Perception, 2010, volume 39, pages 1367 – 1391

doi:10.1068/p6507

The role of expectation and probabilistic learning in auditory boundary perception: A model comparison

Marcus T Pearce, Daniel Müllensiefen, Geraint A Wiggins

Centre for Cognition, Computation and Culture, Goldsmiths, University of London, London SE14 6NW, UK; e-mail: m.pearce@gold.ac.uk

Received 4 July 2009, in revised form 26 August 2010

Abstract. Grouping and boundary perception are central to many aspects of sensory processing in cognition. We present a comparative study of recently published computational models of boundary perception in music. In doing so, we make three contributions. First, we hypothesise a relationship between expectation and grouping in auditory perception, and introduce a novel information-theoretic model of perceptual segmentation to test the hypothesis. Although we apply the model to musical melody, it is applicable in principle to sequential grouping in other areas of cognition. Second, we address a methodological consideration in the analysis of ambiguous stimuli that produce different percepts between individuals. We propose and demonstrate a solution to this problem, based on clustering of participants prior to analysis. Third, we conduct the first comparative analysis of probabilistic-learning and rule-based models of perceptual grouping in music. In spite of having only unsupervised exposure to music, the model performs comparably to rule-based models based on expert musical knowledge, supporting a role for probabilistic learning in perceptual segmentation of music.

1 Introduction

Grouping and boundary perception are central to the understanding and modelling of core tasks in many areas of cognitive science. They are fundamental processes in, for example, natural language processing (eg speech segmentation and word discovery—Brent 1999b; Jusczyk 1997), motor learning (eg identifying behavioural episodes—Reynolds et al 2007; Newtson 1973), memory storage and retrieval (eg chunking—Kurby and Zacks 2007) and visual perception (eg analysing spatial organisation—Marr 1982). Our focus in this paper is on the perception and cognition of music (Krumhansl 1990; Temperley 2001), where the process by which the human perceptual system groups sequential musical elements together is one of the most fundamental issues.

In particular, 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* (Lerdahl and Jackendoff 1983) or *segmentation* (Cambouropoulos 2006). We distinguish this kind of perceptual aggregation of auditory elements from the integration, or *fusion*, of auditory elements that occur simultaneously in time and also from the segregation of parallel auditory *streams* (Bregman 1990). In musical terms, the kinds of groups we consider correspond with motifs, phrases, sections and other aspects of musical form. We use the term *grouping structure* to refer to a piece of music structured in this way. It is taken that, just as speech is perceptually segmented into phonemes, and then words which subsequently provide the building blocks for the perception of phrases and complete utterances (Brent 1999b; Jusczyk 1997), motifs or phrases in music are identified by listeners, stored in memory and made available for inclusion in higher-level structural groups (Lerdahl and 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 demands on memory.

Grouping structure is generally agreed to be logically independent from metrical structure (Lerdahl and Jackendoff 1983). By metrical structure, we mean the hierarchical pattern of cyclically recurring accented notes that one might tap one's foot along to (such that, for example, the first beat of a bar might be a stronger metrical accent than the third, which might be stronger than the second, and so on). 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 structures are often intimately related, and both are likely to serve as inputs to the processing of more complex musical structures (Lerdahl and 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).

In this paper, we propose the hypothesis that grouping in auditory perception is determined by perceptual expectations for auditory events. To test the hypothesis, we first take an existing cognitive model of musical expectation, based on probabilistic learning, and identify grouping boundaries at points of perceptual unexpectedness predicted by the model. We focus on musical melody, the cognitive task being to gather notes together into sequential groups. Second, we collect boundary indications from a group of listeners for fifteen melodies, deemed to have complex grouping structure, and with lyrics, orchestration, and expressive timing and dynamics to restrict the potential influences on perceived grouping to the pitch and timing of notes. As a result, several of the melodies have ambiguous grouping structure affording multiple interpretations which leads to low inter-participant agreement. We cope with this ambiguity by identifying subgroups of participants sharing the same grouping interpretation of each melody. Finally, we test our hypothesis by using the output of the probabilistic model to predict the listeners' responses and comparing its performance with several existing models in the literature.

The cognitive model of expectation has been shown to provide an accurate model of listeners' pitch expectations in melody (Pearce et al 2010b; Pearce and Wiggins 2006). In one experiment, for example, the probabilistic model accounted for 85% of the variance in listener's expectations elicited by fragments of British folk songs (Pearce and Wiggins 2006). It consistently predicts listeners' pitch expectations better than the best-performing rule-based model (Schellenberg 1997) in the literature. Here, we use the model to predict perceptual grouping boundaries at points of perceived unexpectedness and uncertainty, concepts which we formalise quantitatively in information-theoretic terms (Attneave 1959; Barlow 1959; Shannon 1948). In contrast to cognitive models consisting of hard-wired, domain-specific rules, the probabilistic model of melodic expectation and segmentation learns about the structure of an auditory environment through experience and therefore has the potential to account for the acquisition and development of these cognitive processes and to generalise naturally to expectations and sequential segmentation in different musical styles, auditory domains, or sensory modalities. The model learns in an unsupervised manner; it is never told explicitly where boundaries occur. As with word boundaries in speech, there is little evidence that musical boundaries are explicitly marked for the listener; and listeners readily perceive boundaries in the absence of explicit markers (Saffran et al 1999).

In a previous study, Pearce et al (2010a) evaluated the probabilistic segmentation model's ability to correctly predict the boundaries marked by a musicologist in a collection of 1705 German folk songs and compared the model's performance to nine existing grouping models. Here we extend this work to evaluate the model's efficacy in predicting perceived boundaries collected from twenty-five musically trained participants who provided explicit segmentations of fifteen melodies.

2 Background

2.1 *Theoretical approaches*

2.1.1 *A generative theory of tonal music.* Traditional theories of melodic grouping associate boundaries with local discontinuities or changes between events in terms of temporal proximity, pitch, duration, and dynamics. Perhaps the best known examples are the grouping preference rules (GPRs) of the Generative Theory of Tonal Music (GTTM—Lerdahl and Jackendoff 1983), which was inspired by Chomskian linguistics. The most widely studied of the GPRs predict that phrase boundaries will be perceived between two melodic events whose temporal proximity is less than that of the immediately neighbouring events due to a slur, a rest (GPR 2a), or a relatively long interval between the onset of one note and the onset of the next (inter-onset interval or IOI) (GPR 2b), or when the transition between two events involves a greater change in register (GPR 3a), dynamics (GPR 3b), articulation (GPR 3c), or duration (GPR 3d) than the immediately neighbouring transitions. These local GPRs were directly inspired by the principles of proximity (GPR 2) and similarity (GPR 3) developed to account for figural grouping in visual perception by the Gestalt school of psychology (eg Koffka 1935). GPR 4 states that where the effects of GPRs 2 and 3 are relatively more pronounced, a larger level grouping boundary may be placed. GPR 6 predicts that grouping boundaries are perceived in accordance with musical parallelism (eg at the same point in the bar or at similar points with respect to a repeated motif). The GPRs of GTTM have inspired some of the computational models of grouping reviewed in section 2.3.

2.1.2 *The implication-realisation theory.* Narmour (1990) presents the implication-realisation (IR) theory of music cognition which, like GTTM, is intended to be general (although the initial presentation was restricted to melody). However, while GTTM operates statically on an entire piece of music, the IR theory emphasises the dynamic processes involved in perceiving music as it occurs in time. The theory posits two distinct perceptual systems: the bottom–up system is held to be hard-wired, innate and universal, while the top–down system is held to be learnt through musical experience. In the bottom–up system, sequences of melodic intervals vary in the degree of *closure* that they convey. An interval which is unclosed is said to be an *implicative interval* and generates expectations for the following interval, termed the *realised interval*. The expectations generated by implicative intervals for realised intervals are described by Narmour (1990) in terms of several principles of continuation which are, again, influenced by the Gestalt principles of proximity, similarity, and good continuation. Strong closure, however, 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. The IR theory provides the musicological background for hypothesised link between expectation and grouping in perception and the cognitive models based on this hypothesis (see section 2.4).

2.2 *Experimental paradigms and results*

Early studies of musical segmentation (Gregory 1978; Sloboda and Gregory 1980; Stoffer 1985) provided basic evidence that listeners perceptually organise melodies into structural groups, using a click localisation paradigm adapted from research on perceived phrase structure in spoken language (Fodor and Bever 1965; Ladefoged and Broadbent 1960). More recently, two kinds of experimental tasks have been used to study perceptual grouping in music.

The first is a short-term memory recognition paradigm introduced by Dowling (1973), based on studies of phrase perception in language (Bower 1970; Waugh and Norman 1965). In a typical experiment, listeners are first presented with a musical stimulus containing one or more hypothesised boundaries before being presented with a short excerpt (the probe) and asked to indicate whether it appeared in the stimulus.

The critical probes either border on or straddle a hypothesised boundary and it is expected that owing to perceptual grouping, the former will be recalled more accurately or efficiently than the latter. Dowling's original experiment demonstrated that silence contributes to the perception of melodic segment boundaries (cf GPR 2a). Using the same paradigm, Tan et al (1981) demonstrated the influence of harmonic closure, especially for musicians.

In the second paradigm, participants provide explicit judgments of boundary locations while listening to the musical stimulus. The indicated boundaries are subsequently analysed to discover which principles guide perceptual segmentation. Using this approach with short musical excerpts, Deliège (1987) found that musicians and (to a lesser extent) non-musicians identify segment boundaries in accordance with the GPRs of GTTM (Lerdahl and Jackendoff 1983), especially those relating to rests or long notes and changes in timbre or dynamics. These factors were also reflected in large-scale segmentation by musically trained listeners of piano works composed by Stockhausen and Mozart (Clarke and Krumhansl 1990). Frankland and Cohen (2004) collected explicit boundary judgments from participants listening to six melodies (nursery rhymes and classical themes) and compared these to the boundaries predicted by quantitative implementations of GPRs 2a, 2b, 3a, and 3d (see table 1). The results indicated that GPR 2b produced consistently strong correlations with the empirical boundary profiles, while GPR 2a also received support in the one case where it applied. No empirical support was found for GPRs 3a and 3d. However, several instances of perceived boundaries were not predicted by any of the GPRs, but would have been covered by a revised version of attack-point (GPR 2b) that predicts a boundary after an event that is relatively longer than its two predecessors, thereby partially subsuming the function of length change (see also Deliège 1987, experiment 1).

