A Probabilistic Model for Melody Segmentation

Miguel Ferrand[1]**, Peter Nelson[1], and Geraint Wiggins[2]

[1] University of Edinburgh, Scotland, UK,
{mferrand,pwn}@music.ed.ac.uk,
http://www.music.ed.ac.uk/
[2] City University, London, UK
geraint@soi.city.ac.uk,
http://www.soi.city.ac.uk/

Abstract. In this paper we propose that a probabilistic model of music listening may be used to predict segmentation boundaries in melodies,, as perceived by a listener. Existing models of music perception usually achieve a structural segmentation of a music piece based on Gestalt–based local discontinuities and on the detection of parallelism. The assimilation of regularities in music contributes to expectations created during the course of listening, and is reflected in the listener's ability (or inability) to predict what comes next. We conjecture that the expectations associated with intra-opus musical information provide strong hints for segmentation points within a piece. We describe an implementation of this model and analyse a preliminary segmentation experiment, discussing the limitations and the possible developments of this approach.

1 Introduction

When listening to music, subjects often perceive divisions in the musical discourse. The identification of several parts or segments in a piece is an important step for abstracting musical contents. Several theories have recognised music segmentation as an important part of music understanding, and have attempted to explain and formalise how listeners' intuitions account for the identification of the pieces' constituent units such as motives, phrases or sections.

Some of these theories [1–3] employ Gestalt principles to identify discontinuities or create note groupings. Although grouping principles have been found to have a reasonable explanatory power [4, 5], most theories that use Gestalt principles for segmentation often rely on higher-level rules to form larger groupings or to identify parallelisms.

Deliège and Melén [6] argued for the prototypical nature of parallelism and showed that descriptions of sections of a musical piece can be formed and retained by listeners, based on the repetition and salience of small musical patterns. These small patterns constitute indexes for larger sections and their salience is enhanced by the familiarity of the listener with the music. Familiarity with a certain musical material may result from the assimilation of a particular musical

** funded by EPSRC research project GR/N08049/01

style or repertoire, or from recent or repetitive audition and memorisation of a particular piece.

Leonard Meyer affirmed that "suspense is essentially a product of ignorance as to the future course of events" [7] outlining the relationship between acquired knowledge and expectation. Later he recognised the affinity between expectation and information theory and in particular with the notion of entropy [8]. The unpredictability of an event in a sequence of events, can alter its prominence. An unpredictable musical event is more noticeable to the listener and therefore more likely to be remembered. On the same line of thought Huron [9] argues that to establish feature salience in a musical work, one has to identify those characteristics that present intratextual or intertextual distinctiveness since "the mere presence of some element or property does not necessarily make it a good feature. A good feature must in some ways draw attention to itself".

Researchers have used the notion of entropy to model and evaluate musical composition [10, 11] or to measure musical learning [12], but to our best knowledge, none have used this concept to model music segmentation.

This paper proposes the use of a probabilistic approach to predict segmentation boundaries in melodies. We argue that musical segments are not always clearly characterised parts of a musical piece, such as those related by similarity. In the course of listening, the absence of references during a prolonged temporal interval, may lead the listener to recall that interval as a segment in the music for which no clear understanding was experienced. In other words, parts of a musical piece which, to the listener, lack an identity or a particular character, may also be identified as distinct segments.

2 Overview of the Proposed Model

2.1 General principles

We view music as a multidimensional phenomenon, where several features (e.g. pitch related or rhythm related) unfold simultaneously in time. At different moments within a piece, different musical features may be the source of discontinuities or perceived saliences. A piece of music contains more information than a listener can process in a single hearing, therefore listening implies choosing which elements to attend to, from time to time [13].

Krummansl [14] found that in a task of musical segmentation, listeners identified boundaries mainly on the basis of a combination of several different musical characteristics. In the absence of evidence to distinguish quantitatively the contribution of different musical features, we propose only to identify which ones are salient or not salient, at a particular moment. If two or more different features are found to be salient simultaneously, at a particular point, then they will add up to constitute a stronger salient moment.

