolute pitch transition probabilities were the only available cue for segmentation and words and part-words differed only in terms of absolute pitch and not pitch interval; in the second (RP condition), pitch interval transition probabilities were the only available cue for segmentation and part-words consisted of a pair of unfamiliar intervals corresponding to a two pairs of familiar pitches. Eight-month-old infants demonstrated performance significantly above chance in the AP condition but not in the RP condition while adults exceeded chance in the RP condition but not the AP condition (when participants with absolute pitch were removed from the analysis). Saffran (2003) replicated this pattern of results using tones drawn from a diatonic rather than a chromatic alphabet and found that performance was generally better with diatonic than chromatic materials.

Saffran *et al.* (2005) suggested that the failure of the infants to learn in the RP condition might be a result of a general bias towards absolute pitch structure since the part-words in this condition were constructed from overlapping pairs of familiar pitches. They examined this hypothesis by constructing part-words which were transpositions of those used by Saffran & Griepentrog (2001) and therefore entirely novel in terms of absolute pitch. The results demonstrated performance significantly above chance (for infants and adults) suggesting that infants are capable of learning on the basis of interval structure but only when absolute pitch cues are absent.

Tillmann & McAdams (2004) reports two experiments with adult listeners which followed the experimental design Saffran *et al.* (1999) except that the sequences were constructed using sounds which were constant in pitch but differed in timbre. Crucially, the construction of the timbres allowed the systematic manipulation of the acoustic similarity of adjacent sounds in a timbre space. Three conditions were created in which timbral similarity either supported, contradicted or were neutral with respect to the grouping of sounds on the basis of transition probabilities. In contrast to Saffran *et al.* (1999), only one language was used and the responses of the trained participants were compared to a control group with no prior experience of the timbre sequence. In the first experiment, participants had to discriminate on the basis of familiarity words and non-words.

The performance of trained participants exceeded that of the control group by a consistent amount in all three conditions of congruity between statistical and acoustic cues. For both groups, performance was greater in the congruent condition than in the incongruent condition with performance in the neutral condition falling in between these extremes. The failure to find an interaction between learning and congruity condition indicates a general bias to segment on the basis of timbral similarity but also that this bias does not interact with inductive learning which occurs even when transition probabilities conflict with acoustic similarity. Experiment 2 replicated these findings using the more difficult task of discriminating words and part-words and two congruence conditions (congruent and neutral). Although the demands of the task were reflected in generally lower performance, although the performance of the learning group was greater than that of the control group and above chance in both congruence conditions. By contrast, the performance of the control group was only above chance in the congruent condition as in the first experiment.

Creel *et al.* (2004) set out to examine whether the ability to learn statistical dependencies in tone sequences extends to dependencies between non-adjacent events. This was achieved by concatenating triplets of tones into sequences such that the transition probabilities within triplets was higher than that between triplets (1.0 and 0.5 respectively) and then interleaving this sequence with another such that the transition probabilities between adjacent tones was 0.5. The participants were divided into two groups according to the discrimination task performed. In the first task, participants had to discriminate nonadjacent triplets with nonadjacent nontriplets (whose withing triplet transition probabilities were zero). In the second task, participants had to discriminate adjacent sextuplets from non-sextuplets (which had never occurred in training). The results demonstrated chance performance in the nonadjacent test condition but significantly greater than chance performance in the adjacent test condition.

Subsequent experiments demonstrated that listeners were capable of inducing nonadjacent statistical regularities when the temporally nonadjacent triplets were distinguished from the interleaved sequence in term of register (experiment 2) or timbre (experiment 3). In these experiments, chance performance was observed in the adjacent test condition. In a final experiment, Creel *et al.* showed that with moderate timbral cues, both adjacent and nonadjacent dependencies could be learned although performance was lower than in the previous experiments. On the basis of this result, they argue that the improvement yielded by registral and timbral cues in these experiments represent an interaction between these cues and statistical dependencies during learning rather than being simply a result of stream segregation blocking temporal adjacency information. However, the results of Tillmann & McAdams (2004) warrant caution in drawing such a conclusion in the absence of a control group with no learning experience.

3.7 The Independence of Meter and Grouping: Neuroscientific Evidence

In GTTM, although metric and grouping hierarchies interact to yield a time-span reduction, each is constructed in a relatively independent manner: metric accent structures do not posses any inherent grouping and groups do not imply a particular pattern of metrical accent. Some evidence for a distinction between the psychological processing of metre and grouping comes from research on the cognitive neuropsychology of music.

