Melodic Grouping in Music Information Retrieval: New Methods and Applications

Marcus T. Pearce, Daniel Müllensiefen and Geraint A. Wiggins

[1] Marcus T. Pearce, Wellcome Laboratory of Neurobiology, University College London, WC1E 6BT, UK. marcus.pearce@ucl.ac.uk
[2] Daniel Müllensiefen Department of Computing, Goldsmiths, University of London, SE14 6NW, UK. d.mullensiefen@gold.ac.uk
[3] Geraint A. Wiggins Department of Computing, Goldsmiths, University of London, SE14 6NW, UK. g.wiggins@gold.ac.uk

Summary. *

Summary. We introduce the MIR task of segmenting melodies into phrases, summarise the musicological and psychological background to the task and review existing computational methods before presenting a new model, IDyOM, for melodic segmentation based on statistical learning and information-dynamic analysis. The performance of the model is compared to several existing algorithms in predicting the annotated phrase boundaries in a large corpus of folk music. The results indicate that four algorithms produce acceptable results: one of these is the IDyOM model which performs much better than naive statistical models and approaches the performance of the best-performing rule-based models. Further slight performance improvement can be obtained by combining the output of the four algorithms in a hybrid model, although the performance of this model is moderate at best, leaving a great deal of room for improvement on this task.

1 Introduction

The segmentation of music into meaningful units is a fundamental (pre-)processing step for many MIR applications including melodic feature computation, melody indexing, and retrieval of melodic excerpts. Here, we focus on 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 & Jackendoff, 1983) or *segmentation* (Cambouropoulos, 2006) and is 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 groups we shall consider might correspond with motifs, phrases, sections and other aspects of musical form, so the scope is rather general. 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 & 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.

We restrict ourselves primarily to research on symbolic representations of musical structure that take discrete events (individual musical notes in this work) as their musical surface (Jackendoff, 1987). Working at this level of abstraction, the task is to gather events (represented in metrical time as they might be in a musical score) into sequential groups. Research on segmentation from sub-symbolic or acoustic representations of music is not discussed as it generally operates either at the level of larger sections of music differing in instrumentation (e.g., Abdallah *et al.*, 2006) or at the lower level of separating a continuous audio stream into individual note events (e.g., Gjerdingen, 1999; Todd, 1994). 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).

Melodic segmentation is a key task in the storage and retrieval of musical information. The melodic phrase is often considered one of the most important basic units of musical content (Lerdahl & Jackendoff, 1983) and many large electronic corpora of music are structured or organised by phrases, for example, the Dictionary of Musical Themes by Barlow & Morgenstern (1949), the Essen Folksong Collection (EFSC, Schaffrath, 1995) or the RISM collection (RISM-ZENTRALREDAKTION, RISM-ZENTRALREDAKTION). At the same time, melodic grouping is thought to be an important part of the perceptual processing of music (Deliège, 1987; Frankland & Cohen, 2004; Peretz, 1989). It is also fundamental to the phrasing of a melody when sung or played: melodic segmentation is a task that musicians and musical listeners perform regularly in their everyday musical practice.

Several algorithms have been proposed for the automated segmentation of melodies. These algorithms differ in their modelling approach (supervised learning, unsupervised learning, music-theoretic rules), and in the type of information they use (global or local). In this chapter, we review these approaches before introducing a new statistical model of melodic segmentation and comparing its performance to several existing algorithms on a melody segmentation task. The motivation for this model comparison is two-fold: first, we are interested in the performance differences between different types of model; and second, we aim to build a hybrid model that achieves superior performance by combining boundary predictions from different models.

2 Background

The segmentation of melodies is a cognitive process performed by the minds and brains of listeners based on their musical and auditory dispositions and experience. Therefore, an MIR

system must segment melodies in a musically and psychologically informed way if it is to be successful. Before reviewing computational models of melodic segmentation and their use in MIR, we consider it appropriate to survey the musicological and psychological literature that has informed the development of these models.

2.1 Music-theoretic Approaches

A Generative Theory of Tonal Music

Melodic grouping has traditionally been modelled through the identification of local discontinuities or changes between events in terms of temporal proximity, pitch, duration and dynamics (Cambouropoulos, 2001; Lerdahl & Jackendoff, 1983; Temperley, 2001). Perhaps the best known examples are the Grouping Preference Rules (GPRs) of the Generative Theory of Tonal Music (GTTM, Lerdahl & Jackendoff, 1983). The most widely studied of these 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 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. Another rule, GPR 6, predicts that grouping boundaries are perceived in accordance with musical parallelism (e.g., at parallel points in a metrical hierarchy or after a repeated motif). The 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 (e.g., Koffka, 1935).

The Implication-Realisation Theory

Narmour (1990, 1992) 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. The two systems may conflict and, in any given situation, one may over-ride the implications generated by the other.

In the bottom-up system, sequences of melodic intervals vary in the degree of *closure* that they convey. An interval which is unclosed (i.e., one that generates expectations for a subsequent interval) 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 (i.e., a boundary) and the melodic groups formed either side of the boundary thus created can share different amounts of structure depending on the degree of closure conveyed. Furthermore, structural notes 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.