In this work, and as explained previously, we associate feature salience with expectation. We will use the entropy as a measure of unpredictability associated with different musical features. Low entropy usually means high predictability

but if a particular feature (e.g. note duration) is highly predictable throughout the piece then it may well be because it is either highly invariant or because it follows a monotonous variation pattern. For example if a whole melody is layered on semi quavers, we can say that rhythm is highly predictable, but it provides no references for the segmentation of the melody. For this reason we are not interested in measuring the overall entropy of the model, but rather how entropy changes along the piece. We conjecture that transitions between high and low entropy constitute salient moments in a listening experience. Furthermore we argue that musical parameters with varying entropy along the piece are more informative than parameters with consistently high or low entropy values.

2.2 N-gram models

The implementation of the model is based on an n-gram grammar. N–gram grammars are n^{th} order Markov models that assume that the probability of occurrence of a symbol depends on the prior occurrence of $n-1$ other symbols. N–gram models are typically constructed from statistics obtained from a large corpus of data (usually referred to as the training corpus) using the co-occurrences of symbols to determine the probabilities of sequences of symbols.

Hence, the probability $P(s)$ of a sequence $s = w_1...w_l$ of length l is given by,

$$P(s) = \prod_{i=1}^{l} P(w_i | w_{i-n+1}^{i-1}) \tag{1}$$

where w_i^j denotes the sequence $w_i...w_j$ and n is the order of the model [3].

Independently of the size of the training corpus, it is unlikely that all possible symbol sequences will occur. Data sparseness becomes a problem if, when computing probabilities using Equation 1 some of the terms in the product have zero probability. Also, if the training corpus is small, and the order of the model is significantly high, longuer sequences will have relatively lower counts, resulting in less accurate probabilities.

Several methods, usually referred to as *smoothing* methods, have been described in the literature [15, 16] to overcome the data sparseness problem, and estimate probabilities. In this work we are focusing only on intra-opus information meaning that the amount of data to be analysed is substantially lower than if we were using a larger corpus of pieces, so a linear interpolation smoothing method [17] was employed. Using linear interpolation the probabilities of a sequence of length l can be estimated by a weighted sum of n–gram probabilities from models of order $n <= l$. For instance, the probability of a tri-gram is determined by the weighted sum of corresponding uni-gram, bi-gram and tri-gram probabilities,

$$P(w_k | w_{k-2}, w_{k-1}) = \lambda_1 P(w_k) + \lambda_2 P(w_k | w_{k-1}) + \lambda_3 P(w_k | w_{k-2}, w_{k-1}) \tag{2}$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1 < \lambda_2 < \lambda_3$ as it is assumed that longer contexts, being more specific, should have a higher weight.

[3] when $n > i$ padding symbols have to be introduced to provide the necessary contexts

2.3 Entropy

The fundamentals of Information Theory (IT) were first introduced in [18], and set up quantitative ways of measuring the information contained in a message being transmitted, received, or stored. One of the ways of measuring the quantity of information of a particular message is to determine its unpredictability or entropy. For a given N-gram model M, entropy associated with a given context c can be determined by,

$$H_c(M) = -\sum_{\forall e:(c,e)\in M} P(e|c)\log_2 P(e|c) \tag{3}$$

where e denotes all possible successor symbols of the context c. Contexts are sequences of size $N-1$ where N is the order of the model M. Conditional probabilities are calculated using Equation 2.

Since we are interested is observing the changes in entropy along a sequence of symbols, the occurrence of every new symbol in the sequence provides a new context for which the values of $H_c(M)$ can be calculated.

3 A case study

Seeking to compare the proposed model with segmentation data provided by real listeners we used some data described in a segmentation experiment carried out by Deliège [19]. In these experiments subjects listened to a melody (the solo for English Horn, from Wagner's opera Tristan and Isolde) and had to identify segmentation points in real–time. Both musically trained and untrained subjects took part in the experiments. A familiarisation audition of the piece preceded the auditions during which subjects were asked to identify segmentation boundaries. The experiments revealed a set of 8 main segment boundaries (identified by most subjects) and an additional set of 13 weaker boundaries. For the present study only the stronger boundaries were used for comparison. For full details of the experimental procedure the reader is referred to [19].