In a study of unilaterally brain damaged patients, Peretz (1990) found that lesions in either hemisphere impaired performance in discriminating melodies differing only in terms of rhythmic structure (by interchanging the rhythmic values of adjacent notes in order to interrupt grouping on the basis of temporal proximity) whilst performance on a task intended to reflect processing of metric structure (discriminating marches in 4/4 and waltzes in 3/4 metre) remained unimpaired. The lesions were too coarse to infer the precise cortical regions contributing to the two tasks. In a subsequent study, Liegeoise-Chauvel *et al.* (1998) found the opposite pattern of impairment in patients with lesions in the anterior temporal gyrus, providing a double dissociation between the processing of rhythmic grouping and metre.

A recent fMRI study reported by Brochard *et al.* (2000) examined brain activity during two rhythmic pattern discrimination tasks. In the first task, the patterns were metrically irregular and in some trials one test item had an event displaced from one temporal group to another;

in the second task, the patterns were metrically regular and in some trials one test item had an event displaced (within its group) in order to interrupt the regular metric pulses. While there was a large overlap in the areas activated in both tasks, certain areas were exclusively activated by each task (including the anterior temporal gyrus for the metric task).

4 Computational Models

4.1 Gestalt Based Models

Many efforts to construct cognitive models of the perception of grouping structure in music are based on the idea that boundaries are perceived at points where the Gestalt principles of proximity or similarity are violated with respect to some dimension of the musical surface. Tenney & Polansky (1980), for example, present a model which predicts boundaries in a melody between elements whose distance from the previous element is greater than the inter-element distances immediately preceding and following it. The elements and distances in question are determined by the musical parameter and the hierarchical level of interest. In their model, Tenney & Polansky consider the parameters (and distances) duration (inter-onset interval), pitch (pitch interval) and intensity (intensity difference). Boundary strengths computed on the basis of these parameters are combined using a weighted sum of the absolute distances where the relative weights are set by trial and error for each individual composition examined. Boundary strengths between non-atomic elements (i.e., sequences of atomic events) are computed using a weighted mean of the distances between the mean parametric values of the segments at that hierarchical level and at all lower levels. The weights are computed such that lower level boundaries contribute less than higher level boundaries. The implemented system was used to generated structural grouping analyses for three pieces: Varèses *Density 21.5*; Webern's Concerto, Op. 24, 2nd movement; and Debussy's *Syrinx*. For the first two pieces, the analyses were found to be very similar to analyses published in the music-theoretic literature.

In contrast to the Gestalt-based approach of Tenney & Polansky, where boundaries are associated with intervals that are larger than their neighbours, Cambouropoulos (1998, 2001) proposes a more general model in which boundaries are associated with any local change in interval magnitudes. The *Local Boundary Detection Model* (LBDM) consists of a *change* rule, which assigns boundary strengths in proportion to the degree of change between consecutive intervals, and a *proximity* rule, which scales the boundary strength according to the size of the intervals involved. In its most recent form (Cambouropoulos, 2001), the LBDM operates over several independent parametric melodic profiles $P_k = [x_1, x_2, \ldots, x_n]$ where $k \in \{pitch, ioi, rest\}, x_i > 0, i \in \{1, 2, \ldots, n\}$ and the boundary strength at interval x_i is given by:

$$s_i = x_i \times (r_{i-1,i} + r_{i,i+1}) \tag{1}$$

where the degree of change between two successive intervals:

$$r_{i,i+1} = \begin{cases} \frac{|x_i - x_{i+1}|}{x_i + x_{i+1}} & \text{if } x_i + x_{i+1} \neq 0 \wedge x_i, x_{i+1} \geq 0 \\ 0 & \text{if } x_i = x_{i+1} = 0. \end{cases} \tag{2}$$

For each parameter k, the boundary strength profile $S_k = [s_1, s_2, \ldots, s_n]$ is calculated and normalised in the range $[0, 1]$. A weighted sum of the boundary strength profiles is computed using weights derived by trial and error (0.25 for $pitch$ and $rest$, and 0.5 for ioi), and boundaries are predicted where the combined profile exceeds a predefined threshold.

Cambouropoulos (2001) evaluated the LBDM in a series of three experiments. In the first, it was found that a large proportion of the predicted boundaries corresponded to those marked on a score by a musician (between 63% and 74% depending on the threshold and weights used) although the LBDM also produced a large proportion of spurious boundaries (between 45% and 49%). In a second experiment, Cambouropoulos (2001) examined the correspondence between the LBDM boundary profiles and the expressive lengthening of note durations in pianists' performance of seven Mozart piano sonatas. It was found that notes corresponding to supra-threshold boundaries as well as those corresponding to local peaks in the boundary profile were more often lengthened than shortened while notes corresponding to sub-threshold boundaries were lengthened and shortened equally often. The third experiment replicated these findings for 22 performances of the first 20 bars of Chopin's *Etude Op. 10, no. 3*. It was also found that the penultimate notes in the predicted groups were consistently lengthened.

