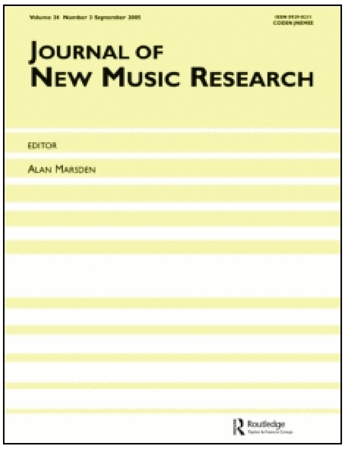


Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713817838

# A Unified Probabilistic Model for Polyphonic Music Analysis

David Temperley[a]
[a] Eastman School of Music, University of Rochester, USA

**To cite this Article** Temperley, David(2009) 'A Unified Probabilistic Model for Polyphonic Music Analysis', Journal of New Music Research, 38: 1, 3 — 18
**To link to this Article: DOI:** 10.1080/09298210902928495
**URL:** http://dx.doi.org/10.1080/09298210902928495

Journal of New Music Research
2009, Vol. 38, No. 1, pp. 3–18

Routledge
Taylor & Francis Group

# A Unified Probabilistic Model for Polyphonic Music Analysis

David Temperley

Eastman School of Music, University of Rochester, USA

## Abstract

This article presents a probabilistic model of polyphonic music analysis. Taking a note pattern as input, the model combines three aspects of symbolic music analysis—metrical analysis, harmonic analysis, and stream segregation—into a single process, allowing it to capture the complex interactions between these structures. The model also yields an estimate of the probability of the note pattern itself; this has implications for the modelling of music transcription. I begin by describing the generative process that is assumed and the analytical process that is used to infer metrical, harmonic, and stream structures from a note pattern. I then present some tests of the model on metrical analysis and harmonic analysis, and discuss ongoing work to integrate the model into a transcription system.

## 1. Introduction

In the last decade, the field of computational music research has seen an explosion of work using probabilistic methods. This work includes models of meter induction (Cemgil et al., 2000a,b; Raphael, 2002; Cemgil & Kappen, 2003), key induction and harmonic analysis (Raphael & Stoddard, 2004; Temperley, 2004), voice separation (Kirlin & Utgoff, 2005), style classification (Chai & Vercoe, 2001; de la Higuera et al., 2005), expectation (Pearce & Wiggins, 2006), and transcription (Kashino et al., 1998; Cemgil et al., 2005; Davy, 2006). These studies have explored a variety of ways of applying probabilistic techniques to musical problems and have yielded impressive achievements. For the most part, however, this research has been highly compartmentalized—focusing on individual problems such as meter induction, harmonic analysis, and transcription in isolation without addressing the connections between these problems. One of the great virtues of the probabilistic approach is that it provides a framework for integrating multiple interacting processes and representations into a single unified model. This not only has the obvious benefits of parsimony and generality—solving several problems at once—but also holds out the promise of improving performance on each individual problem, in relation to what can be achieved by addressing them separately.

In this article I present a unified probabilistic model of polyphonic music analysis. The model is unified in two senses. First of all, it integrates three aspects of symbolic music analysis—meter analysis, harmonic analysis, and stream segregation—that in previous work have only been addressed individually. Secondly, at a higher level, the model unifies the general problem of structural analysis with the problem of estimating the probabilities of note patterns—a problem that, in turn, has implications for the modelling of music transcription. The model builds on earlier work: in Temperley (2007), I presented a model which analyses key and meter in monophonic input and estimates the probabilities of monophonic note patterns. The current model extends this previous research by accommodating polyphonic input, incorporating harmonic and stream analysis, and providing a workable component for a transcription system. The model is intended primarily for traditional Western art music ('classical' music), but may be applicable to other styles as well.[1]

[1] The source code for the implementation of the model, which is written in C, can be downloaded at www.theory.esm.rochester.edu/temperley/melisma2.

*Correspondence*: David Temperley, Eastman School of Music, University of Rochester, 26 Gibbs St., Rochester, NY 14604, USA. E-mail: dtemperley@esm.rochester.edu

DOI: 10.1080/09298210902928495 © 2009 Taylor & Francis

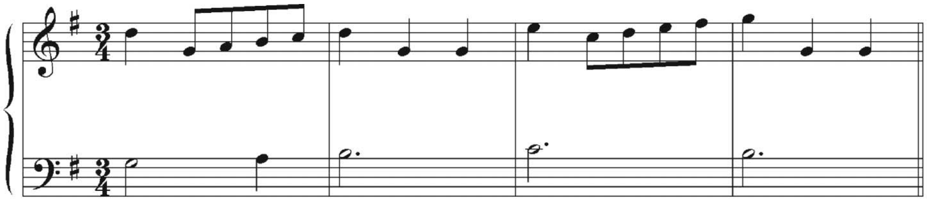
The input to the model is a MIDI or 'piano-roll' representation—a list of notes indicating the on-time and off-time (in milliseconds) and pitch of each note. The model then derives three kinds of musical structure: metrical structure, harmonic structure, and stream structure. Following a widely used convention (Lerdahl & Jackendoff, 1983), I define a metrical structure as a framework of levels of beats, as shown at the left of Figure 1. Generally (at least in Western music), every second or third beat at one level is retained at the next level up; beats at each level tend to be roughly evenly spaced, but not exactly so, at least in human performance. A harmonic structure is a segmentation of a piece into time-spans labelled with chords; for our purposes the labels are simply roots, though they could also carry more specific information, e.g. chord quality (major versus minor) or relationship to the key (e.g. 'I of C major'). Finally, a stream structure is a grouping of the notes of a polyphonic texture into melodic lines (also called streams or voices). Following previous research on stream separation (Temperley, 2001; Kirlin & Utgoff, 2005; Cambouropoulos, 2008), I assume that the number of active streams within a piece may fluctuate; thus streams are allowed to begin and end within the piece. Figure 1 shows the opening of a Bach minuet; below the music notation, the figure shows the piano-roll representation used as input as well as the metrical structure, harmonic structure, and stream structure.