2.2 Psychological Studies

Early studies of musical segmentation (Gregory, 1978; Sloboda & 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 & Bever, 1965; Ladefoged & Broadbent, 1960). More recently, two kinds of experimental task 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 & 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 due 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. Using the same paradigm, Tan *et al.* (1981) demonstrated the influence of harmonic closure (e.g., a cadence to the tonic chord) with an effect of musical training such that musicians were more sensitive to this parameter than non-musicians.

In the second paradigm, subjects provide explicit judgements of boundary locations while listening to the musical stimulus. The indicated boundaries are subsequently analysed to discover what 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 & Jackendoff, 1983) especially those relating to rests or long notes and changes in timbre or dynamics. These factors have also been found to be important in large-scale segmentation by musically-trained listeners of piano works composed by Stockhausen and Mozart (Clarke & Krumhansl, 1990). Frankland & Cohen (2004) collected explicit boundary judgements 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 (Attack-point) produced consistently strong correlations with the empirical boundary profiles, while GPR 2a (Rest) also received support in the one case where it applied. No empirical support was found for GPRs 3a (Register Change) and 3d (Length change).

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 both methods on one set of stimuli (French folk melodies). The judgement paradigm (online, explicit) showed that musicians and non-musicians responded significantly more often in accordance with GPR 3d (Length change) than they did with GPR 3a (Register Change). 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 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 an influence of GPR 3d, but not 3a, on boundary perception. In contrast to these results, however, Frankland & Cohen (2004) found no major difference between the results of their explicit judgement task and a retrospective recognition-memory task using the same materials.

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

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

Many questions remain open and further empirical study is necessary to fully understand perceptual grouping. Nonetheless, psychological research has guided the development of computational models of melodic segmentation, which can be applied to practical tasks in MIR.

2.3 Computational Models

Tenney & Polansky (1980) were perhaps the first to propose formal models of melodic segmentation based on Gestalt-like rules, which became the dominant paradigm in the years to come. In this section,q we review three models developed within this tradition: quantified versions of the GPRs from GTTM (Frankland & Cohen, 2004); the Local Boundary Detection Model (Cambouropoulos, 2001); and Grouper (Temperley, 2001). We also summarise previous studies that have evaluated the comparative performance of some of these models of melodic segmentation. Recently, there has been increasing interest in using machine learning to build models that learn about grouping structure, in either a supervised or unsupervised manner, through exposure to large bodies of data (Bod, 2001; Brent, 1999a; Ferrand *et al.*, 2003; Saffran *et al.*, 1999). The model we present follows this tradition and we include some related work in our review. In another direction, some researchers have combined Gestalt-like rules with higher-level principles based on parallelism and music structure (Ahlbäck, 2004; Cambouropoulos, 2006) in models which are mentioned for the sake of completeness but not reviewed in detail.

Grouping Preference Rules

Inspired by the GTTM, Frankland & Cohen (2004) quantified GPRs 2a, 2b, 3a and 3d as shown in Table 1. Since a slur is a property of the IOI while 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 change) concern perceived duration, they were based on linearly scaled time in accordance with psychoacoustic research (Allan, 1979). Finally, a natural result of the individual quantifications is that they can be combined using multiple regression (a multivariate extension to linear correlation, Howell, 2002) 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.

The Local Boundary Detection Model

Cambouropoulos (2001) proposes 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 $P_k = [x_1, x_2, \ldots, x_n]$ where $k \in \{$pitch, ioi, rest$\}$, $x_i > 0, i \in \{1, 2, \ldots, n\}$ and the boundary strength at interval x_i (a pitch interval in semitones, inter-onset interval, or offset-to-onset interval) 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 (.25 for *pitch* and *rest*, and .5 for *ioi*), and boundaries are predicted where the combined profile exceeds a threshold which may be set to any reasonable value (Cambouropoulos used a value such that 25% of notes fell on boundaries).

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. This was also true of the penultimate notes in the predicted groups.

More recently, Cambouropoulos (2006) proposed a complementary model which identifies instances of melodic repetition (or parallelism) and computes a pattern segmentation profile. While repetitions of melodic patterns are likely to contribute to the perception of grouping (see GPR 6 above), this model is not yet a fully developed model of melodic segmentation as it operates at a "local level (i.e. within a time window rather than [on] a whole piece)" (Emilios Cambouropoulos, personal email communication, 09/2007).

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 (which may be computed using a beat-tracking algorithm or computed from the time signature and bar lines if these are available), 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 IOIs and (b) large offset-to-onset intervals (OOI); 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, p. 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 parallel points in the metrical hierarchy (e.g., both on the first 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 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 judgements rather than continuous boundary strengths. By way of evaluation, Temperley used Grouper to predict the phrase boundaries marked in 65 melodies from the EFSC, a collection of several thousand folk songs with phrase boundaries annotated by expert musicologists, achieving a recall of .76 and a precision of .74.