Given the differences between these two experimental paradigms, it is not certain that they probe the same cognitive systems. Peretz (1989) addressed this question by comparing the two methods on one set of stimuli (French folk melodies). The judgment paradigm (online, explicit) showed that musicians and non-musicians responded significantly more often in accordance with GPR 3d than they did with GPR 3a. However, the recognition-memory paradigm (offline, implicit) showed no effect of boundary type for either group of participants. To test the possibility that this discrepancy is due to a loss of information in the offline probe-recognition task, Peretz carried out

Table 1. The quantification by Frankland and Cohen (2004) of GTTM's grouping preference rules which identify boundaries between notes based on their properties including local proximity to other notes (GPR 2) or the extent to which they reflect local changes in pitch or duration (GPR 3).

GPR	Description	Note property	Boundary strength
2a	rest		Absolute magnitude or rest (semibreve = 1.0)
2b	attack-point	length	$1.0 - \frac{n_1 + n_3}{2 \times n_2}$, if $n_2 > n_3 \wedge n_2 > n_1$,
			$\perp$ otherwise
3a	register change	pitch height	$1.0 - \frac{\lvert n_1 - n_2 \rvert + \lvert n_3 - n_4 \rvert}{2 \times \lvert n_2 - n_3 \rvert}$, if $n_2 \neq n_3 \wedge \lvert n_2 - n_3 \rvert > \lvert n_1 - n_2 \rvert \wedge \lvert n_2 - n_3 \rvert > \lvert n_3 - n_4 \rvert$
			$\perp$ otherwise
3d	length change	length	$1.0 - \begin{cases} n_1/n_3, \text{ if } n_3 \geqslant n_1 \\ n_3/n_1, \text{ if } n_3 < n_1 \end{cases}$

a third experiment 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. As predicted, the results demonstrated that GPR 3d, but not 3a, predicted boundary perception. In contrast, however, Frankland and Cohen (2004) found support for the utility of boundary formation using a retrospective recognition-memory task, but analyses were not presented.

2.3 *Cognitive models*

Frankland and 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 support reported 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 also comply with GPR 2b. To alleviate these problems, one might create artificial stimuli each of which is consistent with just one hypothesis at the expense of ecological validity.

2.3.1 *Grouping preference rules.* To address the issues discussed above, Frankland and Cohen (2004) quantified GPRs 2a, 2b, 3a, and 3d as shown in table 1. The quantification of GPR 2a reflects the presence of rests only and not slurs. A natural result of the individual quantifications is that they can be combined with multiple regression used to quantify the implication contained in GPR 4 that co-occurrences of two or more aspects of GPRs 2 and 3 lead to stronger boundaries. On the basis of their experimental results, Frankland and Cohen suggest a revised version of GPR 2b that applies only to notes that are relatively longer than their two predecessors.

2.3.2 *The local boundary detection model.* Cambouropoulos (2001) proposed a model related to the quantified GPRs 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. The LBDM operates over several independent parametric melodic profiles $P_k = [x_1, x_2, ..., x_n]$ where k is {pitch, IOI, rest}, $x_i > 0$, $i \in \{1, 2, ..., n\}$ 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:

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

For each parameter k, the boundary strength profile $S_k = [s_1, s_2, ..., s_n]$ 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) found that the LBDM obtained a recall of 63%–74% of the boundaries marked on a score by a musician (depending on the threshold and weights used) although precision was lower at 55%. In further experiments, it was demonstrated that notes falling before predicted boundaries were more often lengthened than shortened in pianists' performances of Mozart piano sonatas and a Chopin étude.

More recently, Cambouropoulos (2006) proposed a complementary model, not tested here, which identifies instances of melodic repetition (or parallelism, see GPR 6) and computes a pattern segmentation profile.

2.3.3 *Grouper.* Temperley (2001) introduces a model called 'Grouper' which accepts as input a melody, in which each note is represented by its onset time, off time, chromatic pitch, and level in a metrical hierarchy, and returns a single, exhaustive partitioning of the melody into nonoverlapping 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 IOIs (the interval between the onset of one note and the onset of the next note) and (b) large offset-to-onset intervals (OOI, the interval between the end of one note and the onset of the next note); PSPR 1 is 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 with about 10 notes, achieved by penalising predicted phrases by $|(\log_2 N) - \log_2 10|$ where N is the number of notes in the predicted phrase—the preferred phrase length is chosen ad hoc (see Temperley 2001, page 74), to suit the corpus of music being studied (in this case, Temperley's sample of the EFSC) and therefore may not be general.
PSPR 3 (metrical parallelism rule): prefer to begin successive groups at the same point in the metrical hierarchy (eg on the same beat of the bar).

The first rule is another example of the Gestalt principle of temporal proximity (cf GPR 2 above) while the third is related to GPR 6; the second was determined through an empirical investigation of the typical phrase lengths in a collection of folk songs. The best analysis of a given piece is computed offline by 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. Unlike the other models, this procedure results in binary segmentation judgments rather than continuous boundary strengths. By way of evaluation, Temperley used Grouper to predict the phrase boundaries marked in 65 melodies from the Essen Folk Song Collection of several thousand folk songs with phrase boundaries annotated by expert musicologists, achieving a recall of 0.76 and a precision of 0.74.

2.3.4 *Data oriented parsing.* Bod (2001) argues for a supervised learning approach to modelling melodic grouping structure as an alternative to the rule-based approach. The three grammar-learning methods he proposed learn how to segment melodies by analysing a training set of melodies explicitly segmented by a musicologist. He compares the performance of three methods in matching the musicologists' segmentations in a test set of unseen melodies. The best-performing method, data oriented parsing (DOP—Bod 1998), achieved an F1 score of 0.81. A qualitative examination of the folk song reveals several cases of boundaries which cannot be explained by Gestalt principles of pitch proximity or temporal proximity between notes but which are captured by the DOP parser and, conversely, several cases of non-boundaries which contain large pitch changes or time intervals between notes but which are classified correctly by the DOP parser. Bod uses these observations to argue that grouping involves memory for characteristic patterns rather than simple perceptual or musical principles.

2.3.5 *Transition probabilities and pointwise mutual information.* In discussing probabilistic models, we consider a melody to be a sequence of values $e_1, e_2, ..., e_j$ of length j representing some property (eg pitches taken from an alphabet E of pitch names) of the notes in the melody. We denote a subsequence from index i to k of such a sequence as e_i^k. When $i = k$, e_i^k is a single note and we use the shorthand e_i. A *transition* (or diagram) *probability* (TP) is the conditional probability of an event e_i at index

$i \hat{\mathrm{I}} \{2, ..., j\}$ in a sequence of length j given the preceding element e_{i-1}:

$$p(e_i|e_{i-1}) = \frac{count(e_{i-1}^{i})}{count(e_{i-1})}, \tag{3}$$

where $count(e_i^k)$ indicates the number of times the subsequence e_i^k occurs in some collection of sequences used to estimate the probabilities.

In research on language acquisition, it has been shown that infants and adults reliably identify grouping boundaries in sequences of synthetic syllables on the basis of statistical cues (Saffran et al 1996). In these experiments, participants are exposed to long, isochronous sequences of syllables where the only reliable cue to boundaries between groups of syllables is that transition probabilities are higher within than between groups. Further research using the same experimental paradigm has demonstrated that infants and adults use the implicitly learnt statistical properties of pitch (Saffran et al 1999), pitch interval (Saffran and Griepentrog 2001), and scale degree (Saffran 2003) sequences to identify segment boundaries on the basis of higher digram probabilities within than between groups.

In a comparison of cognitive models for word identification in infant-directed speech, Brent (1999a) quantified these ideas in a model that puts a word boundary between phonemes whenever the transition probability at e_i is lower than at both e_{i-1} and e_{i+1}. Brent also introduced a related model that replaces digram probabilities with pointwise mutual information (PMI), $I(e_i, e_{i-1})$, which measures how much the occurrence of one event reduces the model's uncertainty about the co-occurrence of another event (Manning and Schütze 1999) and is defined as:

$$I(e_i, e_{i-1}) = \log_2 \frac{p(e_{i-1}^{i})}{p(e_i)p(e_{i-1})}. \tag{4}$$

While digram probabilities are asymmetrical with respect to the order of the two events, PMI is symmetrical in this respect.[1] Brent (1999a) found that the PMI model outperformed the transition probability model in predicting word boundaries in phonemic transcripts of infant-directed speech.

Brent (1999a) implemented these models such that a boundary was placed whenever the statistic (TP or PMI) was higher at one phonetic location than in the immediately neighbouring locations. By contrast, here we construct a boundary-strength profile P at each note position i for each statistic $S = \{\mathrm{TP}, \mathrm{PMI}\}$ such that:

$$P_i = \begin{cases} \dfrac{2S_i}{S_{i-1} + S_{i+1}}, & \text{if } S_i > S_{i-1} \wedge S_i > S_{i+1}, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

2.3.6 *Model comparisons.* Most of the models discussed in section 2.3 were evaluated to some extent by their authors and, in some cases, compared quantitatively to other models. Bod (2001), for example, compared the performance of his data-oriented parsing with other closely related methods (Markov and treebank grammars). In addition, however, a handful of studies has empirically compared the performance of different melodic segmentation models. These studies differ in the models compared, the type of ground truth data used and the evaluation metrics. Melucci and Orio (2002), for example, collected the boundary indications of 17 music scholars on melodic excerpts from

[1] Manning and Schütze (1999) note that pointwise mutual information is biased in favour of low-frequency events inasmuch as, all other things being equal, I will be higher for digrams composed of low-frequency events than for those composed of high-frequency events. In statistical language modelling, pointwise mutual information is sometimes redefined as $count(xy)I(x, y)$ to compensate for this bias.

