A COMPARISON OF STATISTICAL AND RULE-BASED MODELS OF MELODIC SEGMENTATION

M. T. Pearce, D. Müllensiefen and G. A. Wiggins
Centre for Computation, Cognition and Culture
Goldsmiths, University of London
{m.pearce,d.mullensiefen,g.wiggins}@gold.ac.uk

ABSTRACT

We introduce a new model for melodic segmentation based on information-dynamic analysis of melodic structure. The performance of the model is compared to several existing algorithms in predicting the annotated phrase boundaries in a large corpus of folk music.

1 INTRODUCTION

The segmentation of melodies into phrases is a fundamental (pre-)processing step for many MIR applications including melodic feature computation, melody indexing, and retrieval of melodic excerpts. In fact, the melodic phrase is often considered one of the most important basic units of musical content [16] and many large electronic corpora of music are structured or organised by phrases, for example, the Dictionary of Musical Themes by Barlow and Morgenstern [2], the Essen Folksong Collection (EFSC) [33] or the RISM collection [28].

At the same time, melodic grouping is thought to be an important part of the perceptual processing of music [11, 14, 27]. 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 paper, we introduce a new statistical model of melodic segmentation and compare 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

2.1 Evaluation Measures

In modern information retrieval, *Precision*, *Recall*, and *F1* have become standard measures for assessing model performance. These measures are usually defined in terms of *True Positives, TP* (i.e. the number of times a model predicts a specific outcome correctly), *False Positives, FP* (i.e. the number of times a model predicts a specific outcome incorrectly), and *False Negatives, FN* (i.e. the number of times a model incorrectly does not predict a specific outcome):

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

2.2 Models of Melodic Segmentation

GTTM: 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 [6, 16, 36]. Perhaps the best known examples are the Grouping Preference Rules (GPRs) of the Generative Theory of Tonal Music (GTTM) [16]. 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. Some of these GPRs have been quantified [14] and studied in psychological experiments [11, 14].

LBDM: Cambouropoulos [6] proposes a related model in which boundaries are associated with any local change in interval magnitudes. The *Local Boundary Detection Model* (LBDM) consists of a *change* rule, which assigns boundary strengths in proportion to the degree of change between consecutive intervals, and a *proximity* rule, which scales the boundary strength according to the size of the intervals involved. The LBDM operates over several independent parametric melodic profiles $P_k = [x_1, x_2, \ldots, x_n]$ where $k \in \{\text{pitch, ioi, rest}\}$, $x_i > 0, i \in \{1, 2, \ldots, n\}$ and the boundary strength at interval x_i is given by:

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

where the degree of change between two successive intervals:

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

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

Grouper: Temperley [36] introduces a model called *Grouper* which accepts a melody, in which each note is represented by its onset time, off time, chromatic pitch and level in a metrical hierarchy, and returns a single, exhaustive partitioning of the melody into 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 $|(\log_2 N) - 3|$ where N is the number of notes in the predicted phrase;

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

The first rule is another example of the Gestalt principle of temporal proximity (cf. GPR 2 above); 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. By way of evaluation, Temperley used Grouper to predict the phrase boundaries marked in 65 melodies from the EFSC achieving a recall of 0.76 and a precision 0.74.

Other Models: Tenney and Polansky [37] were perhaps the first to propose models of melodic segmentation based on Gestalt-like rules. Other authors have combined Gestalt-like rules with higher-level principles based on parallelism and music structure [1, 7]. Ferrand et al. [13] introduce an approach based on the idea of 'melodic density' (i.e., segment at points of low cohesion between notes) and compare the methods performance to the LBDM. In contrast, Bod [3] argues for a supervised learning approach to modelling melodic grouping structure. A model based on data-oriented parsing (DOP) yielded $F1 = 0.81$ in predicting unseen phrase boundaries in the EFSC. A qualitative examination of the data revealed that 15% of the phrase boundaries predicted by the Markov-DOP parser cannot be accounted for by Gestalt principles. These models are mentioned for completeness, but are not included in our comparison.