Data Oriented Parsing

Bod (2001) argues for a supervised learning approach to modelling melodic grouping structure as an alternative to the rule-based approach. He examined three grammar induction algorithms originally developed for automated language parsing in computational linguistics: first, the treebank grammar learning technique which reads all possible context free rewrite rules from the training set and assigns each a probability proportional to its relative frequency in the training set (Manning & Schütze, 1999); second, the Markov grammar technique which assigns probabilities to context free rules by decomposing the rule and its probability by a Markov process, allowing the model to estimate the probability of rules that have not occurred in the training set (Collins, 1999); and third, a Markov grammar augmented with a Data Oriented Parsing (DOP, Bod, 1998) 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 (Rabiner, 1989) was used to generate the most probable parse for each melody in the test set given each of the three models. Bod (2001) evaluated the performance of these three algorithms in predicting the phrase boundaries in the EFSC using F1 scores (Witten & Frank, 1999). The results demonstrated that the treebank technique yielded moderately high precision but very low recall ($F1 = .07$), the Markov grammar yielded slightly lower precision but much higher recall ($F1 = .71$) while the Markov-DOP technique yielded the highest precision and recall ($F1 = .81$). A qualitative examination of the folk song data revealed several cases (15% of the phrase boundaries in the test set) where the annotated phrase boundary cannot be accounted for by Gestalt principles but is predicted by the Markov-DOP parser.

Transition Probabilities and Pointwise Mutual Information

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 consist of

higher transition probabilities within than between groups. A *transition (or digram) probability* (TP) is the conditional probability of an element e_i at index $i \in \{2,\ldots,j\}$ in a sequence e_1^j 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 e_m^n is the subsequence of e between indices m and n, e_m is the element at index m of the sequence e and $count(x)$ is the number of times that x appears in a training corpus. 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 & 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 computational methods for word identification in unsegmented 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 & 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, pointwise mutual information is symmetrical in this respect.[4] Brent (1999a) found that the pointwise mutual information model outperformed the transition probability model in predicting word boundaries in phonemic transcripts of phonemically-encoded infant-directed speech from the CHILDES collection (MacWhinney & Snow, 1985).

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 $S = \{\text{TP, PMI}\}$ such that:

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

Model Comparisons

The models reviewed above differ along several different 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.

[4] Manning & 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.

Most of these models 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 applied. Melucci & Orio (2002), for example, collected the boundary indications of 17 expert musicians and experienced music scholars on melodic excerpts from 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 a fixed (8 and 15) or random (in the range of 10 and 20) numbers of notes. Melucci & Orio report false positives, false negatives and a measure of disagreement which show that the LBDM outperforms the other models.

Melucci & Orio noticed 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 judgements using a probabilistic procedure.

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 & Cohen (2004). The output of each model was convolved with a 2.4s 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 (using a pen on a minimal score while listening to a MIDI file) by 19 musical experts for 10 melodies in a range of styles. In a first experiment, Thom *et al.* examined the average F1 scores between experts for each melody, obtaining values ranging between .14 and .82 for phrase judgements and .35 and .8 for subphrase judgements. The higher consistencies tended to be associated with melodies whose phrase structure was emphasised by rests. In a second 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 a third 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 .6 and .7) outperformed the LBDM (mean F1 between .49 and .56). 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 EFSC. The performance of both algorithms was impaired, most notably for the subphrase judgements of the experts.

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 data set and experimental task. Finally, most of these comparative studies used ground truth segmentations derived from manual annotations by human judges. However, only a limited number of melodies can be tested in this way

(ranging from 6 in the case of Bruderer, 2008 to 20 by Melucci & Orio, 2002). Apart from Thom *et al.* (2002, Experiment D), there has been no thorough comparative evaluation over a large corpus of melodies annotated with phrase boundaries. However, that study did not include the GPRs and to date, no published study has directly compared these rule-based models with learning-based models (as we do here).

2.4 A New Segmentation Model

The IDyOM Model

As we have seen, most existing models of melodic grouping consist of collections of symbolic rules that describe the musical features corresponding to perceived groups. Such models have to be adjusted by hand using detailed *a priori* knowledge of a musical style. Therefore, these models are not only domain-specific, pertaining only to music, but also potentially style specific, pertaining only to Western tonal music or even a certain genre.

We present a new model of melodic grouping (the Information Dynamics Of Music, or IDyOM, model) which, unlike the GPRs, the LBDM and Grouper, uses unsupervised learning from experience rather than expert-coded symbolic rules. The model differs from DOP in that it uses unsupervised, rather than supervised, learning which makes it more useful for identifying grouping boundaries in corpora where phrase boundaries are not explicitly marked. The IDyOM model takes the same overall approach and inspiration from experimental psychology (Saffran, 2003; Saffran & Griepentrog, 2001; Saffran *et al.*, 1999) as the TP/PMI models (see §2.3). 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 statistical 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 §2.1). These proposals may be given precise definitions in an information-theoretic framework (MacKay, 2003; Manning & Schütze, 1999) which we define by reference to a model of sequences, e_i, composed of symbols drawn from an alphabet $\mathscr{E}$. 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 or surprisal associated with an event. The contextual uncertainty of the model's expectations in a given melodic context can be defined as the *entropy* (or average information content) of the predictive context itself:

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

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 the model's 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 (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 speech with some success (Brent, 1999a). Similar strategies for identifying word boundaries have been implemented 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.