20 works by Bach, Mozart, Beethoven, and Chopin. Having combined the boundary indications into a ground truth, they evaluated the performance of the LBDM against three models that inserted boundaries after fixed (8 and 15) or random (in the range of 10 and 20) numbers of notes. Melucci and Orio report false positives, false negatives and a measure of disagreement which show that the LBDM outperforms the other models.

Bruderer (2008) evaluated a broader range of models in a study of the grouping structure of melodic excerpts from six Western pop-songs. The ground truth segmentation was obtained from 21 adults with different degrees of musical training; the boundary indications were summed within consecutive time windows to yield a quasi-continuous boundary strength profile for each melody. Bruderer examined the performance of three models: Grouper, LBDM, and the summed GPRs (GPR 2a, 2b, 3a, and 3d) quantified by Frankland and Cohen (2004). The output of each model was convolved with a Gaussian window to produce a boundary strength profile that was then correlated with the ground truth. Bruderer reports that the LBDM achieved the best and the GPRs the worst performance.

In another study, Thom et al (2002) compared the predictions of the LBDM and Grouper with segmentations at the phrase and subphrase level provided by 19 musical experts for 10 melodies in a range of styles. In one experiment, the performance of each model on each melody was estimated by averaging the F1 scores over the 19 experts. Model parameters were optimised for each individual melody. The results indicated that Grouper tended to outperform the LBDM. Large IOIs were an important factor in the success of both models. In another experiment, the predictions of each model were compared with the transcribed boundaries in several datasets from the EFSC. The model parameters were optimised over each dataset and the results again indicated that Grouper (with mean F1 between 0.6 and 0.7) outperformed the LBDM (mean F1 between 0.49 and 0.56).

To summarise, the few existing comparative studies suggest that more complex models such as Grouper and LBDM outperform the individual GPR rules even when the latter are combined in an additive manner (Bruderer 2008). Whether Grouper or LBDM exhibits a superior performance seems to depend on the stimuli and experimental task. We are not aware of any published study that has directly compared these rule-based models with learning-based models (such as DOP or TP/PMI).

2.4 *The IDyOM model*

We present a new model of melodic grouping (the information dynamics of music, or IDyOM, model) which, unlike the GPRs, the LBDM, and Grouper, acquires knowledge through experience, by using probabilistic learning rather than by using expert-coded symbolic rules. It can, therefore, explain how rule-like responses to music are acquired in melody cognition. The model differs from DOP in that it uses unsupervised, rather than supervised, learning which makes for a more veridical cognitive model because phrase boundaries in music, like word boundaries in speech, are not explicitly marked for the listener. The IDyOM model takes the same overall approach and background in experimental psychology (Saffran 2003; Saffran and Griepentrog 2001; Saffran et al 1999) as the TP/PMI models (see section 2.3.5). In contrast to these models, however, IDyOM uses a range of strategies to improve the accuracy of its conditional probability estimates. Before describing these aspects of the model, we first review related research in musicology, cognitive linguistics, and machine learning that further motivates a probabilistic approach to segmentation.

From a musicological perspective, it has been proposed that perceptual groups are associated with points of closure where the ongoing cognitive process of expectation is disrupted either because the context fails to stimulate strong expectations for any

particular continuation or because the actual continuation is unexpected (Meyer 1957; Narmour 1990, see section 2.1.2). These proposals can be given precise information-theoretic definitions (MacKay 2003; Manning and Schütze 1999) by reference to a model of sequences, e_i, composed of symbols drawn from an alphabet E (see section 2.3.5 for notational conventions). The model estimates the conditional probability of an element at index i in the sequence given the preceding elements in the sequence: $p(e_i|e_1^{i-1})$. Given such a model, the degree to which an event appearing in a given context in a melody is unexpected can be defined as the *information content* (MacKay 2003), $h(e_i|e_1^{i-1})$, of the event given the context:

$$h(e_i|e_1^{i-1}) = \log_2 \frac{1}{p(e_i|e_1^{i-1})} . \tag{6}$$

The information content can be interpreted as the contextual unexpectedness associated with an event. The uncertainty of the model's expectations in a given melodic context can be defined as the *Shannon entropy* (Shannon 1948) computed by averaging the information content over all symbols in the alphabet:

$$H(e_1^{i-1}) = \sum_{e \in E} p(e_i|e_1^{i-1}) h(e_i|e_1^{i-1}) . \tag{7}$$

We hypothesise that boundaries are perceived before events for which the unexpectedness of the outcome (h) and the uncertainty of the prediction (H) are high. These correspond to two ways in which the prior context can fail to inform a listeners' sequential predictions leading to the perception of a discontinuity in the sequence. Segmenting at these points leads to cognitive representations of the sequence (in this case a melody) that maximise likelihood and simplicity with respect to a prior model (cf Chater 1996, 1999). In the current work, we focus on the information content (h), leaving the role of entropy (H) for future work.

There is evidence that related information-theoretic quantities are important in cognitive processing of language. For example, it has recently been demonstrated that the difficulty of processing words is related both to their information content (Levy 2008) and the induced changes in entropy over possible grammatical continuations (Hale 2006). Furthermore, in machine learning and computational linguistics, algorithms based on the idea of segmenting before unexpected events can identify word boundaries in infant-directed speed with some success (Brent 1999a). Similar strategies for identifying word boundaries have been implemented by using recurrent neural networks (Elman 1990). Recently, Cohen et al (2007) proposed a general method for segmenting sequences based on two principles: first, so as to maximise the probability of events to the left and right of the boundary; and second, so as to maximise the entropy of the conditional distribution across the boundary. This algorithm was able to successfully identify word boundaries in text from four languages as well as episode boundaries in the activities of a mobile robot. These results in language and music perception suggest that the relationship between expectation and grouping may generalise across auditory domains.

The IDyOM model is presented in detail elsewhere (Pearce et al 2005; Pearce and Wiggins 2004). It may be considered an extended version of the TP model; both models estimate the conditional probability of a note given the preceding notes. However, IDyOM uses a number of methods to generate more accurate probability estimates. First, whereas TP estimates the conditional probability of a note given only the preceding note (a context of one note), IDyOM uses longer contexts (the actual length of the context used in a given situation varies). As with human listeners, the model's expectations are uncertain at the start of a melody but become more accurate with longer contexts. Second, while TP simply uses a static corpus of music to estimate its

probabilities, IDyOM combines these static long-term predictions (reflecting the prior musical experience of a listener) with short-term predictions learned dynamically while listening to the current piece of music (reflecting the perception of local structure particular to that composition). Third, while TP only applies to pitch, IDyOM also has the ability to model many different features of music and combine them into a single prediction.

The probabilities generated by IDyOM have been shown to predict listeners' expectations for melodic pitch better than existing rule-based models based on the IR theory (Pearce et al 2010b; Pearce and Wiggins 2006). Here we take the same model with two differences. First, rather than predicting just pitch, the model predicts three basic features of notes: their pitch, IOI and OOI, and multiplies the probabilities to reach an overall probability for a note. Second, while Pearce and Wiggins (2006) used derived features (such as pitch interval, pitch contour, tonal scale degree, etc.) to predict pitch, here we use only basic features to predict basic features. Based on our hypothesised relationship between expectation violation and boundary perception, we treat the sequence of information contents (or negative log probability, described above). We emphasise that the model never has access to boundary information in the melodies it is exposed to; it simply generates expectations and is in no way optimised to predict grouping boundaries.

2.5 *Peak picking*

With the exception of Grouper, all models used in our analysis (including IDyOM) produce continuous boundary strength values for each note in a melody, which reflects the likelihood that the note is followed by a boundary. However, boundaries are usually experienced as categorical phenomena between two groups of musical events (Lerdahl and Jackendoff 1983). In keeping with this observation and also with previous empirical studies of melodic boundary perception, the participants in our study (described below) were asked to indicate categorically the locations of boundaries on the grounds that this reflects the experience of boundary perception in music. Therefore, we need to identify categorical boundaries in the boundary strength profiles returned by the models, so as to be able to compare their output with the boundaries indicated by the participants.

We do this using three principles. First, the note following a boundary should have a boundary strength greater than, or equal to, the note following it: $S_n \geqslant S_{n+1}$. Second, the note following a boundary should have a greater boundary strength than the note preceding it: $S_n > S_{n-1}$. Third, the note following a boundary should have a high boundary strength relative to the local context. We implement this principle by requiring the boundary strength S_n to exceed by k standard deviations the mean boundary strength computed in a linearly weighted window measured in notes from the beginning of the piece to the preceding event:

$$S_n > k\left[\frac{\sum_{i=1}^{n-1}\left(w_i S_i - \bar{S}_{w,\,1\ldots n-1}\right)^2}{\sum_{i=1}^{n-1} w_i} + \frac{\sum_{i=1}^{n-1} w_i S_i}{\sum_{i=1}^{n-1} w_i}\right]^{1/2}. \tag{8}$$

The threshold therefore depends on the variance of the boundary strength profile up to the current event such that the strength of recent events have a greater weight than that of less recent ones. Thus the variance estimates should be more accurate towards the end of a melody, although, since more distant contributions have low weights, we do not expect any significant effect of melody length (which would in any case be consistent across the models). To create a level playing field, we apply this procedure to the boundary profiles

of all models (except Grouper) and choose a value of k to optimise the performance of each model individually. This is the only way in which the output of any of the models is optimised to fit the behavioural data; we do this so as not to unfairly impair the performance of any model when peak picking from the boundary profiles.