The melody information was translated into an event–based representation. All events are numbered sequentially and gather information about pitch (Midi note code), duration and onset time. From these basic event attributes, four other features were extracted and associated with each event:

- Pitch step (PS): expresses the interval distance to following event in semi–tones.
- Pitch contour (PC): expresses the sign of the pitch step; takes value -1,+1 or 0 if PS is also 0.
- Duration ratio (DR): expresses the ratio between the durations of the present and the following event.
- Duration contour (DC): expresses duration ratio changes; takes values -1 if $DR > 1$, 1 if $DR < 1$ or 0 if $DR = 1$.

Table 1. The first 14 events of the melody and features extracted: pitch step (PS); pitch change (PC); duration contour (DC) and duration ratio (DR)

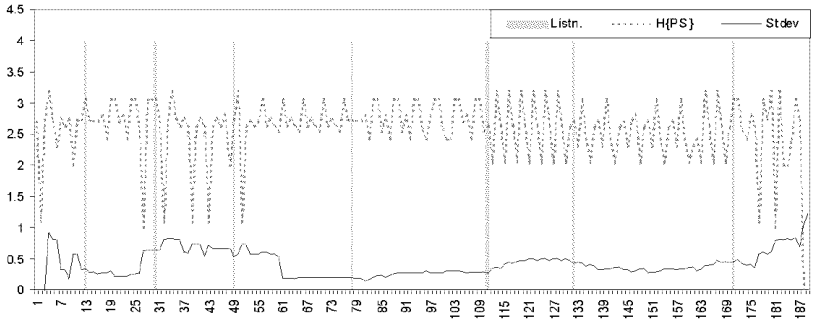
No	Midi	Dur	OnSet	PS	PC	DC	DR
1	53	2.000	0.000	7	1	1	1.250
2	60	2.500	2.000	3	1	-1	0.200
3	63	0.500	4.500	-2	-1	1	3.000
4	61	1.500	5.000	-5	-1	-1	0.333
5	56	0.500	6.000	7	1	1	3.000
6	63	1.500	6.500	-8	-1	-1	0.333
7	55	0.500	8.000	5	1	0	1.000
8	60	0.500	8.500	-7	-1	0	1.000
9	53	0.500	9.000	5	1	1	3.000
10	58	1.500	9.500	2	1	-1	0.333
11	60	0.500	10.500	-2	-1	1	4.000
12	58	2.000	11.000	-2	-1	0	1.000
13	56	2.000	13.000	-1	-1	-1	0.167
14	55	0.333	15.000	-2	-1	0	1.000
15	53	0.333	15.333	-2	-1	0	1.000
..	..	...	...	.	.	.	...

Table 1 shows an extract of the encoded melody of the Horn solo, with the additional four attributes.

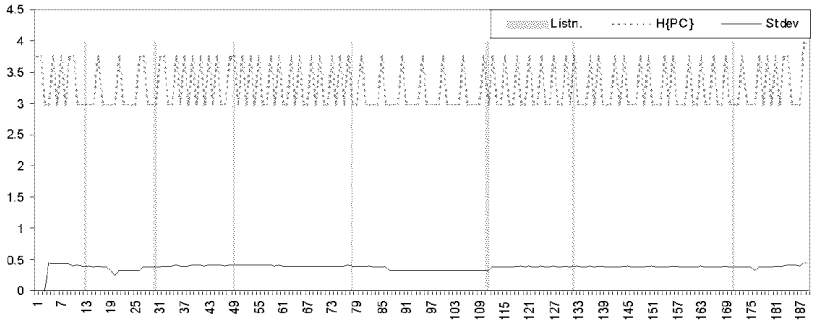
The values obtained for the attributes PS, PC, DC and DR, constitute four sequences of symbols from which sequence probabilities can be generated. A tri–gram, bi–gram and uni–gram model was generated for each one of the four sequences.