The digram models used by TP and PMI are specific examples of a larger class of models called n-gram models (Manning & Schütze, 1999). An n-gram is a sequence of n symbols consisting of a *context* of $n-1$ symbols followed by a single symbol *prediction*. A digram, for example, is a sequence of two symbols ($n=2$) with a single symbol context and a single symbol prediction. An n-gram model is simply a collection of n-grams each of which is associated with a frequency count. The quantity $n-1$ is known as the *order* of the model and represents the number of symbols making up the sequential context within which the prediction occurs. During the *training* of the statistical model, these counts are acquired through an analysis of some corpus of sequences (the training set) in the target domain. When the trained model is exposed to an unseen sequence drawn from the target domain, it uses the frequency counts associated with n-grams to estimate a probability distribution governing the identity of the next symbol in the sequence given the $n-1$ preceding symbols. Therefore, an assumption made in n-gram modelling is that the probability of the next event depends only on the previous $n-1$ events:

$$p(e_i|e_1^{i-1}) \approx p(e_i|e_{(i-n)+1}^{i-1})$$

However, n-gram models suffer from several problems, both in general and specifically when applied to music. The TP and PMI models are conceptually simple but, as models of musical structure, they have at least two major shortcomings. The first is general: probabilities are estimated purely on the basis of digram (first order) statistics collected from some existing corpus. The second problem is representational and specific to music: in estimating the probability of a note, only its pitch (and that of its predecessor) are taken into consideration - the timing of the note is ignored. In the IDyOM model, we address these shortcomings as described below.

Regarding the first problem, that of probability estimation, IDyOM uses several methods drawn from the literature on text compression (Bell *et al.*, 1990; Bunton, 1997) and statistical language modelling (Manning & Schütze, 1999) to improve the prediction performance of the model. The following is a brief description of the principal methods used; technical details can be found elsewhere (Conklin & Witten, 1995; Pearce *et al.*, 2005; Pearce & Wiggins, 2004).

Since the model is based on *n*-grams, one obvious improvement would be to increase the model order (i.e., *n*). However, while low-order models fail to provide an adequate account of the structural influence of the context, increasing the order can prevent the model from capturing much of the statistical regularity present in the training set (an extreme case occurring when the model encounters an *n*-gram that does not appear in the training set and returns an estimated probability of zero). To address this problem (and maximise the benefits of both low- and high-order models) the IDyOM model maintains frequency counts during training for *n*-grams of all possible values of *n* in any given context. This results in a large number of *n*-grams; the time and space complexity of both storage and retrieval are rendered tractable through the use of suffix trees augmented with frequency counts (Bunton, 1997; Larsson, 1996; Ukkonen, 1995). During prediction, distributions are estimated using a weighted sum of all models below an order bound that varies depending on the context (Cleary & Teahan, 1997; Pearce & Wiggins, 2004). This bound is determined in each predictive context using simple heuristics designed to minimise uncertainty (Cleary & Teahan, 1997). The combination is designed such that higher-order predictions, which are more specific to the context, receive greater weighting than lower-order predictions, which are more general (Witten & Bell, 1991).

Another problem with many *n*-gram models is that a static (pre-trained) model will fail to make use of local statistical structure in the music it is currently analysing. To address this problem, IDyOM includes two kinds of model: first, a static *long-term* model that learns from the entire training set before being exposed to the test data; and second, a *short-term* model that is constructed dynamically and incrementally throughout each individual melody to which it is exposed (Conklin & Witten, 1995; Pearce & Wiggins, 2004). The distributions returned by these models are combined using an entropy-weighted multiplicative combination scheme corresponding to a weighted geometric mean (Pearce *et al.*, 2005) in which greater weights are assigned to models whose predictions are associated with lower entropy (or uncertainty) at that point in the melody.

A final issue regards the fact that music is an inherently multi-dimensional phenomenon. Musical events have many perceived attributes including pitch, onset time (the start point of the event), duration, timbre and so on. In addition, *sequences* of these attributes may have multiple relevant emergent dimensions. For example, pitch interval, pitch class, scale degree, pitch contour (rising, falling or unison) and many other derived features are important in the perception and analysis of pitch structure. To accommodate these properties of music into the model, we use a multiple viewpoint approach to music representation (Conklin & Witten, 1995). The modelling process begins by choosing a set of basic properties of musical events (e.g., pitch, onset, duration, loudness etc) that we are interested in predicting. As these basic features are treated as independent attributes, their probabilities are computed separately and the probability of a note is simply the product of the probabilities of its attributes. Each basic feature (e.g., pitch) may then be predicted by any number of models for different derived features (e.g., pitch interval, scale degree) whose distributions are combined using the same entropy-weighted scheme (Pearce *et al.*, 2005).