3 Method

3.1 *Participants*

The participants were twenty-five adults with a mean age of 28.4 years (SD = 8 years). As the experimental task required musical training, they were recruited from the Institute of Musicology at Hamburg University including graduate and postgraduate students as well as senior academics.[2] They had played a musical instrument for an average of 16.4 years (SD = 9 years), had played an average of 36.6 paid performances (SD = 60.6 years) and received a mean number of 100.8 months (about 8 years) of instrumental lessons (SD = 72.3 months). All participants were able to read and write musical notation very fluently. Participation was on a voluntary basis.

3.2 *Stimulus materials*

The stimulus set comprised fifteen monophonic vocal melodies taken from popular songs (in folk or pop styles) ranging in length from 39 to 131 note events with a mean length of 83 notes. More detailed information about the stimuli can be found in Appendix A. These melodies were selected on the basis of an informal pilot study in which three expert judges were asked to segment 36 popular melodies into melodic phrases.[3] Melodies with low agreement between judges were chosen as stimuli for the present study on the grounds that they would provide challenging test cases for the segmentation models.

The melodies were obtained as MIDI files without lyrics, expressive timing or dynamics. They were synthesised to CD quality audio files at their natural tempi with a Roland Sound Canvas software synthesiser (Grand Piano timbre, MIDI patch 1). An audio CD was created containing the entire stimulus set, including repeated presentations of each melody and intervening silence.

3.3 *Procedure*

The participants were briefed that the purpose of the experiment was to indicate phrase boundaries for monophonic melodies. They were not provided with a definition of a melodic phrase or a phrase boundary and were assured that there were no right or wrong answers in this task. To put this task in a musically meaningful context, it was suggested that they imagine themselves as a choir director marking phrase boundaries in the score for an amateur choir, not formally trained in music: the boundaries should be inserted in such a way that singers could learn, remember, and reproduce the melodies quickly and accurately. The participants were informed that the marked boundaries should fit best with their interpretation of the musical structure.

The participants were provided with paper scores of the melodies and were asked to indicate strong and weak phrase boundaries by inserting symbols between notes on the score. The melodies were presented aurally over high quality loudspeakers at approximately 70 dB. Each melody was presented twice with a silent interval of 2 s between each presentation and 5 s between each melody. Participants were encouraged to insert boundary marks during or between listenings and were explicitly allowed to correct erroneously placed marks.

[2] Musically trained participants were used for two reasons. First, it was expected that they would be better able to accurately introspect about their perception of music than nonmusicians. Second, they would be able to accurately mark their perceptions on a score.

[3] One of these three judges was a co-author of the paper (DM) but none took part in the main experiment.

Participants were tested in three groups of 9, 11, and 5 in a quiet and acoustically damped room. The experimenter was present throughout the experiment to read the instructions, distribute and collect response sheets, ensure the CD ran correctly, and to answer questions and debrief the participants. The participants were seated individually at desks some distance apart and no interaction of any kind between participants was observed.

Following the experimental session, participants reported their age, musical training, and experience. Each test session lasted $\sim\frac{1}{2}$ h in total after which participants were provided an opportunity to ask questions and give feedback. Overall, they indicated that they were able to maintain good concentration levels. When asked if they were familiar with any of the melodies, some participants indicated that they recognised melody 14.

3.4 *Model implementations*

While Grouper marks each note with a binary indicator (1 = boundary, 0 = no boundary), the other models output a positive real number for each note which can be interpreted as a boundary strength. To create a level playing field for these models, we applied the same peak-picking method (see section 2.5) to produce binary boundary indications. An optimal value of k was chosen from the set $\{0.25, 0.5, 1.0, 1.5, 2.0\}$ separately for each model. This is the only way in which the output of any model was influenced by the behavioural data: the boundary strength profiles themselves were never fit to the participants' responses. For all models, the last note of a melody was taken to be an implicit phrase boundary. The DOP method was not included in the comparison, as an implementation that could straightforwardly be applied to MIDI files was not available. The following model implementations were tested.

GPR 2a, GPR 2b, GPR 3a, GPR 3d: implemented as described by Frankland and Cohen (2004); Frankland and Cohen's revised version of GPR 2b is also included (using the obvious implementation). In peak picking, $k = 0.25$ was used for all except GPR 2a where $k = 0.5$ and GPR 2b where $k = 1.0$.

GPRs: the sum of GPR 2a, GPR 2b revised, GPR 3a, GPR 3d. In peak picking, $k = 1.5$ was used.

LBDM: implemented and weighted as described in Cambouropoulos (2001). In peak picking, $k = 0.5$ was used.

Grouper: using the implementation by Temperley and Daniel Sleator, parameterised as described in Temperley (2001).[4]

TP/PMI: in peak picking, $k = 0.25$ for TP and $k = 2$ for PMI. The digram models were trained on the collection of folk songs and hymn melodies shown in table 2.

IDyOM: the model was trained on the melodies shown in table 2. In peak picking, $k = 1$.

Always, Never: the former always predicts a boundary at every event, the latter never predicts a boundary; provided for baseline comparison.

Table 2. The collections of melodies used for training the probabilistic models. Pearce and Wiggins (2006) showed that IDyOM models trained with this corpus can accurately predict human pitch expectations.

Description	Number of compositions	Number of events	Mean (events/composition)
Canadian folk songs/ballads	152	8553	56.27
Lutheran chorale melodies	185	9227	49.88
German folk songs	566	33087	58.46
Total	903	50867	56.33

[4] Source code available at http://www.link.cs.cmu.edu/music-analysis/grouper.html. The code was compiled and implemented in a software tool by Klaus Frieler at the University of Hamburg.

4 Results

4.1 *Aggregating participants' responses*

Strong and weak boundaries were aggregated for all subsequent analyses.[5]

Most previous studies of melodic segmentation have aggregated the responses of participants in some way. This often involves summing the boundary indications over all participants (Bruderer 2008; Deliège 1998; Ferrand et al 2003; Melucci and Orio 2002; de Nooijer et al 2008) based on the assumption that perceived boundaries can be identified at points where a majority of participants perceive a boundary. However, this kind of participant aggregation can only be justified when there is high inter-participant agreement. Most studies have not assessed inter-participant differences in perceived grouping before aggregating, exceptions being Frankland and Cohen (2004) and Melucci and Orio (2002) who found no significant inter-participant differences. Thom et al (2001), however, found much variability in agreement between participants for each melody (F1 ranging between 0.14 and 0.82 for phase judgments and 0.35 and 0.8 for subphrase judgments). Melodies whose phrase structure was emphasised by rests tended to produce higher inter-participant agreement.

Here we measured inter-participant agreement using Fleiss's κ (Fleiss 1971) with a threshold of 0.6 representing the lower bound for 'substantial agreement' (Landis and Koch 1977). Agreement was above the threshold for just seven of our fifteen melodies, so aggregation would have meant excluding more than half of the data (eight out of fifteen melodies, 622 out of 1250 notes) from the analysis. However, subthreshold inter-rater agreement need not imply complete disagreement between the participants, but simply that some used different segmentation strategies from others, in response to ambiguous stimuli. It is possible that different groups of participants perceived distinct, equally valid, segmentations of the melodies. Such ambiguity seems likely from the point of view of musical practice: choral and orchestral conductors frequently differ on where to place phrasing marks, and musical scores often allow for individual expressive performance in phrasing. It, no doubt, also reflects the fact that these are unfamiliar melodies with no lyrics, orchestration, harmony, or expressive timing and dynamics to guide grouping perception. Furthermore, we deliberately selected melodies with complex grouping structure to create a challenging test for the models.

If there is genuine structural ambiguity, aggregating participants' responses is likely to lead to incomplete or fictitious segmentation solutions. Figure 1 illustrates this situation. The mean number of boundaries perceived in this melody is seven. However, six boundaries generate high agreement while the seventh shows much lower agreement. One interpretation is that there are different strategies for placing a seventh boundary and that votes get split between notes 21, 36, and 68. Melodies 7 and 13 also seem to encourage different segmentation strategies; in both cases, majority vote produces only one boundary. Figures 2 and 3 show that there are regions (eg note numbers 4 – 7 in melody 7 and note numbers 28 – 31 in melody 13) where participants place boundaries such that votes are split between neighbouring locations, complicating the selection of boundary locations.

One solution would be to use a wider time window for summing votes (Spiro 2006), but, in the case of genuine ambiguity, this would only exacerbate the problem.

[5] The overall number of boundaries reported was consistent across participants (a mean ratio of 0.11 boundaries to nonboundaries over all melodies and participants; SEM = 0.007; SD = 0.033). However, preliminary analysis of the results revealed differences between participants in the way strong and weak boundaries were perceived or reported: the variance of the ratio of strong to weak boundaries was quite high (mean ratio = 0.47; SEM = 0.06; SD = 0.32). Therefore, since both weak and strong boundaries indicate the end of a group, and since there was good agreement on boundary locations (see section 4.1), the two categories were aggregated into one, and the analysis carried out on the basis of boundary location, regardless of boundary strength.

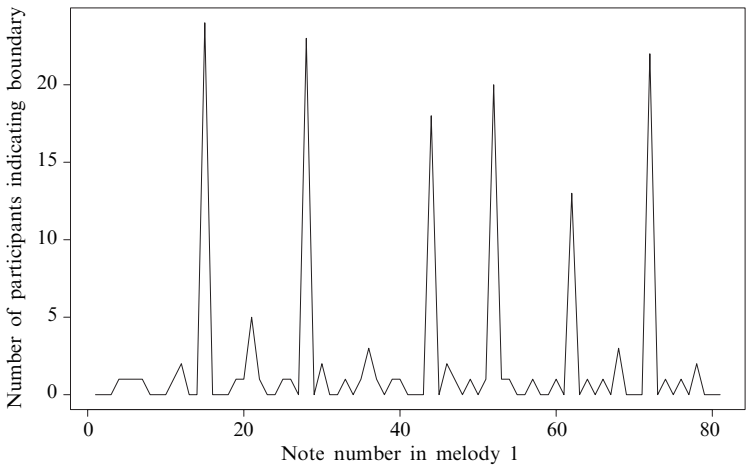

Figure 1. The sum of participants' boundary indications for each note in melody 1.