The entropy values were obtained from each one of the models, and for all events in the melody. In Figure 1 we show the entropy profiles obtained from a tri–gram model. The vertical grey lines overlapped in the graph indicate the locations of the stronger boundaries indicated by the listeners in Deliège's experiment [19]. The standard deviation (*stdev*) of the entropy is also depicted in the lower part of each graph. Standard deviation gives a good measure of the spread of the entropy values along the graph and since we are interested in measuring the changes in entropy along the piece, the *stdev* is calculated with a sliding window. In this experiment we used a fixed size window of 10 events although we suggest that the size of window could be determined in terms of time and not in number of events. This would seem perceptually more realistic and the sliding window, with a fixed duration, could be seen as a short–term memory time frame within which changes can be perceived by the listener. The number of events that would fit in this window would depend on the tempo assigned to the piece. Further research is necessary to corroborate this idea.

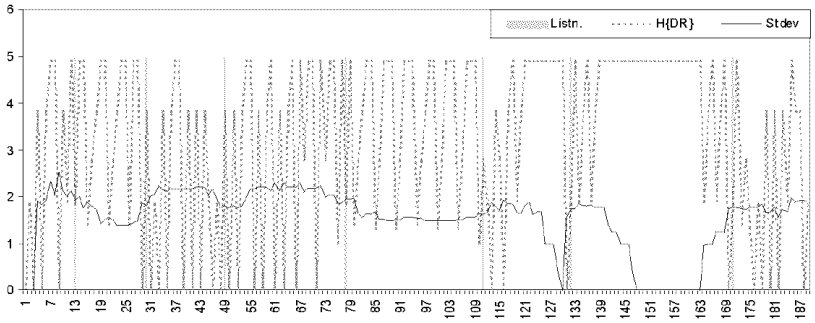
In a first observation of the graphs of Figure 1 it seems clear that duration based features register a much higher entropy variance along the melody than pitch based features. Following our conjecture, time based features are then



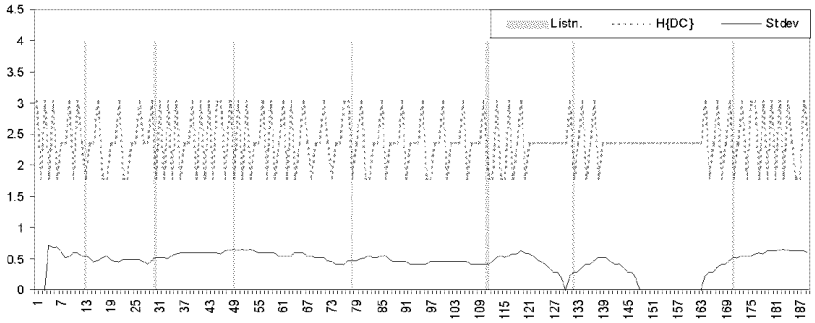
a) Pitch step entropy



b) Pitch contour entropy



c) Duration ratio entropy



d) Duration contour entropy

Fig. 1. Entropy(dotted line) and entropy moving standard deviation (solid line) for PS, PC, DC and DR *vs.* event number. Listeners' segmentation boundaries are indicated by grey vertical lines

likely to convey more information to the listener regarding segmentation. In fact it can be observed that the DR entropy graph exhibits accentuated variations, many of which occur in the vicinity of the boundaries indicated by listeners. This is confirmed by Deliège in the analysis of her experimental results, where it is reported that the listeners' decision were dominated by Gestalt principles of proximity which are more evident in the rhythmic content of this melody. The graphs of PS and PC, although with overall low *stdev* display different entropy variation patterns within many of the segments identified by the listeners in Deliège's experiment.

At present we do not yet have a method to automatically interpret and extract these boundaries from the entropy profiles obtained from the melodies. This part of the model will be contemplated in further developments.

4 Discussion

The use of n–gram models is often criticised for the underlying assumption that a state depends only on the previous states. This assumption seems to be over–simplistic if we are analysing musical sequences, however it is known that human memory limitations impose a limit on the ability to establish large–span temporal relations. This has been acknowledged in [1, 20] where it has been suggested that listeners perceive a musical surface by focusing on successive zones, along the musical piece.

