UNSUPERVISED LEARNING OF MELODIC SEGMENTATION: A MEMORY-BASED APPROACH

Miguel Ferrand, Peter Nelson
University of Edinburgh, UK

Geraint Wiggins
City University, London, UK

ABSTRACT

In this paper we propose a memory-based model for melodic segmentation. We argue that the perception of segment boundaries is related to the unpredictability of certain musical features and that feature salience can be learned from a corpus of non-annotated musical data. We describe the implementation of this model and how it uses the acquired information to predict the location of segment boundaries for a given melody. Finally we present some experimental results to show that the model has a significant predictive power regarding the location of segment boundaries, when compared with segment boundaries obtained with listeners.

1. BACKGROUND

When listening to a piece of music, listeners often identify distinct sections or segments within the piece. Music segmentation is recognised as an important step in the abstraction of musical contents and researchers have attempted to explain how listeners perceive and identify the boundaries of these segments.

Existing theories on music segmentation have employed Gestalt principles to identify discontinuities and create groupings between musical events [1,2]. The perception of parallelism and similarities are also known to influence the listener to relate different passages within a musical piece. In fact many Gestalt-based approaches rely on higher-level grouping rules or similarity functions to identify larger scale segment boundaries.

Often, it is suggested that Gestalt principles operate independently of the listeners musical knowledge. When familiarised with a certain musical repertoire, listeners memorise recurrent features in the music and use this knowledge to carry out musical analytical tasks. Empirical evidence has shown that large sections of a musical piece can be recalled by listeners based on the recurrence of small musical cells [3], which act as markers within the piece. Leonard Meyer in his theory on music expectation also outlines the importance of learning in music understanding and relates expectation with information theoretical notions such as entropy [4].

The entropy (or unpredictability) associated with the occurrence of a musical event can change its prominence and hence make it salient to the listener, within a sequence of events. The notion of salience has also been referred as being associated with features that present intra-textual or inter-textual distinctiveness [5]. Probabilistic methods have been widely used to acquire regularities in large sets of data, with many successful applications in natural language and speech processing [6]. Some of these methods have migrated into the music domain however, probabilistic modelling has been used mostly for music prediction and generation [7-9] and seldom to model musical analytical tasks [10] or listening behaviour.

2. AIMS

We seek the development of a system to learn and perform melodic segmentation in an unsupervised way. Learning from raw musical data (without annotations) and avoiding the use of *a priori* musical knowledge or musical rules are central motivations of the present work.

Applications for automatic music segmentation include the support to other analytical methods such as music search and pattern finding. The segments found can set the initial search points within a large piece, thus providing a reduction of the initial search space for complex algorithms.

3. A MODEL FOR MELODIC SEGMENTATION

We propose the implementation of a memory-based model to automatically predict the location of segmentation boundaries in a melody. The three main aspects of this model are described in this section. The first relates to the input of the model and deals with the representation of melodic information. The second and central part of the model is the feature learning module, which is implemented based on Markov models. The third relates to the output of the model and describes how it generates predictions about the location of segment boundaries, given a test melody.

3.1. Melody Representation

Music can be seen as a temporal process where sound events unfold in time. In this work, melody information is converted directly from a Midi source into an event-based symbolic representation including the pitch, duration and the inter-onset interval between events. From these basic attributes we obtained two additional melodic features:

- Pitch step (PS): the interval distance between consecutive events (in semitones).
- Duration ratio (DR): the ratio between the duration of consecutive events.

In Table 1 we show the melodic representation for an extract of Debussy's Syrinx, together with derived features PS and DR. For practical reasons the original DR values (in parenthesis) were converted into a logarithmic scale.

ISBN 3-931852-67-9
ISSN 1617-6847

No.	Pitch	Onset	Dur	PS	DR
1	82	3998	1000	-1	-5 (0.14)
2	81	4998	143	+2	-1 (0.80)
3	83	5141	115	-3	6 (8.70)
4	80	5256	1000	-1	-5 (0.17)
5	79	6256	167	+2	0 (0.99)
6	81	6423	166	-2	1 (1.64)
7	78	6589	273	-3	0 (1.18)
8	77	6862	322	…	…

Table 1. Melody representation and two derived features.

These two features have the advantage of representing melodic information in a relative manner, thus avoiding the use of absolute pitch values or absolute durations. The latter is particularly important since in an expressive (non-mechanical) performance the durations of Midi events do not correspond exactly to the notated durations.