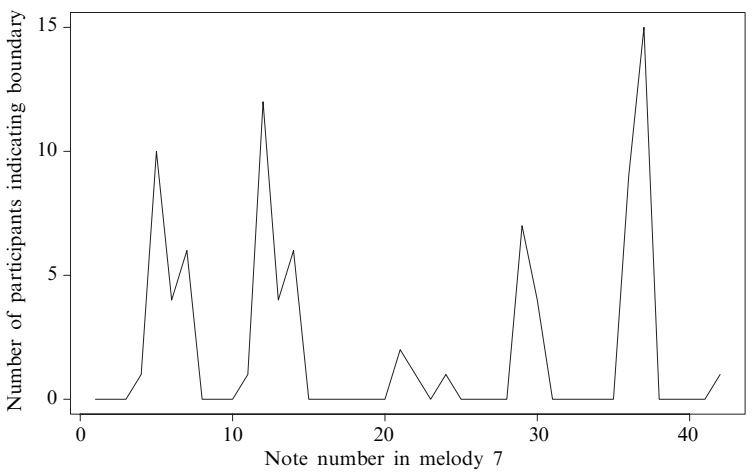

Figure 2. The sum of participants' boundary indications for each note in melody 7.

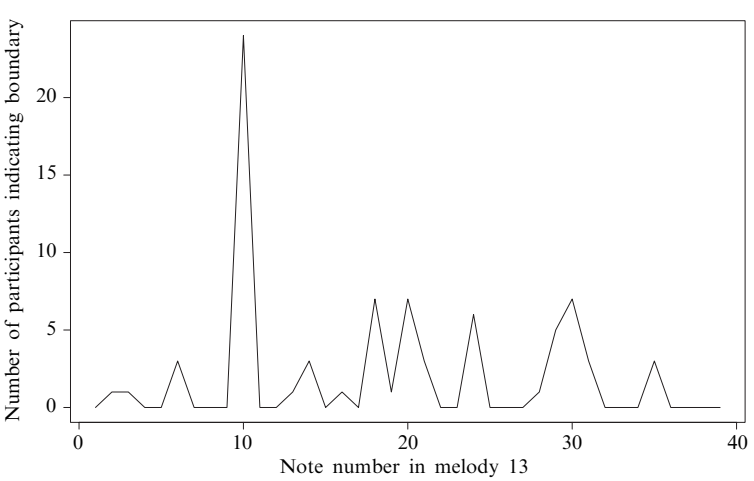

Figure 3. The sum of participants' boundary indications for each note in melody 13.

Another approach, adopted by Thom et al (2002), is to compute the performance of each model for each participant and then average over all participants. Again, however, this makes the averaged performance value dependent on the relative number of participants who used a segmentation strategy similar to that of the model. Consider an extreme example in which an equal number of participants chose either of two valid but nonoverlapping segmentation strategies with the same number of boundaries (ie all boundaries are in different places). A model that follows one of these strategies perfectly would only achieve 50% accuracy. A second model that outputs the union of the set of boundaries from the two strategies (producing twice as many boundaries than any of the participants) will also achieve 50% performance accuracy in spite of being clearly incorrect.

As a result of these considerations, we address our low inter-participant agreement by explicitly investigating the possibility that each melody admits multiple segmentation solutions, each of which is musically and perceptually valid.

4.2 *Clustering participants with similar strategies*

Melucci and Orio (2002) noted a certain amount of disagreement between the segmentation markings of their participants. However, as they did not observe clear distinctions between participants when their responses were scaled by MDS and subjected to a cluster analysis, they aggregated all participants' boundary markings to binary judgments using a probabilistic procedure. We follow a similar approach, using a clustering algorithm to identify potential groups (clusters) of participants that exhibit similar segmentation strategies for each melody.

First, we use hierarchical agglomerative clustering (Everitt and Dunn 2001) with a complete linkage distance metric using the Kulczynski distance (Kulczynski 1927) between pairs of segmentations of particular participants on particular melodies. Figure 4 shows an example of the resulting cluster dendrogram of participants for melody 8.

Second, for each melody, we cluster participants by slicing the dendrogram horizontally at a height determined by several constraints which are intended to achieve a balance between a small number of clusters for each melody (up to five), the inclusion of a large number of participants (at least three per cluster) and high agreement within clusters (Kulczynski distance < 0.5 and $\kappa > 0.6$).

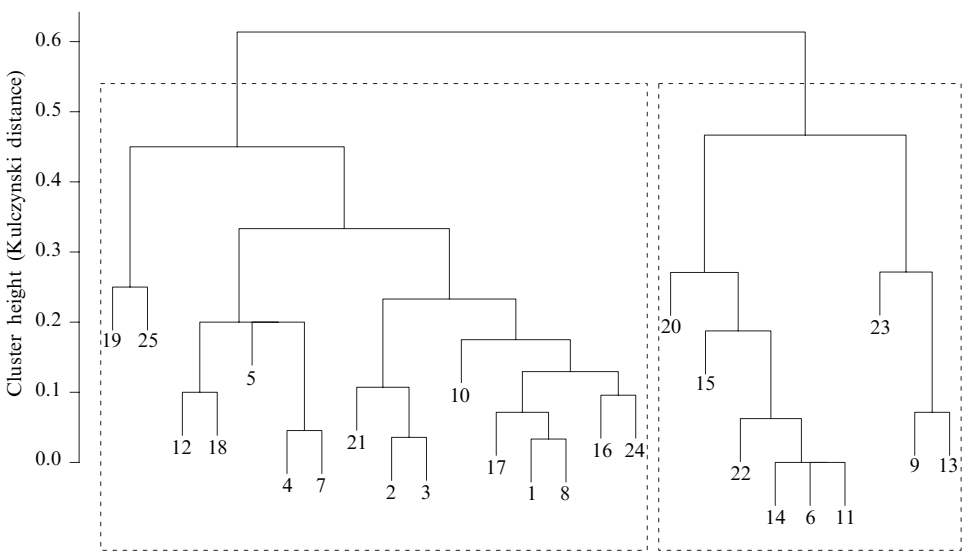

Figure 4. A cluster dendrogram of participants' responses for melody 8. The dashed boxes show the two clusters produced by slicing the dendrogram horizontally at a given height.

Across all fifteen melodies, the cluster analysis produced 78 clusters of which we excluded 31 on the grounds that they contained only one or two participants and a further 6 on the grounds of low inter-participant agreement ($\kappa < 0.6$), leaving a total of 41 clusters of participants for the fifteen melodies. Table 3 gives an overview of the clustering solutions for each melody, including the total number of clusters, the number of participants and inter-participant agreement within each cluster and the total number of participants excluded either for being part of a small cluster or a cluster with low agreement.

Table 3. A summary of participant clustering for each melody. The second column shows the total number of participants who responded to each melody (not all participants completed the task for all melodies) while the third column gives the total number of clusters generated by the cluster analysis. The following five columns give the values of κ and numbers of participants (in brackets) for the clusters with three or more participants. Values of κ above the threshold of 0.6 are in bold to distinguish them from clusters with low inter-participant agreement. The final column indicates the number of participants excluded for each melody, either through being in a cluster of one or two participants or through being in a cluster with low inter-participant agreement.

Melody number	Total number of participants	Total number of clusters	Clusters with three or more participants					Participants excluded
			κ_1	κ_2	κ_3	κ_4	κ_5	
1	24	5	**0.81**(19)					5
2	24	3	**0.87**(13)	**0.80** (7)	**0.61** (4)			0
3	23	6	**0.79** (9)	**0.81** (4)	**0.90** (5)	**0.76**(3)		2
4	23	8	**0.74** (6)	**0.64** (5)	0.45 (3)	**0.60**(4)		9
5	24	6	**0.81**(11)	**0.68** (3)	**0.81** (7)			3
6	23	6	**0.87**(11)	**0.66** (4)	**0.81** (3)			5
7	22	6	**0.76** (4)	**0.76** (6)	**0.80** (5)	**0.64**(3)		4
8	25	2	**0.77**(16)	**0.78** (9)				0
9	25	3	**0.75**(21)	0.55 (3)				4
10	23	7	**0.73 (5)**	**0.64(10)**	**0.55 (3)**			**8**
11	25	6	**0.72** (7)	**0.74 (4)**	**0.71** (7)	**0.70**(3)	**0.70**(3)	1
12	25	3	**0.83**(22)					3
13	25	8	0.54 (3)	**0.88** (6)	**0.74** (3)	**0.79**(7)		9
14	25	3	**0.89** (7)	**0.73** (5)	**0.84**(13)			0
15	25	6	0.57 (5)	**0.63** (7)	**0.68** (5)	**0.60**(4)	0.59(3)	9

Finally, for each cluster of participants on each melody, we generated a single segmentation solution to represent that cluster. This was achieved by summing the boundary indications of all participants within a cluster for each note in a melody and using the k-means algorithm (Everitt and Dunn 2001) with $k = 2$ to classify each note in a melody as a boundary or nonboundary.

4.3 *Comparative evaluation of model performance*

To evaluate the cognitive models, we compare the segmentations of each model with the segmentations of one cluster of participants for each melody. We follow a *best-only* strategy where, for each melody, the cluster that matches the model's segmentation most closely (according to F1) is selected (on the grounds that each model can only represent a single cognitive segmentation strategy at a given time). Since each model's parameters, as chosen by its author, are held fixed, the only degree of freedom in this comparison is the parameter k in the peak picker, which is optimised on the behavioural data individually for each model.

The results of model comparison using the best-only approach are shown in table 4 in terms of the mean and standard deviation of the F1 values over the fifteen melodies (mean precision and recall values are also shown for comparison). Although the combined

Table 4. The mean performance of each model over all fifteen melodies.