2.3 The IDyOM Model

We present a new model of melodic grouping (the Information Dynamics Of Music model) that is inspired by previous research in musicology, music perception, computational linguistics and machine learning.

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 [21, 22]. These proposals may be given precise definitions in an information-theoretic framework which we define by reference to a model of unsupervised inductive learning of melodic structure. Briefly, the models we propose output conditional probabilities of an event e, given a preceding sequential context c. 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*, $h(e|c)$, of the event given the context:

$$h(e|c) = \log_2 \frac{1}{p(e|c)}.$$

The information content can be interpreted as the contextual unexpectedness or surprisal associated with an event. Given an alphabet $\mathcal{E}$ of events which have appeared in the prior experience of the model, the uncertainty of the model's expectations in a given melodic context can be defined as the *entropy* or average information content of the events in $\mathcal{E}$:

$$H(c) = \sum_{e \in \mathcal{E}} p(e|c) h(e|c).$$

We propose that boundaries will occur before events for which unexpectedness (h) and uncertainty (H) are high.

In addition to the musicological basis, there is a precedent for these ideas in experimental psychology. Empirical research has demonstrated that infants and adults use the implicitly learnt statistical properties of pitch [32], pitch interval [30] and scale degree [29] sequences to identify segment boundaries on the basis of higher digram ($n = 2$) transition probabilities within than between groups.

There is also 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 [17] and the induced changes in entropy of grammatical continuations [15]. More specifically, experimental work has demonstrated that infants and adults reliably identify grouping boundaries in sequences of synthetic syllables [31] on the basis of higher transition probabilities within than between groups.

Furthermore, research in machine learning and computational linguistics has demonstrated that algorithms that segment before unexpected events can successfully identify word boundaries in infant-directed speech [4]. Similar strategies for identifying word boundaries have been implemented using recurrent neural networks [12]. Recently, Cohen et al. [8] proposed a general method for segmenting sequences based on two principles: first, so as to maximise *n*-gram frequencies to the left and right of the boundary; and second, so as to maximise the entropy of the conditional distribution across the boundary. The algorithm was able to successfully identify word boundaries in text from four languages and episode boundaries in the activities of a mobile robot.

IDyOM itself is based on n-gram models commonly used in statistical language modelling [18]. An n-gram is a sequence of n symbols and an n-gram model is simply a collection of such sequences each of which is associated with a frequency count. 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 a 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. The quantity $n - 1$ is known as the *order* of the model and represents the number of symbols making up the context within which a prediction is made.

However, n-gram models suffer from several problems, both in general and specifically when applied to music. The first difficulties arise from the use of a fixed-order. Low-order models fail to provide an adequate account of the structural influence of the context. However, 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). In order to address these problems, the IDyOM model maintains frequency counts during training for n-grams of all possible values of n in any given context. During prediction, distributions are estimated using a weighted sum of all models below a variable order bound. This bound is determined in each predictive context using simple heuristics designed to minimise uncertainty. 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).

Another problem with n-gram models is that a trained model will fail to make use of local statistical structure of the music it is currently analysing. To address this problem, IDyOM includes two kinds of model: first, the *long-term* model that was trained over the entire training set in the previous step; and second, a *short-term* model that is trained incrementally for each individual melody being predicted. The distributions returned by these models are combined using an entropy weighted multiplicative combination scheme [26] 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 attributes including pitch, onset time, duration, timbre and so on. In addition, sequences of these attributes may have multiple relevant dimensions. For example, pitch interval, pitch class, scale degree, contour and many other derived features are important in the perception and analysis of pitch structure. In order to accommodate these properties, the modelling process begins by choosing a set of basic features that we are interested in predicting. As these basic features are treated as independent attributes, their probabilities are computed separately and in turn, 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.

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 additive combination scheme all improve the performance of IDyOM in predicting the pitches of unseen melodies [24, 26]. Full details of the model and its evaluation can be found elsewhere [9, 23, 24, 26].