3.2. Feature Learning with Markov Models

Markov 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 adjacent symbols to determine the probabilities of sequences of symbols.

An nth order n-gram model (a class of Markov models) assumes that the probability of occurrence of a symbol depends on the prior occurrence of *n-1* other symbols. Given a sequence $s=w_1...w_l$ of length l , the probability *P(s)* is given by,

$$P(s) = \prod_{i=1}^{l} P(w_i \mid w_{i-1},..., w_{i-n+1}) \qquad (1)$$

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

Another disadvantage of n-gram models is that their size increases rapidly with an increase in their order since we may need to store the probabilities of all combinations of fairly long sequences.

Next, we describe Mixed-order Models, which were used in the present work to overcome the increased order and data sparseness problems.

Mixed-order Markov Models

Mixed-order Markov Models (MMM) provide a representation of higher-order models by combining several lower order models [11]. Thus an nth order model over a random variable *S* (with *k* possible values) can be expressed as:

$$P(w_i \mid w_{i-1},..., w_{i-n}) = \sum_{\mu=1}^{n} \phi(\mu) a^{\mu}(w_i \mid w_{i-\mu}) \qquad (2)$$

where $a^{\mu}(w_i|w_{i-\mu})$ is a $k\times k$ transition matrix containing the probabilities of the occurrence of a symbol at position *i* given the occurrence of a symbol at position *i*-μ.

The mixing coefficients $\phi(\mu)$ are estimated using an iterative procedure, using the initial counts in the transition. Due to space restrictions we omit here the description of this procedure. For a detailed explanation of MMMs and parameter estimation methods the reader is referred to [11,12].

The MMM is trained with all pairwise dependencies found in the feature sequences generated from the training set and then the corresponding mixing coefficients are estimated.

2.3. Entropy and Boundary Prediction

Following our initial assumption, we propose that some segmentation boundaries are likely to occur close to accentuated changes in entropy, associated with some melodic features.

Shannon [13] showed that one of the ways of measuring the quantity of information of a particular message is to determine its unpredictability or entropy. We can determine the entropy associated with a given context *c* as,

$$H_c = \sum_{\forall w} P(w \mid c) \log_2 P(w \mid c) \qquad (3)$$

where *w* denotes all symbols that can be successors of the context *c*. Context *c* is a sequence of size *n-1*, where *n* is the order of the model. Conditional probabilities are obtained from Equation 2 and will reflect the statistics of the training set.

Entropy vectors are then calculated by taking all the successive context sequences from the feature vectors of the target melody. As mentioned earlier we are interested only in more prominent entropy changes across the melody. For every entropy vector we first determine the mean and standard deviation. Then all values outside the standard deviation are filtered from the vectors. Finally, from the remaining values in the vector, we considered only those that register a contiguous low-high or high-low variation with respect to the mean.

3. RESULTS

The experimental part of this work has two components. The first is an empirical study on melodic segmentation carried out with listeners on some melodies. The aim of this study was to collect segmentation information from a real listening experience and to provide comparison data for our computational model of melody segmentation. In the second part we used our computational model with some of the examples provided to the listeners, to predict the locations of the segment boundaries.

3.1. A Listening Study

A total of 48 subjects took part in this listening study. Participants were all 3rd/4th-year undergraduate or postgraduate students, split between musically trained and non-musically trained subjects.

ISBN 3-931852-67-9
ISSN 1617-6847

The set of melodies used in this study included 3 folk songs from the Essen Folk Song Collection (initiated by Prof. H. Schaffrath), 2 melody excerpts from Mozart Piano Sonatas and Debussy's Syrinx. All melodies were provided as deadpan MIDI files, with the exception of Syrinx, which was obtained from an expressive performance (performed by Peter-Jan van Dijk), thus including ornaments, tempo fluctuations and dynamics.

For each melody subjects had two familiarisation auditions, a trial segmentation audition and a final segmentation audition. Data collection was performed by a computer program designed to guide the listeners through the whole listening session. Listeners were able to indicate a segment boundary by pressing the mouse button, while the melody was being played. To minimise the effects of priming, the program also guaranteed that no two listeners heard the melodies in the same order.