Model	Mean precision	Mean recall	Mean F1	SD F1
Grouper	0.86	0.82	0.83	0.16
LBDM	0.79	0.81	0.78	0.11
IDyOM	0.57	0.73	0.64	0.31
GPR 2a	0.70	0.54	0.58	0.41
GPRs	0.38	0.68	0.47	0.16
GPR 2b revised	0.47	0.45	0.43	0.26
GPR 2b	0.46	0.42	0.40	0.27
PMI	0.24	0.49	0.31	0.14
TP	0.25	0.45	0.31	0.18
GPR 3a	0.26	0.43	0.30	0.15
Always	0.13	1.0	0.23	0.08
GPR 3d	0.17	0.11	0.11	0.16
Never	0.00	0.00	0.00	0.00

GPRs model correctly predicts more boundaries (higher recall) than any of its component rules, this is outweighed in the overall F1 score by poor precision caused by the higher number of false positives. Four models (Grouper, LBDM, IDyOM, and GPR 2a) achieved F1 values greater than 0.5 and were selected on this basis for further analysis.

Owing to significant inhomogeneity of variance of F1 values between the four models (Bartlett's $K^2 = 24.97$, df $= 3$, $p < 0.01$), sign-tests were used as a non-parametric alternative to the t test. The high standard deviation in F1 (and relatively low recall) for GPR 2a probably reflect the fact that this model indicates no boundaries for melodies that do not contain rests. At an α level of 0.05, two-tailed sign tests indicated no significant differences in performance between these four models, and low effect sizes compared to a chance model ($p = 0.5$), except between Grouper and GPR 2a (Grouper versus GPR 2a: Grouper's F1 value was higher than that of GPR 2a on twelve out of fifteen melodies, so the effect size of the sign test $g = 12/15 - 0.5 = 0.3$, $p = 0.04$; Grouper versus LBDM: $g = 0.17$, $p = 0.3$; Grouper versus IDyOM: $g = 0.29$, $p = 0.06$; LBDM versus IDyOM, $g = 0.21$, $p = 0.18$; LBDM versus GPR 2a: $g = 0.1$, $p = 0.61$; GPR 2a versus IDyOM: $g = 0.12$, $p = 0.58$).

In the previous analysis, individual models could have been tested on different clusters of participants for a given melody, corresponding with the assumption that these models correspond with different segmentation strategies. We complement this approach with an alternative analysis using only melodies 1 and 12, each of which has a single cluster, meaning that all models are evaluated against the same segmentation of each melody. Note that this is a more stringent criterion of inter-participant agreement than a threshold of Fleiss's $\kappa > 0.6$ used in section 4.1. The values of k used in peak picking were the same as those used in the previous analysis (see section 3.4). The results are shown in table 5 and are broadly comparable with those shown in table 4. The same four models achieve F1 values greater than 0.5 although in this case IDyOM outperformed LBDM, particularly in terms of recall. Grouper achieved perfect performance on both melodies.

4.4 *Combining models*

The cognitive models discussed in this paper differ along several general dimensions. For example, the GPRs, LBDM, and Grouper use rules derived from expert musical knowledge, while DOP and TP/PMI rely on learning from musical examples. Looking in more detail, DOP uses supervised training while TP/PMI uses unsupervised induction of statistical regularities. Along another dimension, the GPRs, LBDM, and TP/PMI predict phrase boundaries locally, while Grouper and DOP attempt to find the best segmentation of an entire melody. The models also differ in terms of the musical features they use:

Table 5. The mean performance of each model over melodies 1 and 12.

Model	Mean precision	Mean recall	Mean F1	SD F1
Grouper	1.0	1.0	1.0	0.0
IDyOM	0.72	0.94	0.82	0.06
LBDM	0.72	0.69	0.69	0.13
GPR 2a	1.0	0.48	0.57	0.49
GPR 2b revised	0.48	0.52	0.43	0.25
GTTM	0.34	0.54	0.42	0.31
GPR 2b	0.26	0.52	0.35	0.33
GPR 3a	0.26	0.38	0.31	0.10
PMI	0.18	0.46	0.26	0.15
Always	0.10	1	0.18	0.06
TP	0.11	0.25	0.15	0.21
GPR 3d	0.0	0.0	0.0	0.0
Never	0.0	0.0	0.0	0.0

LBDM and IDyOM use pitch, IOI and OOI; Grouper uses IOI, OOI and metrical structure; GPR 3a and TP/PMI use pitch; GPR 2a uses IOI; GPR 2b uses OOI; and so on. Finally, the models differ in idiosyncratic ways: IDyOM differs from TP/PMI, for example, in using longer contexts and more sophisticated methods to estimate note probabilities. These differences between the models suggest that most are incomplete in some way and that they might operate best at different timescales (ie frequent but weak low-level boundaries or strong but infrequent high-level boundaries). If this is the case, we might be able to achieve a better fit to the behavioural data by combining the best performing models into a weighted hybrid model.

Accordingly, logistic regression models were implemented using Grouper, LBDM, IDyOM, and GPR 2a as predictors. Apart from Grouper, which returns binary segmentation judgments, the raw boundary strength profiles were used (ie without peak picking). Rather than selecting the clusters on which the hybrid model performs best (as in the previous analysis), we tested the hybrid model against two plausible segmentation strategies. Two dependent variables were created, by selecting participant clusters for each melody, which consisted, respectively, of a high-level phrase segmentation and a lower-level subphrase segmentation. For each melody, we chose the cluster with most boundaries for the low-level segmentation and the cluster with fewest boundaries for the high-level segmentation. In the case of ties, we chose the cluster that represented more participants and then, randomly, if ties could not be resolved in that way.

The logistic regression models are shown in tables 6 and 7 for the phrase- and subphrase-level dependent variables, respectively. The tables indicate the fitted coefficients and their standard errors (obtained by maximum-likelihood estimation) which are used to compute a z-score for the contribution of each predictor to the model.

Table 6. Logistic regression results for the four predictors making up the hybrid model on the phrase-level segmentation strategy.

Predictor	Estimate	SE	z value	p
(Intercept)	−6.26	0.05	−120.50	<0.01
Grouper	2.56	0.05	52.51	<0.01
IDyOM	0.16	0.01	28.46	<0.01
LBDM	5.69	0.14	39.84	<0.01
GPR 2a	1.89	0.28	6.72	<0.01

Null deviance: 61395 on 1249 degrees of freedom
Residual deviance: 18939 on 1245 degrees of freedom

Table 7. Logistic regression results for the four predictors making up the hybrid model on the subphrase-level segmentation strategy.

Predictor	Estimate	SE	z value	p
(Intercept)	−5.00	0.04	−132.62	<0.01
Grouper	2.63	0.04	64.12	<0.01
IDyOM	0.18	0.01	33.93	<0.01
LBDM	5.68	0.14	40.21	<0.01
GPR 2a	19.54	0.52	37.30	<0.01
Null deviance: 87601 on 1249 degrees of freedom				
Residual deviance: 32664 on 1245 degrees of freedom				

For both phrase and subphrase levels, all component predictors make a significant unique contribution to the regression model, and backwards stepwise elimination using the Bayes Information Criterion (BIC) failed to remove any of the predictors from the overall model. Although IDyOM's coefficient is smaller than that of the other models, it also has a smaller standard error and makes a significant unique contribution to the model, which is why it was retained in the stepwise elimination procedure.

In table 8 (phrase level) and table 9 (subphrase level) the performance of the hybrid model is compared with that of Grouper. In both cases, sign tests demonstrated that the hybrid model achieved (marginally) significantly better performance than Grouper (phrase level: $g = 0.3$, $p = 0.05$; subphrase level: $g = 0.32$, $p = 0.03$), suggesting that Grouper, LBDM, and IDyOM all contribute different information.

Table 8. The mean performance of the hybrid model on the phrase-level segmentation strategy across the fifteen melodies.

Model	Mean precision	Mean recall	Mean F1
Hybrid	0.78	0.70	0.74
Grouper	0.60	0.77	0.66

Table 9. The mean performance of the hybrid model on the subphrase-level segmentation strategy across the fifteen melodies.

Model	Mean precision	Mean recall	Mean FT
Hybrid	0.88	0.66	0.73
Grouper	0.77	0.62	0.66

5 Discussion

The research presented here makes three novel contributions. First, we hypothesised a relationship between expectation and grouping in auditory perception. To test this hypothesis, we introduced a new cognitive model of melodic segmentation (IDyOM) derived from an existing cognitive model of pitch expectations based on unsupervised probabilistic learning and information-theoretic prediction (Pearce and Wiggins 2006). The segmentation model identifies segment boundaries before unexpected events in a melodic sequence.

Second, we collected perceptual segmentations of fifteen folk and pop melodies by twenty-five musically trained participants. Because we selected unfamiliar melodies with complex grouping structure and presented them without lyrics, harmonic accompaniment, or expressive timing and dynamics, some of our stimuli admitted different interpretations of their grouping structure leading to low inter-participant agreement.

Rather than discarding or aggregating our data, we used clustering methods to identify groups of participants sharing distinct segmentation strategies for each melody. Each melody gave rise to between one and five clusters and, in general, inter-participant agreement was high within clusters. In auditory research, the possibility of different participants perceiving stimuli in different ways is often accounted for (although not explicitly investigated) on using latent variable models (eg McAdams et al 1995). Here, by contrast, we examine such ambiguity explicitly. Most studies of ambiguity in visual perceptual research either look at within-participants reversals in multistable figures (eg Leopold and Logothetis 1999) or use various factors to bias participants towards a particular interpretation of an ambiguous stimulus (eg Balcetis and Dunning 2006). In language, it has been shown that individual differences in resolving syntactic ambiguity can be related to working memory constraints (Pearlmutter and Macdonald 1995). Here we have investigated stimuli which afford multiple perceptual interpretations in terms of grouping structure.