The use of long- and short-term models, incorporating models of derived features, the entropy-based weighting method and the use of a multiplicative (as opposed to a weighted linear or additive) combination scheme all improve the performance of IDyOM in predicting the pitches of unseen melodies; technical details of the model and its evaluation can be found elsewhere (Conklin & Witten, 1995; Pearce *et al.*, 2005; Pearce & Wiggins, 2004). The goal in the current work, however, is to test its performance in retrieving segmentation boundaries in large corpora of melodies. Here, we use the model to predict the pitch, IOI and OOI associated with melodic events, multiplying the probabilities of these attributes together to yield the

overall probability of the event. For simplicity, we use no derived features. We then focus on the unexpectedness of events (information content, h) using this as a boundary strength profile from which we compute boundary locations, as described in §2.4.

Peak Picking

To convert the boundary strength profile produced by IDyOM into a concrete segmentation, we devised a simple method that achieves this using three principles. First, given a vector S of boundary strengths for each note in a melody, the note following a boundary should have a greater or equal boundary strength than the note following it: $S_n \geq S_{n+1}$. 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 to be k standard deviations greater than the mean boundary strength computed in a linearly weighted window from the beginning of the piece to the preceding event:

$$S_n > k\sqrt{\frac{\sum_{i=1}^{n-1}(w_i S_i - \overline{S}_{w,1\ldots n-1})^2}{\sum_1^{n-1} w_i}} + \frac{\sum_{i=1}^{n-1} w_i S_i}{\sum_1^{n-1} w_i}. \tag{8}$$

where w_i are the weights associated with the linear decay (triangular window) and the parameter k is allowed to vary depending on the nature of the boundary strength profile.

3 Method

3.1 The Ground Truth Data

The IDyOM model was tested against existing segmentation models on a subset of the EFSC, database `Erk`, containing 1705 Germanic folk melodies encoded in symbolic form with annotated phrase boundaries which were inserted during the encoding process by folk song experts. The dataset contains 78,995 note events at an average of about 46 events per melody and overall about 12% of notes fall before boundaries (a boundary occurs between two notes). There is only one hierarchical level of phrasing and the phrase structure exhaustively subsumes all the events in a melody.

3.2 The Models

The models included in the comparison are as follows:

Grouper: as implemented by Temperley (2001);[5]
LBDM: as specified by Cambouropoulos (2001) with $k = 0.5$;
IDyOM: as specified in §2.4 with $k = 2$;
GPR2a: as quantified by Frankland & Cohen (2004) with $k = 0.5$;
GPR2b: as quantified by Frankland & Cohen (2004) with $k = 0.5$;
GPR3a: as quantified by Frankland & Cohen (2004) with $k = 0.5$;
GPR3d: as quantified by Frankland & Cohen (2004) with $k = 2.5$;

[5] Adapted for use with Melconv 2 by Klaus Frieler.

TP: as defined in §2.3 with $k = 0.5$;
PMI: as defined in §2.3 with $k = 0.5$;
Always: every note falls on a boundary;
Never: no note falls on a boundary.

The Always model predicts a boundary for every note while the Never model never predicts a boundary for any note. Grouper outputs binary boundary predictions. These models, therefore, do not use the peak-picking and are not associated with a value of k. The output of every other model was processed by Simple Picker using a value of k chosen from the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$ so as to maximise F1 (and secondarily Recall in the case of ties).

The DOP method (Bod, 2001) is not included due to the complexity of its implementation and lack of any third party software that is straightforwardly applicable to musical data.

The IDyOM, TP and PMI models were trained and evaluated on melodies taken from the `Erk` dataset. In order to demonstrate generalisation, we adopted a cross-validation strategy in which the dataset is divided into k disjoint subsets of approximately equal size. The model is trained k times, each time leaving out a different subset to be used for testing. A value of $k = 10$ was used which has been found to produce a good balance between the bias associated with small values of k and the high variance associated with large values of k (Kohavi, 1995).

3.3 Making Model Outputs Comparable

The outputs of the algorithms tested vary considerably. 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. In contrast to Bruderer (2008) we chose to make all segmentation algorithms comparable by picking binary boundary indications from the boundary strength profiles.

To do so, we applied the peak-picking procedure described in S2.4 to the boundary profiles of all models (except Grouper which produces binary boundary judgements) and chose a value of k to optimise the performance of each model individually. In practice, the optimal value of k varies between algorithms depending on the nature of the boundary strength profiles they produce.

In addition, we modified the output of all models to predict an implicit phrase boundary on the last note of a melody.

3.4 Evaluation Measures

It is common to represent a segmentation of a melody using a binary vector with one element for each event in the melody indicating, for each event, whether or not that event falls on a grouping boundary. An example is shown in Figure 1.

Given this formulation, we can state the problem of comparing the segmentation of a model with the ground truth segmentation in terms of computing the similarity or distance between two binary vectors. Many methods exist for comparing binary vectors. For example, version 14 of the commercial statistical software package SPSS provides 27 different measures for determining the similarity or distance between binary variables. Additional measures have been proposed in the areas of data mining and psychological measurement. The appropriate measure to use depends on the desired comparison and the nature of the data (Sokolova & Lapalme, 2007). Here we introduce and compare five methods that are widely used in psychology, computer science and biology.