The segment locations (time stamps) collected from the listeners were later synchronised with the MIDI data to associate them with the events in the melody. A boundary is matched with an event if it occurs between the onset times of that event and the next. In Figure 1 we show the histogram of segment boundary counts per event for Syrinx, for all subjects. The analysis of the separate boundary histograms for musician and non-musician subjects indicates that the differences between the two are not significant. This relatively low influence of the factor musical training in a segmentation task has previously been reported in [14].

Observing the graph of Figure 1 it is clear that listeners agreed on several segmentation locations, within the melody. The graph, also suggests that there is a delay or anticipation in some of the subjects' responses, particularly visible around the main segment boundaries at events 80, 112, 139 and 252. There are also areas in the graph that show a considerable number of responses that span over a fairly large number of consecutive events (e.g. 47-50 or 303-306).

3.3. Automatic Segmentation

We now look at the results obtained with our segmentation model on Syrinx, the larger melody of the study set. This melody was used both as the training set and the test set.

In Figure 2 we plot the outstanding entropy transitions (white markings) for PS and DR, overlapped with the boundaries indicated by the listeners (we will refer to the latter as *L-boundaries*). In some cases the entropy variations stretch across more than two consecutive events and these, similarly to the *L-boundaries*, are depicted as several overlapped markings.

We considered that a prediction is correct if it indicates an existing *L-boundary* location within a distance of ±1 event. From a total of 14 *L-boundaries* considered, 11 were predicted correctly by the model (5 from *H(PS)* and 6 from *H(DR)*). The model generated also 5 excessive boundaries, 3 from *H(PS)* and 2 from *H(DR)*. Excessive boundaries are those that have no correspondence with any of the *L-boundaries*.

3.4. Discussion

After analysing the nature of the boundaries predicted by the model we observe that most of them can be explained by Gestalt-based principles of proximity both in the pitch and time domains. This idea is reinforced by the fact that most *L-boundaries* coincide with the location of breath marks (see Figure 2), and these often follow longer notes or large pitch intervals marking the end of phrases. Nevertheless, it is a fact that no rules were previously provided to the model, so they have in fact been derived from the data and reflected on the results.

Although we generated models of up to order 6, the weighting coefficients of Equation 2 show that the MMM approximation for this particular melodic data is equivalent to a model of order between 2 and 3. This means that patterns acquired by the model involve at most 3 to 4 events. This imposes a limit on the discovery of pattern similarities. Nevertheless, boundaries at 15, 252 and 268 were predicted based on the rhythmic re-ocurrence of the opening motif of this piece (represented in Table 1). Boundaries 15 and 252 were two of the most voted by listeners. Boundary 268 also marks the beginning of the same rhythmic motif, but was not selected by the listeners. This suggests that although a low order model cannot store large patterns, smaller partial patterns can be retained as indexes of longer ones. Some theories have argued for the prototypical nature of parallelism and have shown that patterns are often remembered by the repetition of smaller cells, often their initial section [16].

The length of Syrinx seems to provide enough redundant information to highlight most of its recurrent features, but not enough to prevent the model from being fairly sensitive to the less frequent ones.

For example, pitch intervals of 1, 2 or 3 semitones are very frequent throughout the whole melody. In comparison most other intervals will seem very improbable, and thus will be responsible for large variations in entropy.

The use of relative measurements for the melodic features used to train the model, increased the redundancy of the data, and to some extent can be seen as a form of representing approximate similarity. It is remarkable that the majority of the *L-boundaries* could be predicted only with the information contained in this one piece. However, for very short melodies this would not be possible due to the lack of redundant information. In the following stage of this research we plan to train the model with a set of melodies and then use a target melody not included in the training set, but that is somehow represented by the training set. More specifically, we plan to use a subset of songs from the Essen Database as our training set. Then use will take the melodies used in the listening study as our test set, and re-evaluate the ability of the model to predict the boundary locations.

5. CONCLUSIONS

We presented a memory-based model of music learning and melodic segmentation. The model relates feature salience with expectation and uses entropy measurements to evaluate the occurrence of pitch and time-based melodic features.

We presented some experimental results that seem to corroborate the idea that outstanding variations in entropy constitute salient moments in a listening experience. The results so far suggest that intra-opus information can greatly influence the perception of segmentation boundaries. It was found that most boundaries predicted by the model could be explained with Gestalt-based principles, but these principles were captured from non-annotated melodic data and reproduced in the segmentation predictions.