Third, we evaluated IDyOM in comparison with existing computational models by examining the similarity between the segmentations of all the models and those of the participants. We explicitly allowed for multiple interpretations by treating each model as a possible perceived segmentation. For each melody, each model was evaluated on the cluster of participants whose segmentation it most closely matched. The results indicated that Grouper, LBDM, and IDyOM performed well in matching the perceived segmentations of the participants, while GPR 2a performed less well but much better than the other local models based on single rules. These four successful models were entered into a logistic regression analysis to create a hybrid model that outperformed Grouper, the best-performing model, in predicting two sets of participant clusters designed to represent phrase and subphrase level grouping strategies. The fact that stepwise selection failed to remove any predictors, and that the hybrid model outperforms Grouper, indicate that the other three models complement Grouper in accounting for different aspects of the participants' segmentations. However, the results (F1 values between 0.74 and 0.83) still leave room for improved fit between the segmentations produced by listeners and the models.

These results are broadly comparable with those of Pearce et al (2010a) who found that the models ranked in the same order when tested against the expert segmentations of a musicologist on 1705 German folk songs. However, the performance of all models was worse in that study, perhaps owing to the fact that here models were tested on the optimal cluster of participants. Pearce et al also obtained slightly better performance with a hybrid model containing the same component models as those reported here. Furthermore, de Nooijer et al (2008) have recently corroborated our present findings in a separate comparative study.

To the best of our knowledge, this is the first published comparative evaluation of unsupervised-learning models of melodic segmentation to explicitly examine perceptual ambiguity. IDyOM outperformed the simple unsupervised probabilistic segmentation models (TP and PMI) and performed comparably with the best of the rule-based segmentation models (Grouper and LBDM). This is encouraging, given that the underlying model was developed to account for human pitch expectations (Pearce and Wiggins 2006); we used it to predict segmentation simply by placing boundaries before surprising notes (the model was in no way optimised for melodic segmentation).

We argue that a learning model such as IDyOM has several other advantages over models such as Grouper and LBDM. Unlike these models, IDyOM provides an account of how cognitive mechanisms of expectation and grouping might be acquired in development through interaction with the auditory environment (Bruce et al 2003; Wiggins 2007). Furthermore, while rule-based models such as those examined here implement a very specific musical (even style-specific) cognitive function, IDyOM suggests the possibility that the same cognitive mechanisms can be parsimoniously deployed in

different domains, such as language and music, flexibly adapting to the structure of the input during learning; preliminary evidence for this possibility is supplied by Wiggins (2010). Building on experimental studies of statistical segmentation of syllable and tone sequences by infants and adults (eg Saffran et al 1999) and computational research on word segmentation (eg Brent 1999a; Cohen et al 2007), these results provide further evidence that auditory grouping is influenced by expectations based on probabilistic learning. Finally, we suggest that probabilistic models of perceptual processes, such as expectation and segmentation, have a more natural neurobiological interpretation than static domain-specific rules in terms of current theories of predictive coding in neural processing of perceptual stimuli (Barlow 1959; Friston 2010; Smith and Lewicki 2006).

The present research suggests various avenues for further investigation. In terms of experimental method, we chose musicians as our participants on the grounds that they would be best able to introspect reliably where they perceive boundaries. In this way, we hoped to reduce noise in our data, although it remains to be seen whether these results generalise to individuals without particular musical skills (although note that we only asked our participants *where* they perceived boundaries not *how* they perceived them). We followed Deliège (1987) in asking our participants to indicate musical boundaries in writing, on a visual representation of the musical notes, although unlike Deliège, we used a score as our participants were musical. Given the possibility that the visuo-spatial representation of music could have influenced the participants' segmentations, this method should be compared to others where boundaries are indicated with a mouse click (eg Frankland and Cohen 2004) in future research.

Another general recommendation for future research would be to focus on the boundaries not indicated by rests, which seem to be the most reliable indicators of boundaries. The greater ambiguity in perceived segmentation that this will almost certainly produce could be handled by the clustering approach presented here. Future research should also examine the factors that bias a given listener towards a particular grouping of a particular melody. We have assumed here that a segmentation strategy is determined by the interaction between a particular listener and a particular melody (ie there is no reason why the same listener should adopt the same strategy for all songs). It remains to be seen how stable segmentation strategies are across time and different classes of stimuli. Are there demographic factors, skills, experiences, or personality traits on the one hand, and aspects of musical structure on the other, that determine the selection of a particular segmentation strategy? Our approach could also usefully be developed and applied to other areas where genuine ambiguity in perception of the stimuli presents a challenge to traditional methods of analysis.

The results provide evidence that perceptual segmentation may be based on unsupervised probabilistic learning of expectations, so further developments of the learning-based approach to segmentation are worth pursuing. One option would be to explore other kinds of learning model, in particular alternatives to local probabilistic segmentation in the form of explicit Bayesian models of boundary identification (Brent 1999a; Goldwater 2007). We prefer an unsupervised approach to segmentation to a supervised approach (eg Bod 2001) given the lack of evidence that segmentation boundaries are explicitly marked in the environment (both musical and linguistic). However, it would be worthwhile examining in more detail the effect of additional information, such as expressive timing, on learning to detect phrase boundaries in music. In human development, early exposure to this information might take place via social, expressive, or linguistic cues in infant-directed singing and be used to learn melodic formulae that could subsequently be used to segment music without the explicit cues. This mechanism could be tested with the IDyOM model, by simulating an infant who learns to associate pitch patterns with natural boundaries (eg the end of a melody or a long

pause). In a training phase, the model would learn to predict natural boundaries on the basis of the preceding pitch patterns and would subsequently be tested on its ability to predict perceptual boundaries not associated with natural temporal markers. It should be noted, however, that pre-linguistic infants can perceive phrase boundaries in music (Jusczyk and Krumhansl 1993) and can do so on the basis of pitch statistics alone (Saffran et al 1999).

More generally, it will be interesting to investigate developmental trends in boundary perception and whether these can be modelled by IDyOM's changes of behaviour with increasing exposure to music. It is quite possible that the small training set used here does not approximate the experience of a trained adult listener and that further improvements in performance could be obtained by increasing the size of the training set (or tailoring it to the experience of particular individuals).

It will also be useful to examine ways of tailoring IDyOM specifically for segmentation, including a metrically-based rather than an event-based representation of time, optimising the derived features that it uses to make event predictions, and using other information-theoretic measures such as entropy or predictive information (Abdallah and Plumbley 2009). One of the attractive features of the model is that such measures (and the learning models on which they rely) are in no sense domain-specific and so can be applied generally. A promising avenue for future research will involve applying the approach to other cognitive domains (eg language, memory, vision) to examine its validity as a general cognitive model of segmentation. There is evidence, for example, that probabilistically surprising events attract visual attention (Itti and Baldi 2006) suggesting the possibility that these events may provide markers for the cognitive segmentation of visual scenes.

Acknowledgments. We would like to thank David Temperley and Daniel Sleator for making their implementation of Grouper publicly available, Klaus Frieler for support applying Grouper to our stimuli, Christian Hennig for statistical advice, and the participants for taking part in the study. This research was supported by EPSRC via grant numbers GR/S82220/01 and EP/D038855/1.

References

Abdallah S A, Plumbley M D, 2009 "Information dynamics: Patterns of expectation and surprise in the perception of music" *Convection Science* **21** 89 – 117

Attneave F, 1959 *Applications of Information Theory to Psychology* (New York: Holt)

Balcetis E, Dunning D, 2006 "See what you want to see: Motivational influences on visual perception" *Journal of Personality and Social Psychology* **91** 612 – 625

Barlow H B, 1959 "Sensory mechanisms, the reduction of redundancy, and intelligence", in *Proceedings of a Symposium on the Mechanisation of Thought Processes* (London: Natural Physical Laboratory, Teddington — London: Her Majesty's Stationery Office) pp 537 – 559

Bod R, 1998 *Beyond Grammar: An Experience-based Theory of Language* (Stanford, CA: CSLI Publications)

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

Bower G, 1970 "Organizational factors in memory" *Cognitive Psychology* **1** 18 – 46

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

Brent M R, 1999a "An efficient, probabilistically sound algorithm for segmentation and word discovery" *Machine Learning* **34** 71 – 105

Brent M R, 1999b "Speech segmentation and word discovery: A computational perspective" *Trends in Cognitive Sciences* **3** 294 – 301

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

Bruce V, Green P, Georgeson M, 2003 *Visual Perception: Physiology, Psychology and Ecology* (London: Psychology Press)

Bruderer M J, 2008 *Perception and Modeling of Segment Boundaries in Popular Music* PhD thesis, J F Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven, The Netherlands