As mentioned before, parallelism has a strong influence in the establishment of boundaries in a melody so the perception of similarities cannot be ignored if we want to model segmentation. A probabilistic model, like the one presented here, can capture some parallelisms but it is limited by the order of the model, which determines the length of the patterns that can be stored and recalled as similar. Because only identical sequences are recorded as repetitions, this probabilistic model can, in principle, only deal with exact similarity. However, it can be argued that because several features were separated into more general descriptors (e.g. duration ratio or pitch contour as opposed to absolute duration or pitch values) we can accommodate some form of approximate similarity. For example two sequences can have the same relative durations, although different absolute note durations. Although large patterns cannot be stored as a whole by a low order model, we argue that parallelism may still be established based on smaller parts of these patterns. Empirical studies [21] have shown that primacy effects were present in the recognition of similar patterns, and that patterns are often remembered by the repetition of smaller patterns, which are often their initial sections. This evidence supports the conjecture that a short–term memory model can capture some structural information based on parallelism, provided there is some regularity in the acquired musical data.

In this work, only intra-opus information was used, meaning that the model only capture regularities within a given piece. The results shown were obtained with a model of order 3. As expected, and due to the fairly small set of input data, increasing the size of the sequences stored by the model decreases the

overall pattern count, and therefore compromising probability estimation. The use of interpolation and the unfolding of the melodic information in several features provided additional redundancy but not enough to accommodate very long context models. Additional experiments are necessary to find out how the order of the model influences the granularity of the entropy profiles.

The use of inter–opus in conjunction with intra-opus information was suggested in [10], where two separate models are combined to make predictions, although the way these two context models were actually combined has not been described in detail. Intuitively, it seems that regularities particular to a musical piece could override the intuitions resulting from long term established rules. The long–term model provides the 'norms' (obtained from a large corpus of pieces) and the short–term model, obtained from the piece being heard, provides the listener with confirmations or deviations from the rules.

4.1 Related work

An important contribution of the present approach is that it attempts to predict segmentation boundaries from non–annotated musical data. Most other related approaches include style–dependent knowledge or use pre-annotated training data.

Conklin and Witten's [10] multiple viewpoint system for generation of Bach chorales uses a training corpus which includes score based information such as time signatures, fermatas, location of bar lines, *etc.* Ponsford et al. [11] use a probabilistic approach based on N-grams to capture and generate harmonic sequences, but also assume from the start the use of a score–based music representation.

Bod [22] proposes the use of a Markov Grammar to learn and predict phrase boundaries in folk songs. Learning is based on a training set of pre-annotated pieces obtained from the Essen folk song database. The phrase boundaries indicated in the Essen folk songs have not been validated with listeners, and the author acknowledges that the correction of the annotations should preferably be established by an independent psychological experiment with more than one subject. This raises the question whether the model is really predicting listening behaviour or just predicting boundaries according to particular analytical criteria, reflected in the annotations of the pieces in the database? To answer this question it would be necessary to test a sample of songs from the Essen database with listeners, and find out how the phrase boundaries indicated by listeners would differ from the annotated ones. In short and simple pieces, where parallelism is more obvious, it is likely that the structural segments perceived by listeners correspond to the sections obtained by simple analysis of the pieces. However in longer pieces, and when parallelism is more difficult to establish, either by the temporal separation between motives or by the subtlety of the similarities, models should be compared with results obtained by listeners.

Reis' [12] research aimed to determine to what degree a system without any *a priori* stylistic information, is able to gain proficiency in a given musical style, as measured by its ability to predict the music. The author extends the approach

based on context models [10] to simulate the on–line process of music learning and capture the stylistic information present in a musical surface. He argues for a more cognitively pertinent way of inducing the contexts, using Gestalt–based perceptual cues (e.g. changes of direction, or large jumps in either pitch or time domain) to restrict the number of sequences that are extracted and stored from the training set.

There are advantages in including contextual information in an analytical process. For example, Bod [22] has shown how to improve the performance of his model by limiting the maximum number of phrase boundaries the parser can identify within a song (this maximum may be obtained directly from the Essen database). The drawback of this sort of approach is that the model becomes too biased towards a particular repertoire, so it is likely that the predictive power may drop when parsing pieces from other repertoires. In fact, in the particular case of the Essen folk song database, it is likely that the number of segments may vary significantly according to the origin or type of the songs.