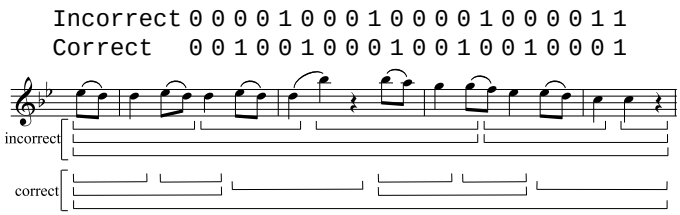

Fig. 1. An example showing the binary vectors representing the segmentation of a melody.

Table 2. A summary of the outcomes of comparing prediction and ground truth binary data.

		Ground Truth	
		P	N
Prediction	p	TP	FP
	n	FN	TN

These methods enable us to compute the similarity between phenomenal data encoded as a binary vector, the *ground truth*, and the output of a model of the process generating that data, the *prediction*, encoded in the same way.

All methods start with the 2 x 2 table shown in Table 2 which summarises the co-occurrences of binary events between the ground truth and the prediction. The ground truth positives (P) and negatives (N), respectively, are the numbers of positions where the ground truth vector contains 1 and 0. The predicted positives (p) and negatives (n) indicate numbers of positions where the prediction vector contains 1 and 0 respectively. The *true positives (TP)* is the number of positions where both ground truth and prediction vectors indicate 1 while the *true negatives (TN)* is the number of positions where both vectors contain 0. *False positives (FP)* and *false negatives (FN)* are the numbers of locations where the ground truth and prediction vectors differ. In the former case, the prediction contains 1 where the ground truth contains 0, and *vice versa* for the latter.

One of the most intuitive measures for comparing binary vectors is *accuracy*, defined as the number of times the prediction vector and ground truth vector agree as a proportion of the total number of entries in the vector:

$$accuracy = \frac{TP+TN}{P+N} \tag{9}$$

However, this measure of *accuracy* can be misleading when the ground truth data is skewed. For example, if the proportion of negative cases in the ground truth is .8, a model that always gives a negative answer will achieve an accuracy of 80%. The following measures take into account the proportion of positive and negative instances in the ground truth data which means that the values are comparable across the distributions occurring in different datasets.

Psychologists are often interested in the agreement between human raters or judges when they assess the same items and Kappa (κ) has become one of the most frequently used measures for assessing inter-rater agreement. It is conceptually related to the accuracy measure but takes the distribution of the two binary classes into account and thus resembles the well-known χ^2 distribution. The variant known as *Fleiss'* κ (Fleiss, 1971) is formulated for multiple-class

ratings and multiple raters. Reducing κ to binary markings from only two sources (raters) and using the notation introduced above, κ is defined as the difference between the proportions of actual agreement ($Pr = accuracy$) and expected agreement (Pr_e):

$$\kappa = \frac{Pr - Pr_e}{1 - Pr_e} \tag{10}$$

where:

$$Pr = \frac{TP + TN}{P + N}, \quad Pr_e = Pr_1^2 + Pr_0^2, \tag{11}$$

$$Pr_1 = \frac{P + p}{2 \cdot (P + N)}, \quad Pr_0 = \frac{N + n}{2 \cdot (P + N)}. \tag{12}$$

$$\tag{13}$$

Another measure, d' (Green & Swets, 1966), was developed in psychophysics and is often used to measure human ability to detect a particular cue in a signal or distinguish two stimuli differing along some dimension. It has been also widely used to analyse experimental data in other areas of cognitive psychology such as memory. It is defined as:

$$d' = z(\frac{TP}{TP + FN}) - z(\frac{FP}{FP + TN}) \tag{14}$$

where $z()$ is the cumulative distribution function of the normal probability distribution.

In modern data mining, the following three measures are standard methods for evaluating query-based systems for document retrieval (Witten & Frank, 1999). *Precision* reflects the true positives as a proportion of the positive output of the prediction while *Recall* reflects the true positives as a proportion of the positive data in the ground truth. *F1* is the harmonic mean of the two.

$$\begin{aligned} Precision &= \frac{TP}{TP + FP}, \\ Recall &= \frac{TP}{TP + FN}, \\ F1 &= \frac{2 \cdot precision \cdot recall}{precision + recall}. \end{aligned}$$

4 Results

Before comparing the performance of the models, it is instructive to consider the problem of how to evaluate quantitatively the degree of correspondence between two segmentations of a melody. To do so, we compute the Pearson correlation coefficients between the different evaluation measures described in §3.4 for each pairwise comparison between each models output for each melody in the dataset. The results are shown in Table 3.

Precision and Recall each only take into consideration one kind of error (i.e., FP or FN) and show low or moderate correlations with the other measures (and very low correlations with each other as expected). Here, however, we want a measure that takes into account both kinds of error. κ, F1 and d' all correlate very highly with each other because they all reflect TP in relation to FP and FN. Although κ is also influenced by TN, the proportion of true

Table 3. Correlations between evaluation measures over models and melodies.