Temperley (2001) introduces a model of grouping called *Grouper* which accepts a melody, in which each note is represented in terms of its onset time, off time, chromatic pitch and level in a metric hierarchy, and returns a single, exhaustive partitioning of the melody into non-overlapping groups. The model operates through the application of three *Phrase Structure Preference Rules* (PSPRs):

PSPR 1 (Gap Rule): prefer to locate phrase boundaries at (a) large inter-onset intervals (IOI) and (b) large offset-to-onset intervals (OOI), calculated as the sum of the IOI and OOI divided by the mean IOI of all previous notes;

PSPR 2 (Phrase Length Rule): prefer phrases to have roughly 8 notes, achieved by penalising predicted phrases by $|(\log_2 N) - 3|$ where N is the number of notes in the predicted phrase;

PSPR 3 (Metrical Parallelism Rule): prefer to begin successive groups at parallel points in the metrical hierarchy.

The first of these rules is another example of a Gestalt principle of temporal proximity (e.g., GPR 2 in Table 1), the second was determined through an empirical investigation of the typical phrase lengths in a collection of folk songs, and the third is related to GPR 6 in Table 1. The best analysis of a given piece is computed offline using a dynamic programming approach where candidate phrases are evaluated according to a weighted combination of the three rules. The weights were determined through trial and error. By way of evaluation, Temperley used Grouper to predict the phrase boundaries marked in 65 melodies from the Essen Folk Song Collection (Schaffrath, 1992, 1994). The model correctly identified 75.5% of transcribed phrase boundaries (although it also generated a significant number of false positives).

Thom *et al.* (2002) compared the predictions of the LBDM (Cambouropoulos, 2001) and Grouper (Temperley, 2001) to segmentations at the phrase and subphrase level provided by 19 musical experts for 10 melodies in a range of styles. For each melody, the correspondence between the boundary profile of each model and that provided by each expert was calculated using an F score (Manning & Schütze, 1999):

$$F = \frac{1}{1 + \frac{fn+fp}{2 \times tp}} \in [0, 1] \tag{3}$$

where fn is the number of false negatives, fp is the number of false positives and tp is the number of true positives. The performance of each model on each melody was estimated by averaging the F scores over the 19 experts. In a first experiment, Thom *et al.* examined the average F scores between experts for each melody, obtaining values ranging between 0.14 and

0.82 for phrase judgements and 0.35 and 0.8 for subphrase judgements. The higher consistencies tended to be associated with melodies whose phrase structure was emphasised by rests. In a second experiment, each model was examined with parameters optimised for each individual melody. The results indicated that Grouper tended to outperform the LBDM. Inter-onset interval was an important factor in the success of both models. In a third experiment, the predictions of each model was compared with the transcribed boundaries in several datasets from the Essen Folk Song Collection (Schaffrath, 1992, 1994). The model parameters were optimised over each dataset and the results indicated that Grouper outperformed the LBDM. Finally, in order to examine the stability of the two models, each was used to predict the expert boundary profiles using parameters optimised over the Essen data. The performance of both algorithms was impaired, most notably for the subphrase judgements of the experts.

4.2 Machine Learning Models

Large *et al.* (1995) have examined the ability of Recursive Auto-Associative Memory (RAAM) to acquire reduced representations of Western children's melodies represented as tree structures according to music-theoretic predictions (the time-span reductions of Lerdahl & Jackendoff, 1983). An auto-associative neural network is simply one which is trained to reproduce on its output layer a pattern presented to its input layer, generally forming a compressed representation of the input on its hidden layer. For example, training a network with eight-unit input and output layers separated by a three-unit hidden layer with the eight 1-bit-in-8 patterns typically results in a 3-bit binary code on the hidden units (Rumelhart *et al.*, 1986). Pollack (1990) introduced an extension of auto-association called Recursive Auto-Associative Memory (RAAM) which is capable of learning fixed-width representations for compositional tree structures through repeated compression. The RAAM architecture consists of two separate networks: first, an encoder network which constructs a fixed-dimensional code by recursively processing the nodes of a symbolic tree from the bottom up; and second, a decoder network which recursively decompresses this code into its component parts until it terminates in symbols, thus reconstructing the tree from the top down. The two networks are trained in tandem as a single auto-associator.