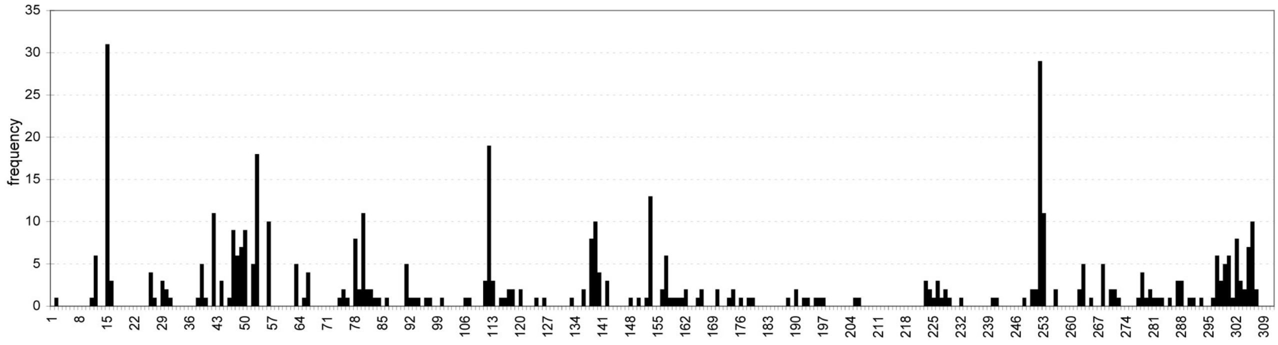

Figure 1 – Histogram of segment boundaries (for all subjects) for Syrinx. Bins correspond to Midi events.

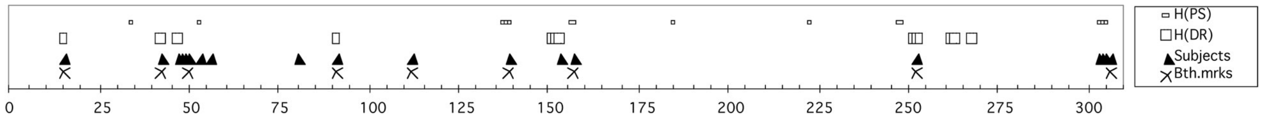

Figure 2 – Segment boundary locations indicated by listeners, notated breath marks and boundaries predicted by the model.

ACKNOWLEDGMENTS

This research was funded by EPSRC research project GR/N08049/01. Thanks to Taylan Cemgil for his advice on Markov modelling and to Marcus Pierce for helpful discussions on the listening study.

5. REFERENCES

1. Lerdahl, F., Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. M.I.T. Press, Cambridge (Mass.).

2. Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. PhD thesis, University of Edinburgh.

3. Deliège, I., Melén, M. (1997). Cue abstraction in the representation of musical form. In Deliège, I., Sloboda, J. (eds.) *Perception and Cognition of Music*. Psychology Press (pp. 387-412).

4. Meyer, L.B. (1967). *Music, The Arts, And Ideas - Patterns and Predictions in Twentieth-Century Culture*. University of Chicago Press, Chicago.

5. Huron, D. (2001). What is a musical feature? Forte's analysis of Brahms's opus 51, no. 1, revisited. *Music Theory On-line* **7**.

6. Manning, C.D., Schüttze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.

7. Conklin, D., Witten, I. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research* **24** (pp. 51-73).

8. Ponsford, D., Wiggins, G., Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research* **28**

9. Reis, Y.B. (1999). 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* (pp. 58-63).

10. Bod, R. (2001). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research* **30**.

11. Saul, Lawrence, Jordan, Michael (1999). Mixed Memory Markov Models: Decomposing Complex Stochastic Processes as Mixtures of Simpler Ones. *Machine Learning*. **30** (1). (pp. 75-87).

12. Ney, Hermann and Essen, Ute and Kneser, Reinhard (1994). Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer Speech and Language*. **8** (1). (pp. 1-38).

13. Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27** (pp. 379-423,623-656).

14. Deliège, I. (1998). Wagner "alte weise": Une approche perceptive. *Musica Scientiæ* **Special Issue** (pp. 63-90)

15. Deliège, I. (2001). Prototype effects in music listening: An empirical approach to the notion of imprint. *Music Perception* **18** (pp. 371-407).