5 Conclusions

This study suggests that some structural information about a melody can be associated and induced by changes in expectation. It was found that distinct changes in entropy associated with different musical features were coincident with melody segment boundaries indicated by listeners. It was also shown that the statistical properties of the entropy profiles may be used to indicate which parts of a melody, or more generally which features of a melody are more informative and therefore more likely to contribute to the perception of segmentation boundaries.

A central motivation of this work is to develop a model that can predict segmentation boundaries by learning from non–annotated data. Preliminary results reveal that the model has a significant predictive power, concerning the location of segmentation boundaries and thus encourages further developments and experimental research.

References

1. Lerdahl, F., Jackendoff, R.: A Generative Theory of Tonal Music. M.I.T. Press, Cambridge (Mass.) (1983)
2. Narmour, E.: The Analysis and Cognition of Basic Melodic Structures: The Implication–Realisation Model. University of Chicago Press, Chicago (1990)
3. Cambouropoulos, E.: Towards a General Computational Theory of Musical Structure. PhD thesis, University of Edinburgh (1998)
4. Deliège, I.: Grouping conditions in listening to music: an approach to Lerdahl and Jackendoff's grouping preference rules. Music Perception **4** (1987) 325–360
5. Bigand, E., Lerdhal, F., Pineau, M.: Deux approches expérimentales des quatre composants de la théorie generative de la musique tonale. In Deliége, I., ed.: Proceedings of the 3rd International Conference on Music Perception and Cognition, Liège, Belgium, European Society for the Cognitive Sciences of Music (1994) 259–260

6. Deliège, I., Melén, M.: Cue abstraction in the representation of musical form. In Deliège, I., Sloboda, J., eds.: Perception and Cognition of Music. Psychology Press (1997) 387–412
7. Meyer, L.B.: Emotion and Meaning in Music. University of Chicago Press (1956)
8. Meyer, L.B.: Music, The Arts, And Ideas – Patterns and Predictions in Twentieth–Century Culture. University of Chicago Press, Chicago (1967)
9. Huron, D.: What is a musical feature? Forte's analysis of Brahms's opus 51, no. 1, revisited. Music Theory On-line **7** (2001)
10. Conklin, D., Witten, I.: Multiple viewpoint systems for music prediction. Journal of New Music Research **24** (1995) 51–73
11. Ponsford, D., Wiggins, G., Mellish, C.: Statistical learning of harmonic movement. Journal of New Music Research **28** (1999)
12. Reis, Y.B.: Simulating music learning: On–line, perceptually guided pattern induction of context models for multiple-horizon prediction of melodies. In: Proceedings of AISB'99 - Symposium on Musical Creativity. (1999) 58–63
13. Aiello, R.: Can listening to music be experimentally studied? In Aiello, R., Sloboda, J., eds.: Musical Perceptions. Oxford University Press (1994) 273–282
14. Krumhansl, C.L.: Cognitive Foundations of Musical Pitch. Number 17 in Oxford Psychology Series. Oxford University Press, Department of Psychology, Cornel University (1990)
15. Manning, C.D., Schüttze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, Mass. (1999)
16. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modelling. In: Proceedings of the 34th Annual Meeting of the ACL. (1996)
17. Jelinek, F., Mercer, R.: Interpolation estimation of Markov source parameters for sparse data. Pattern Recognition in Practice (1980) 381–397
18. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379–423,623–656
19. Deliège, I.: Wagner "alte weise": Une approche perceptive. Musica Scientiæ**Special Issue** (1998) 63–90
20. Bigand, E.: Contributions of music to research on human auditory cognition. In McAdams, S., Bigand, E., eds.: Thinking in Sound: The Cognitive Psychology of Human Audition. Oxford University Press (1993) 231–277
21. Deliège, I.: Prototype effects in music listening: An empirical approach to the notion of imprint. Music Perception **18** (2001) 371–407
22. Bod, R.: Memory–based models of melodic analysis: Challenging the gestalt principles. Journal of New Music Research **30** (2001)