The conditional probabilities output by IDyOM in a given melodic context may be interpreted as contextual expectations about the nature of the forthcoming note. Pearce and Wiggins [25] compare the melodic pitch expectations of the model with those of listeners in the context of single intervals [10], at particular points in British folk songs [34] and throughout two chorale melodies [19]. The results demonstrate that the statistical system predicts the expectations of listeners as least as well as the two-factor model of Schellenberg [35] and significantly better in the case of more complex melodic contexts.

In this work, 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 don't use any 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 below). The role of entropy (H) will be considered in future work. The IDyOM model differs from the GPRs, the LBDM and Grouper in that it is based on statistical learning rather than symbolic rules and it differs from DOP in that it uses unsupervised rather than supervised learning.

2.4 Comparative evaluation of melody segmentation algorithms

Most of the models described above were evaluated to some extent by their authors and, in some cases, compared quantitatively to other models. In addition, however, there exist a small number of studies that empirically compare the performance of different models of melodic grouping. These studies differ in the algorithms compared, the type of ground truth data used, and the evaluation metrics applied. Melucci and Orio [20], for example, collected the boundary indications of 17 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 baseline models that created groups containing fixed (8 and 15) or random (between 10 and 20) numbers of notes. Melucci and Orio report false positives, false negatives, and a measure of disagreement which show that the LBDM outperforms the other models.

Bruderer [5] presents a more comprehensive 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 examines the performance of three algorithms: Grouper, LBDM and the summed GPRs quantified in [14] (GPR 2a, 2b, 3a and 3d). The output of each algorithm is convolved with a Gaussian window to produce a boundary strength profile that is then correlated with the ground truth. Bruderer reports that the LBDM achieved the best and the GPRs the worst performance.

Another study [38] compared the predictions of the LBDM and Grouper to segmentations at the phrase and subphrase level provided by 19 musical experts for 10 melodies in a range of styles. The performance of each model on each melody was estimated by averaging the F1 scores over the 19 experts. Each model was examined with parameters optimised for each individual melody. The results indicated that Grouper tended to outperform the LBDM. Large IOIs were an important factor in the success of both models. In another experiment, the predictions of each model were compared with the transcribed boundaries in several datasets from the EFSC. The model parameters were optimised over each dataset and the results indicated that Grouper (with mean F1 between 0.6 and 0.7) outperformed the LBDM (mean F1 between 0.49 and 0.56).

All 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 [5] to 20 by [20]). Apart from Thom et al. [38], Experiment D, there has been no thorough comparative evaluation over a large corpus of melodies annotated with phrase boundaries.

3 METHOD

3.1 The Ground Truth Data

We concentrate here on the results obtained for 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 sounding events at an average of about 46 events per melody and overall about 12% of notes fall before boundaries. There is only one hierarchical level of phrasing and the phrase structure exhaustively subsumes all the events in a melody.

3.2 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 [5] we chose to make all segmentation algorithms comparable by picking binary boundary indications from the boundary strength profiles.

To do so, we devised a method called *Simple Picker* that uses three principles. First, 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: $S_n > S_{n-1}$. Whilst these principles, simply ensure that a point is a local peak in the profile, the third specifies how high the point must be, relative to earlier points in the profile, to be considered a peak. Thus the note following a boundary should have a boundary strength greater than a threshold based on the linearly weighted mean and standard deviation of all notes preceding it:

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

The third principle makes use of the parameter k which determines how many standard deviations higher than the mean of the preceding values a peak must be to be picked. 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.

3.3 The Models

The models included in the comparison are as follows:

Grouper: as implemented by [36]; [1]

LBDM: as specified by [6] with $k = 0.5$;

IDyOM: with $k = 2$;

GPR2a: as quantified by [14] with $k = 0.5$;

GPR2b: as quantified by [14] with $k = 0.5$;

GPR3a: as quantified by [14] with $k = 0.5$;

GPR3d: as quantified by [14] with $k = 2.5$;

Always: every note falls on a boundary;

Never: no note falls on a boundary.