	Accuracy	Precision	Recall	F1	d'	κ
Accuracy	1					
Precision	0.56	1				
Recall	-0.31	0.08	1			
F1	0.45	0.69	0.63	1		
d'	0.52	0.48	0.64	0.91	1	
κ	0.86	0.70	0.17	0.83	0.84	1

Table 4. The model comparison results in order of mean F1 scores. See text for details of the Hybrid model.

Model	Precision	Recall	F1
Hybrid	0.87	0.56	0.66
Grouper	0.71	0.62	0.66
LBDM	0.70	0.60	0.63
IDyOM	0.76	0.50	0.58
GPR2a	0.99	0.45	0.58
GPR2b	0.47	0.42	0.39
GPR3a	0.29	0.46	0.35
GPR3d	0.66	0.22	0.31
PMI	0.16	0.32	0.21
TP	0.17	0.19	0.17
Always	0.13	1.00	0.22
Never	0.00	0.00	0.00

negatives is constrained given a fixed number of data points (i.e. if we know TP, FP, and FN and the total number of notes then TN is fixed; we have 3 degrees of freedom and not 4 for pairs of vectors of the same length). Accuracy exhibits only small correlations with these three measures (except κ to which it is closely related) and is not appropriate here due to the unequal proportions of positive and negative values in the data (see §3.4). The results of the correlational analysis suggest that we could have used any one of d', F1 or κ for evaluating our models against the ground truth. Following common practice in data mining and information retrieval, we use *F1* to compare model performance.

The results of the model comparison are shown in Table 4. The four models achieving mean F1 values of over 0.5 (Grouper, LBDM, GPR2a, IDyOM) were chosen for further analysis. Sign tests between the F1 scores on each melody indicate that all differences between these models are significant at an alpha level of 0.01, with the exception of that between GPR2a and LBDM. In order to see whether further performance improvement could be achieved by a combined model, we constructed a logistic regression model including Grouper, LBDM, IDyOM and GPR2a as predictors. Backwards stepwise elimination using the Bayes Information Criterion (BIC) failed to remove any of the predictors from the overall model (Venables & Ripley, 2002). The performance of the resulting model is shown in the top row of Table 4. Sign tests demonstrated that the Hybrid model achieved better F1 scores on significantly more melodies than each of the other models (including Grouper, in spite of the fact that the average perfor-

mance, shown in Table 4, was the same). Compared to Grouper and LBDM, the hybrid model has slightly worse recall but much better precision; compared to IDYOM, the hybrid model has better precision and recall; while compared to GPR2a, the lower precision achieved by the hybrid model is balanced by it's better recall.

5 Discussion

We would like to highlight four results of this evaluation study. First, we were surprised by the strong performance of one of the GTTM preference rules, GPR2a. This points to the conclusion that rests, perhaps above all other melodic parameters, have a large influence on boundaries for this set of melodies. Consequently, all of the high-performing rule-based models (Grouper, LBDM, GPR2a) make use of a rest or temporal gap rule while IDyOM includes rests in its probability estimation. Future research should undertake a more detailed qualitative comparison of the kinds of musical context in which each model succeeds or fails to predict boundaries. This suggests that future research should focus on boundaries not indicated explicitly by rests.

Second, it is interesting to compare the results to those reported in other studies. In general, the performance of Grouper and LBDM are comparable to their performance on a different subset of the EFSC reported by Thom *et al.* (2002). The performance of Grouper is somewhat lower than that reported by Temperley (2001) on 65 melodies from the EFSC. The performance of all models is lower than that of the supervised learning model reported by Bod (2001).

Third, the hybrid model which combines Grouper, LBDM, GPR2a and IDyOM generated better performance values than any of its components. The fact that the *F1* value seems to be only slightly better than Grouper is due to the fact that logistic regression optimises the log-likelihood function for whether or not a note is a boundary given the boundary indications of the predictor variables (models). It therefore uses information about positive boundary indications (*P*) and negative boundary indications (*N*) to an equal degree, in contrast to *F1*. This suggests options, in future research, for assigning different weights to *P* and *N* instances or including the raw boundary profiles of LBDM and IDyOM (i.e., without peak-picking) in the logistic regression procedure. Another possibility is to use boosting (combining multiple weak learners to create a single strong learner, Schapire, 2003) to combine the different models which may lead to better performance enhancements than logistic regression.

Finally, it is interesting to note that an unsupervised learning model (IDyOM) that makes no use of music-theoretic rules about melodic phrases performed as well as it does. It not only performs much better than simple statistical segmenters (the TP and PMI models) but also approaches the performance of sophisticated rule-based models. In fact, IDyOM's precision is better than LBDM and Grouper although it's Recall is worse (this is a common tradeoff in MIR). In comparison to supervised learning methods such as DOP, IDyOM does not require pre-segmented data as a training corpus. This may not be an issue for folk-song data where we have large corpora with annotated phrase boundaries but is a significant factor for other musical styles such as pop. IDyOM learns regularities in the melodic data it is trained on and outputs probabilities of note events which are ultimately used to derive an information content (unexpectedness) for each note event in a melody. In turn, this information-theoretic quantity (in comparison to that of previous notes) is used to decide whether or not the note falls on a boundary.