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* (San Francisco, CA: ICMA) pp 17 – 22
Cambouropoulos E, 2006 "Musical parallelism and melodic segmentation: A computational approach" *Music Perception* **23** 249 – 269
Chater N, 1996 "Reconciling simplicity and likelihood principles in perceptual organisation" *Psychological Review* **103** 566 – 581
Chater N, 1999 "The search for simplicity: A fundamental cognitive principle?" *Quarterly Journal of Experimental Psychology A* **52** 273 – 302
Clarke E F, Krumhansl K L, 1990 "Perceiving musical time" *Music Perception* **7** 213 – 252
Cohen P R, Adams N, Heeringa B, 2007 "Voting experts: An unsupervised algorithm for segmenting sequences" *Intelligent Data Analysis* **11** 607 – 625
Deliège I, 1987 "Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's grouping preference rules" *Music Perception* **4** 325 – 360
Deliège I, 1998 "Wagner 'alte weise': Une approche perceptive" *Musicæ Scientiæ Special Issue* 63 – 90
Dowling W J, 1973 "Rhythmic groups and subjective chunks in memory for melodies" *Perception & Psychophysics* **14** 37 – 40
Elman J L, 1990 "Finding structure in time" *Cognitive Science* **14** 179 – 211
Everitt B S, Dunn G, 2001 *Applied Multivariate Data Analysis* (London: Hodder Arnold)
Ferrand M, Nelson P, Wiggins G, 2003 "Unsupervised learning of melodic segmentation: A memory-based approach", in *Proceedings of the 5th Triennial ESCOM Conference* Eds R Kopiez, A C Lehmann, I Wolther, C Wolf (Hanover: Hanover University of Music and Drama) pp 141 – 144
Fleiss J L, 1971 "Measuring nominal scale agreement among many raters" *Psychological Bulletin* **76** 378 – 382
Fodor J A, Bever T G, 1965 "The psychological reality of linguistic segments" *Journal of Verbal Learning and Verbal Behavior* **4** 414 – 420
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** 499 – 543
Friston K, 2010 "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* **11** 127 – 138
Goldwater S, 2007 *Nonparametric Bayesian Models of Lexical Acquisition* PhD thesis, Department of Cognitive and Linguistic Sciences, Brown University, Providence, USA
Gregory A H, 1978 "Perception of clicks in music" *Perception & Psychophysics* **24** 171 – 174
Hale J, 2006 "Uncertainty about the rest of the sentence" *Cognitive Science* **30** 643 – 672
Itti L, Baldi P, 2006 "Bayesian surprise attracts human attention", in *Advances in Neural Information Processing Systems 18* Eds Y Weiss, B Schölkopf, J Platt (Cambridge, MA: MIT Press) pp 547 – 554
Jusczyk P W, 1997 *The Discovery of Spoken Language* (Cambridge, MA: MIT Press)
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** 627 – 640
Koffka K, 1935 *Principles of Gestalt Psychology* (New York: Harcourt, Brace and World)
Krumhansl C L, 1990 *Cognitive Foundations of Musical Pitch* (Oxford: Oxford University Press)
Kulczynski S, 1927 "Die Planzenassoziationen der Pienien" *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles B* 57 – 203
Kurby C A, Zacks J M, 2007 "Segmentation in the perception and memory of events" *Trends in Cognitive Sciences* **12** 72 – 79
Ladefoged P, Broadbent D E, 1960 "Perception of sequences in auditory events" *Journal of Experimental Psychology* **12** 162 – 170
Landis J R, Koch G G, 1977 "The measurement of observer agreement for categorical data" *Biometrics* **33** 159 – 174
Leopold D A, Logothetis N K, 1999 "Multistable phenomena: changing views in perception" *Trends in Cognitive Sciences* **3** 254 – 264
Lerdahl F, Jackendoff R, 1983 *A Generative Theory of Tonal Music* (Cambridge, MA: MIT Press)
Levy R, 2008 "Expectation-based syntactic comprehension" *Cognition* **16** 1126 – 1177
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** 1853 – 1867
McAdams S, Winsberg S, Donnadieu S, Soete G D, Krimphoff J, 1995 "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities and latent subject classes" *Psychological Research* **58** 177 – 192
MacKay D J C, 2003 *Information Theory, Inference, and Learning Algorithms* (Cambridge, UK: Cambridge University Press)

Manning C D, Schütze H, 1999 *Foundations of Statistical Natural Language Processing* (Cambridge, MA: MIT Press)
Marr D, 1982 *Vision* (San Francisco, CA: W H Freeman)
Melucci M, Orio N, 2002 "A comparison of manual and automatic melody segmentation", in *Proceedings of the Third International Conference on Music Information Retrieval* Ed. M Fingerhut (Paris: IRCAM) pp 7 – 14
Meyer L B, 1957 "Meaning in music and information theory" *Journal of Aesthetics and Art Criticism* **15** 412 – 424
Narmour E, 1990 *The Analysis and Cognition of Basic Melodic Structures: The Implication-realisation Model* (Chicago: University of Chicago Press)
Newtson D, 1973 "Attribution and the unit of perception of ongoing behavior" *Journal of Personality and Social Psychology* **28** 28 – 38
Nooijer J de, Wiering F, Volk A, Tabachneck-Schijf H J M, 2008 "An experimental comparison of human and automatic music segmentation", in *Proceedings of the 10th International Conference on Music Perception and Cognition* Eds K Miyazaki, M Adachi, Y Hiraga, Y Nakajima, M Tsuzaki (Adelaide, Australia: Causal Productions) pp 399 – 407
Pearce M T, Conklin D, Wiggins G A, 2005 "Methods for combining statistical models of music", in *Computer Music Modelling and Retrieval* Ed. U K Wiil (Berlin: Springer) pp 95 – 312
Pearce M T, Müllensiefen D, Wiggins G A, 2010a "Melodic grouping in music information retrieval: New methods and applications", in *Advances in Music Information Retrieval* Eds Z W Ras, A Wieczorkowska (Berlin: Springer) pp 364 – 388
Pearce M T, Ruiz M H, Kapasi S, Wiggins G A, Bhattacharya J, 2010b "Unsupervised statistical learning underpins computational, behavioural and neural manifestations of musical expectation" *NeuroImage* **50** 302 – 313
Pearce M T, Wiggins G A, 2004 "Improved methods for statistical modelling of monophonic music" *Journal of New Music Research* **33** 367 – 385
Pearce M T, Wiggins G A, 2006 "Expectation in melody: The influence of context and learning" *Music Perception* **23** 377 – 405
Pearlmutter N J, Macdonald M C, 1995 "Individual differences and probabilistic constraints in syntactic ambiguity resolution" *Journal of Memory and Language* **34** 521 – 542
Peretz I, 1989 "Clustering in music: An appraisal of task factors" *International Journal of Psychology* **24** 157 – 178
Peretz I, 1990 "Processing of local and global musical information by unilateral brain-damaged patients" *Brain* **113** 1185 – 1205
Reynolds J R, Zacks J M, Braver T S, 2007 "A computational model of event segmentation from perceptual prediction" *Cognitive Science* **31** 613 – 643
Saffran J R, 2003 "Absolute pitch in infancy and adulthood: The role of tonal structure" *Developmental Science* **6** 37 – 49
Saffran J R, Aslin R N, Newport E L, 1996 "Statistical learning by 8-month old infants" *Science* **274** 1926 – 1928
Saffran J R, Griepentrog G J, 2001 "Absolute pitch in infant auditory learning: Evidence for developmental reorganization" *Developmental Psychology* **37** 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** 27 – 52
Schaffrath H, 1995 "The Essen folksong collection", in *Database Containing 6,255 Folksong Transcriptions in the Kern Format and a 34-Page Research Guide [Computer Database]* Ed. D Huron (Menlo Park, CA: CCARH)
Schellenberg E G, 1997 "Simplifying the implication-realisation model of melodic expectancy" *Music Perception* **14** 295 – 318
Shannon C E, 1948 "A mathematical theory of communication" *Bell System Technical Journal* **27** 379 – 423 and 623 – 656
Sloboda J A, Gregory A H, 1980 "The psychological reality of musical segments" *Canadian Journal of Psychology* **34** 274 – 280
Smith E C, Lewicki M S, 2006 "Efficient auditory coding" *Nature* **439** 978 – 982
Spiro N, 2006 "A new method for assessing consistency of real-time identification of phrase-parts and its initial application", in *Proceedings of the 9th International Conference of Music Perception and Cognition* Eds M Baroni, R Addessi, R Caterina, M Costa (Bologna: SMPC and ESCOM) pp 793 – 800
Stoffer T H, 1985 "Representation of phrase structure in the perception of music" *Music Perception* **3** 191 – 220
Tan N, Aiello R, Bever T G, 1981 "Harmonic structure as a determinant of melodic organization" *Memory and Cognition* **9** 533 – 539

Temperley D, 2001 *The Cognition of Basic Musical Structures* (Cambridge, MA: MIT Press)
Thom B, Spevak C, Höthker K, 2002 "Melodic segmentation: Evaluating the performance of algorithms and musical experts", in *Proceedings of the International Computer Music Conference* (San Francisco, CA: ICMA) pp 65 – 72
Waugh N, Norman D A, 1965 "Primary memory" *Psychological Review* **72** 89 – 104
Wiggins G A, 2007 "Models of musical similarity" *Musicæ Scientiæ* (Discussion Forum 4a) 315 – 337
Wiggins G A, 2010 "A cross-domain model: grouping of phonemes into syllables by a model of melodic segmentation", in *Proceedings of the International Conference on Music Perception and Cognition* Eds S M Demorest, S J Morrison, P S Campbell (Seattle, WA: ICMPC and Causal Productions) page 75

Appendix: The melodies

Details of the fifteen melodies presented to the participants for segmentation are shown in table A1. Melodies 3, 7, and 10 were taken from the entries identified by E0356, K0059 and K0690, respectively, in the Essen Folk Song Collection (Schaffrath 1995).

Table A1. Details of the fifteen melodies used in the experiment.

Number	Title	Artist	Genre	Tempo	Time sig.	Length
1	Children of the night	Richard Marx	Pop	75	4/4	81
2	Longer	Dan Fogelberg	Pop	80	4/4	70
3	Ihr Franzosen geht nach Haus	unattributed/trad	Folk	100	2/4	51
4	Ich wünsch mir was	J Ziegner	Pop	95	4/4	130
5	Enjoy your life	Funky Be	HipHop	85	4/4	129
6	Freiheit	M Witte	Rock/Pop	120	4/4	75
7	Pripe Ninne Sause	unattributed/trad	Folk	80	2/4	42
8	Hand in hand	H Hofbauer	Ballad	80	4/4	113
9	Please stay	H Hofbauer	Ballad	90	4/4	87
10	Ruru Rinneken	unattributed/trad	Folk	85	2/4	48
11	Let me be your only one	Funky Be	HipHop	100	4/4	106
12	Will you go, Lassie	F McPeake/trad	Folk	120	4/4	62
13	Deus	Sugarcubes	Rock	120	4/4	39
14	Here, there, and everywhere	The Beatles	Rock	120	4/4	86
15	Deep in my dreams	M Röbenack	Rock	120	4/4	131

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 39 2010

www.perceptionweb.com