Grouper outputs binary boundary predictions but the output of every other model was processed by Simple Picker using a value of k was chosen from the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$ so as to maximise F1 (and secondarily Recall).

4 RESULTS

The results of the model comparison are shown in Table 1. The four models achieving mean F1 values of over 0.5

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

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
GPR2a	0.99	0.45	0.58
IDyOM	0.76	0.50	0.58
GPR2b	0.47	0.42	0.39
GPR3a	0.29	0.46	0.35
GPR3d	0.66	0.22	0.31
Always	0.13	1.00	0.22
Never	0.00	0.00	0.00

Table 1. The model comparison results in order of mean F1 scores.

(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 improvements 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. The performance of the resulting model is shown in the top row of Table 1. Sign tests demonstrated that the Hybrid model achieved better F1 scores on significantly more melodies than each of the other models.

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 rule, GPR2a. This points to the conclusion that rests, perhaps above all other melodic parameters, have a large influence on boundaries in this melodic style. 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.

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. [38]. The performance of Grouper is somewhat lower than that reported by Temperley [36] on 65 melodies from the EFSC. The performance of all models is lower than that of the supervised learning model reported by Bod [3].

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 in the logistic regression procedure. Another possibility is to use boosting 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, in comparison to sophisticated rule-based models. 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.

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 segmentation performance by providing the statistical model with more sophisticated melodic representations and examining the role of entropy (uncertainty) in melodic boundary detection.

6 REFERENCES

[1] S. Ahlbäck. *Melody beyond notes: A study of melody cognition*. PhD thesis, Göteborg University, Göteborg, Sweden, 2004.

[2] H. Barlow and S. Morgenstern. *A dictionary of musical themes*. Ernest Benn, 1949.

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

[4] Michael R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105, 1999.

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

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

[7] E. Cambouropoulos. Musical parallelism and melodic

segmentation: A computational approach. *Music Perception*, 23(3):249–269, 2006.

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

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

[10] L. L. Cuddy and C. A. Lunny. Expectancies generated by melodic intervals: Perceptual judgements of continuity. *Perception and Psychophysics*, 57(4):451–462, 1995.

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

[12] J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[13] M. Ferrand, P. Nelson, and G. Wiggins. Memory and melodic density: a model for melody segmentation. In N. Giomi F. Bernardini and N. Giosmin, editors, *Proceedings of the XIV Colloquium on Musical Informatics*, pages 95–98, Firenze, Italy, 2003.

[14] B. W. Frankland and A. J. Cohen. 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, 2004.

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

[16] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.

[17] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 16(3):1126–1177, 2008.

[18] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[19] L. C. Manzara, I. H. Witten, and M. James. On the entropy of music: An experiment with Bach chorale melodies. *Leonardo*, 2(1):81–88, 1992.

[20] M. Melucci and N. Orio. A comparison of manual and automatic melody segmentation. In *Proceedings of the International Conference on Music Information Retrieval*, pages 7–14, 2002.

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

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

[23] M. T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, UK, 2005.

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

[25] M. T. Pearce and G. A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006.

[26] M. T. Pearce, D. Conklin, and G. A. Wiggins. Methods for combining statistical models of music. In U. K. Wiil, editor, *Computer Music Modelling and Retrieval*, pages 295–312. Springer Verlag, Heidelberg, Germany, 2005.

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

[28] RISM-ZENTRALREDAKTION. Répertoire international des scources musicales (rism). URL `http://rism.stub.uni-frankfurt.de/index\_e.htm`.

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

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

[31] J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month old infants. *Science*, 274: 1926–1928, 1996.

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

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

[34] E. G. Schellenberg. Expectancy in melody: Tests of the implication-realisation model. *Cognition*, 58(1): 75–125, 1996.

[35] E. G. Schellenberg. Simplifying the implication-realisation model of melodic expectancy. *Music Perception*, 14(3):295–318, 1997.

[36] D. Temperley. *The Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.

[37] J. Tenney and L. Polansky. Temporal Gestalt perception in music. *Contemporary Music Review*, 24(2): 205–241, 1980.

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