These findings have been corroborated by a recent study comparing computational models of melodic segmentation to perceived segmentations indicated by human listeners for 10

popular melodies (de Nooijer *et al.*, 2008). The results showed that IDyOM's segmentations did not differ significantly from those of the listeners and, furthermore, that the segmentations of IDyOM, LBDM and Grouper did not differ.

We argue that the present results provide preliminary evidence that the notion of expectedness is strongly related to boundary detection in melodies. In future research, we hope to achieve better performance by tailoring IDyOM specifically for segmentation including a metrically-based (i.e., we represent whatever is happening in each metrical time slice) 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 & Plumbley, 2009).

Acknowledgement. This research was supported by EPSRC via grant numbers GR/S82220/01 and EP/D038855/1.

References

Abdallah, S. & Plumbley, M. (2009). Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, *21*(2-3), 89–117.

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.

Ahlbäck, S. (2004). *Melody beyond notes: A study of melody cognition.* Doctoral dissertation, Göteborg University, Göteborg, Sweden.

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

Barlow, H. & Morgenstern, S. (1949). *A dictionary of musical themes*. Ernest Benn.

Bell, T. C., Cleary, J. G., & Witten, I. H. (1990). *Text Compression*. Englewood Cliffs, NJ: Prentice Hall.

Bod, R. (1998). *Beyond Grammar: An experience-based theory of language*. Standford, CA: CSLI Publications.

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

Bower, G. (1970). Organizational factors in memory. *Cognitive Psychology*, *1*, 18–46.

Bregman, A. S. (1990). *Auditory Scene Analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71–105.

Brent, M. R. (1999b). 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.

Bruderer, M. J. (2008). *Perception and Modeling of Segment Boundaries in Popular Music*. Doctoral dissertation, J.F. Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven, Nederlands.

Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, *40*(2/3), 76–93.

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.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organisation. *Psychological Review*, *103*(3), 566–581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology*, *52A*(2), 273–302.

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

Cleary, J. G. & Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*(2/3), 67–75.

Cohen, P. R., Adams, N., & Heeringa, B. (2007). Voting experts: An unsupervised algorithm for segmenting sequences. *Intelligent Data Analysis*, *11*(6), 607–625.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Doctoral dissertation, Department of Computer and Information Science, University of Pennsylvania, USA.

Conklin, D. & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73.

de Nooijer, J., Wiering, F., Volk, A., & Tabachneck-Schijf, H. J. M. (2008). An experimental comparison of human and automatic music segmentation. In K. Miyazaki, M. Adachi, Y. Hiraga, Y. Nakajima, & M. Tsuzaki (Eds.), *Proceedings of the 10th International Conference on Music Perception and Cognition* (pp. 399–407). Adelaide, Australia: Causal Productions.

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.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Ferrand, M., Nelson, P., & Wiggins, G. (2003). Memory and melodic density: a model for melody segmentation. In N. G. F. Bernardini & N. Giosmin (Eds.), *Proceedings of the XIV Colloquium on Musical Informatics* (pp. 95–98). Firenze, Italy.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 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*(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.

Green, D. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

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

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672.

Howell, D. C. (2002). *Statistical methods for pscyhology*. Pacific Grove, CA: Duxbury.

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

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt, Brace and World.

Kohavi, R. (1995). *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. Doctoral dissertation, Department of Computer Science, Stanford University, USA.

Ladefoged, P. & Broadbent, D. E. (1960). Perception of sequences in auditory events. *Journal of Experimental Psychology*, *12*, 162–170.

Larsson, N. J. (1996). Extended application of suffix trees to data compression. In J. A. Storer & M. Cohn (Eds.), *Proceedings of the IEEE Data Compression Conference* (pp. 190–199). Washington, DC: IEEE Computer Society Press.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *16*(3), 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*(10), 1853–1867.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–296.

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

Melucci, M. & Orio, N. (2002). A comparison of manual and automatic melody segmentation. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval* (pp. 7–14). Paris, France: IRCAM.

Meyer, L. B. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

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.

Pearce, M. T., Conklin, D., & Wiggins, G. A. (2005). Methods for combining statistical models of music. In U. K. Wiil (Ed.), *Computer Music Modelling and Retrieval* (pp. 295–312). Berlin: Springer.

Pearce, M. T. & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385.

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.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–285.

RISM-ZENTRALREDAKTION. Répertoire international des sources musicales (rism).

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

Schaffrath, H. (1995). The Essen folksong collection. In D. Huron (Ed.), *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide [computer database]*. Menlo Park, CA: CCARH.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear Estimation and Classification*. Berlin: Springer.

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

Sokolova, M. & Lapalme, G. (2007). Performance measures in classification of human communication. In Z. Kobti & D. Wu (Eds.), *Advances in Artificial Intelligence: 20th Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 159–170). Berlin: Springer.

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 International Computer Music Conference* (pp. 65–72). San Francisco: ICMA.

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

Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica*, *14*(3), 249–260.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Waugh, N. & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89–104.

Witten, I. H. & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, *37*(4), 1085–1094.

Witten, I. H. & Frank, E. (Eds.). (1999). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.