Large *et al.* (1995) found that trained RAAM models acquired compressed representations of the melodies in which structurally salient events are represented more efficiently (and reproduced more accurately) than other events. Furthermore, the trained network showed some ability to generalise beyond the training examples to variant and novel melodies although, in general, performance was affected by the depth of the tree structure used to represent the input melodies with greater degrees of hierarchical nesting leading to impaired reproduction of input melodies. However, the certainty with which the trained network reconstructed events correlated well with music-theoretic predictions of structural importance (Lerdahl & Jackendoff, 1983) and cognitive representations of structural importance as assessed by empirical data on the events retained by trained pianists across improvised variations on melodies.

Bod (2001) argues for a memory-based approach to modelling melodic grouping structure as an alternative to the Gestalt-based approach. Bod uses grammar learning techniques to induce the annotated phrase structure of the Essen Folk Song Collection (Schaffrath, 1992, 1994) and test the trained models on a held-out test set using F scores (Manning & Schütze, 1999) as an evaluation metric. Three grammar induction algorithms are examined: first, the treebank grammar learning technique which read all possible context free rewrite rules from the training set and assigns each a probability proportional to its relative frequency in the training set; second, the Markov grammar technique which assigns probabilities to context free rules by decomposing the rule and its probability by an nth order Markov process, allowing the model to estimate the probability of rules that have not occurred in the training set; and third, a Markov grammar augmented with a Data-oriented parsing (DOP) method for conditioning the probability of a rule over the rule occurring higher in the parse tree. A best-first parsing algorithm based on Viterbi optimisation was used to generate the most probable parse for each melody in the test set given each of the three models. The results demonstrated that the treebank technique yielded moderately high precision but very low recall ($F = 0.065$), the Markov grammar yielded slightly lower precision but much higher recall ($F = 0.706$) while the Markov-DOP technique yielded the highest precision and recall ($F = 0.810$). A qualitative examination of the folk song data reveals a number of cases (15% of the phrase boundaries in the test set) where the annotated phrase boundary cannot be accounted for by Gestalt principles but that are predicted by the Markov-DOP parser.

The models examined by Large *et al.* (1995) and Bod (2001) require the training data to be annotated with phrase structure information. In contrast, Ferrand (2005) has examined an unsupervised learning approach to modelling perceived grouping structure in melodic music.

4.3 Grouping from the Acoustic Signal

- Todd (1994)
- Šerman & Griffith (2004)
- Gjerdingen (1999)
- Abdallah *et al.* (2006)

4.4 Other Models

5 General Discussion

References

Abdallah, S., Sandler, M., Rhodes, C., & Casey, M. (2006). Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, *65*(2-3), 485–515.

Allan, L. G. (1979). The perception of time. *Perception and Psychophysics*, *26*(5), 340–354.

Bod, R. (2001). Memory-based models of melodic analysis: Challenging the Gestalt principles. *Journal of New Music Research*, *30*(3), 27–37.

Boltz, M. G. (1989a). Perceiving the end: Effects of tonal relationships on melodic completion. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 749–761.

Boltz, M. G. (1989b). Time judgements of musical endings: Effects of expectancies on the 'filled interval effect'. *Perception and Psychophysics*, *46*(5), 409–418.

Boltz, M. G. (1991). Some structural determinants of melody recall. *Memory and Cognition*, *19*(3), 239–251.

Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception and Psychophysics*, *53*(6), 585–600.

Boltz, M. G. (1999). The processing of melodic and temporal information: Independent or unified dimensions? *Journal of New Music Research*, *28*(1), 67–79.

Boltz, M. G. & Jones, M. R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, *18*(4), 389–431.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, *3*, 294–301.

Brochard, R., Dufour, A., Drake, C., & Scheiber, C. (2000). Functional brain imaging of rhythm perception. In C. Woods, G. Luck, R. Brochard, F. Seddon, & J. A. Sloboda (Eds.), *Proceedings of the Sixth International Conference of Music Perception and Cognition*. Keele, UK: University of Keele.

Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. Doctoral dissertation, The University of Edinburgh, Faculty of Music and Department of Artificial Intelligence.

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference* (pp. 17–22). San Francisco: ICMA.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation: A computational approach. *Music Perception*, *23*(3), 249–269.

Clarke, E. F. & Krumhansl, K. L. (1990). Perceiving musical time. *Music Perception*, *7*(3), 213–252.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*(5), 1119–1130.

Cross, I. (1995). Review of *The analysis and cognition of melodic complexity: The implication-realization model*, Narmour (1992). *Music Perception*, *12*(4), 486–509.

Deliège, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's grouping preference rules. *Music Perception*, *4*(4), 325–360.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception and Psychophysics*, *28*(5), 381–389.

Deutsch, D. & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, *88*(6), 503–522.

Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception and Psychophysics*, *14*(1), 37–40.

Ferrand, M. (2005). *Data-driven, Memory-based Computational Models of Human Segmentation of Musical Melody*. Doctoral dissertation, School of Arts, Culture and Environment, University of Edinburgh, UK. Submitted.

Frankland, B. W. & Cohen, A. J. (2004). Parsing of melody: Quantification and testing of the local grouping rules of Lerdahl and Jackendoff's *A Generative Theory of Tonal Music*. *Music Perception*, *21*(4), 499–543.

Gjerdingen, R. O. (1999). Apparent motion in music? In N. Griffith & P. M. Todd (Eds.), *Musical Networks: Parallel Distributed Perception and Performance* (pp. 141–173). Cambridge, MA: MIT Press/Bradford Books.

Gregory, A. H. (1978). Perception of clicks in music. *Perception and Psychophysics*, *24*(2), 171–174.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jones, M. R., Boltz, M. G., & Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Perception and Psychophysics*, *32*(3), 211–218.

Jusczyk, P. W. & Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infant's sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(3), 627–640.

Krumhansl, C. L. (1995). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, *17*, 53–90.

Krumhansl, C. L. & Jusczyk, P. W. (1990). Infant's perception of phrase structure in music. *Psychological Science*, *1*(1), 70–73.

Krumhansl, C. L. & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368.

Large, E. W., Palmer, C., & Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science*, *19*(1), 53–96.

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

Liegeoise-Chauvel, C., Peretz, I., Babai, M., Laguitton, V., & Chauvel, P. (1998). Contribution of different cortical areas in the temporal lobes to music processing. *Brain*, *121*(10), 1853–1867.

London, J. (1997). Lerdahl and Jackendoff's strong reduction hypothesis and the limits of analytical description. *In Theory Only*, *13*(1–4), 3–29.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-realisation Model*. Chicago: University of Chicago Press.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication-realisation Model*. Chicago: University of Chicago Press.

Palmer, C. & Krumhansl, C. L. (1987a). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 116–126.

Palmer, C. & Krumhansl, C. L. (1987b). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing and familiarity. *Perception and Psychophysics*, *41*(6), 505–518.

Peretz, I. (1989). Clustering in music: An appraisal of task factors. *International Journal of Psychology*, *24*(2), 157–178.

Peretz, I. (1990). Processing of local and global musical information by unilateral brain-damaged patients. *Brain*, *113*(4), 1185–1205.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1), 77–105.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations through error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, volume 1 (pp. 25–40). Cambridge, MA: MIT Press.

Saffran, J. R. (2003). Absolute pitch in infancy and adulthood: The role of tonal structure. *Developmental Science*, *6*(1), 37–49.

Saffran, J. R. & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74–85.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., Reeck, K., Niebuhr, A., & Wilson, D. (2005). Changing the tune: The structure of the input affects infant's use of absolute and relative pitch. *Developmental Science*, *8*(1), 1–7.

Schaffrath, H. (1992). The ESAC databases and MAPPET software. *Computing in Musicology*, *8*, 66.

Schaffrath, H. (1994). The ESAC electronic songbooks. *Computing in Musicology*, *9*, 78.

Schmuckler, M. A. & Boltz, M. G. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception and Psychophysics*, *56*(3), 313–325.

Sloboda, J. A. & Gregory, A. H. (1980). The psychological reality of musical segments. *Canadian Journal of Psychology*, *34*(3), 274–280.

Stoffer, T. H. (1985). Representation of phrase structure in the perception of music. *Music Perception*, *3*(2), 191–220.

Tan, N., Aiello, R., & Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Memory and Cognition*, *9*(5), 533–539.

Temperley, D. (2001). *The Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press.

Tenney, J. & Polansky, L. (1980). Temporal Gestalt perception in music. *Contemporary Music Review*, *24*(2), 205–241.

Thom, B., Spevak, C., & Höthker, K. (2002). Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the 2002 International Computer Music Conference*. San Francisco: ICMA.

Thompson, W. F. (1996). Eugene Narmour: *The analysis and cognition of basic musical structures* (1990) and *The analysis and cognition of melodic complexity* (1992): A review and empirical assessment. *Journal of the American Musicological Society*, *49*(1), 127–145.

Tillmann, B. & McAdams, S. (2004). Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustic (dis)similarities. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*(5), 1131–1142.

Todd, N. P. M. (1994). The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research*, *23*(1), 25–70.

Šerman, M. & Griffith, N. J. L. (2004). Investigating melodic segmentation through the temporal multi-scaling framework. *Musicæ Scientiæ, Special Tenth Anniversary issue on musical creativity*.