Before proceeding, it may be helpful to elaborate on a point made in the first paragraph—that addressing problems of musical information-processing in a unified fashion can yield better results than addressing them individually. Prior research on metrical, harmonic, and stream analysis has shown that these three problems are very intimately related. Regarding the interaction of meter and harmony, it can be seen from almost any harmonic analysis that changes of harmony tend to occur on relatively strong beats: generally at strong tactus beats, and very rarely below the level of the tactus. (The tactus is an intermediate level in the metrical hierarchy, usually corresponding to what is informally called the 'beat'—most often the quarter-note.) Table 1 shows some evidence on this point, gathered from the Kostka–Payne corpus, a harmonically-annotated corpus of classical excerpts which is discussed further below. If we define the tactus level as level 2, the level immediately above as level 3, and the level below as level 1, changes of harmony occur on about 71% of level 3 beats, 22% of level 2 beats (which are not level 3 beats), and only 2% of level 1 beats. Several models of harmonic analysis, such as those of Maxwell (1992) and Temperley (2001), have explicitly included metrical information in the input. [Others, such as Raphael and Stoddard (2004), have finessed the problem by limiting the possible points of harmonic change to a high metrical level such as bars or half-bars.] But the influence also goes the other way: harmony also affects meter. In Temperley (2001), in discussing the metrical analysis model presented there, I noted that many of the model's errors were due to its ignorance of harmonic structure, specifically the fact (noted above) that harmonic changes rarely occur on very weak beats. From a probabilistic viewpoint, this two-way interaction makes sense: if a strong beat indicates a high probability of a harmonic change, it is hardly surprising that the clear presence of a harmonic change would suggest a strong beat.

The interaction of stream structure with meter and harmony is perhaps less obvious, but nonetheless important. An example is seen in the effect of note length on metrical analysis. It is well known that there is a tendency for long notes to coincide with strong beats; most metrical analysis models define the length of a note as its 'inter-onset-interval' (IOI), the interval between the note's onset and the onset of the following note (see e.g. Povel & Essens, 1985; Rosenthal, 1992). But as noted in Temperley (2001), determining the IOI of a note in polyphonic music is a non-trivial task. In Figure 1, the IOI of the left-hand note of bar 2 is just one quarter-note, as the next note occurs on the second beat of the bar (the G4 in the right-hand); but the perceived length of the note is actually one bar. Intuitively, what seems to matter is the IOI of a note in relation to the next note *within the same voice*; obviously, this assumes knowledge of stream structure. Similarly, a well-known principle of harmony is that non-chord-tones (notes not part of the current chord) tend to resolve by step; but this generally implies resolution to another note within the same voice, which again requires the grouping of notes into voices.

Elsewhere (Temperley, 2007) I have proposed a distinction between 'structural' processes of music cognition—those that involve inferring higher-level structures from a pattern of notes, such as metrical, harmonic, and stream structures—and 'surface' processes, those that involve the identification and projection of the note pattern itself. Transcription—the extraction of a polyphonic note pattern from an audio signal—is a complex and difficult problem that has received much attention in recent years [see Klapuri and Davy (2006) for a survey of work in this area]. Other important surface processes include expectation, the projection of future notes based on a prior context (Schellenberg, 1997; Jones et al., 2002), and error-detection, the identification and correction of erroneous notes. For the most part, these surface processes have been addressed in isolation from structural problems. But again, structural and surface problems are very intimately related. Here, a Bayesian probabilistic framework is particularly helpful. In Bayesian terms, the problem of transcription is one of identifying the most probable note pattern $N$ given the signal. According to Bayes' rule,

$$P(N|Signal) \propto P(Signal|N)P(N). \qquad (1)$$



```
    0        G     x x x x . . . . . . . . . . . . .1. . . . . . . . . 2 . . . . . . .
  100        G     x             .           .      |    .           . |         .
  200        G     x x           .           .      |    .           . |         .
  300        G     x             .           .      |    .           . |         .
  450        G     x x x   . . . . . . . . . . . . .|. . . . . .2. . . . . . . . . . .
  550        G     x             .           .      |    .      |    .           .
  650        G     x x           .           .      |    .        2  .           .
  750        G     x             .           .      |    .        |  .           .
  850        G     x x x   . . . . . . . . . . . . . .1. . . . . . .2. . . . . . . . .
  950        G     x             .           .        |  .          |.           .
 1100        G     x x           .           .        |  .           2           .
 1200        G     x             .           .        |  .           |           .
 1300        G     x x x x . . . . . . . . . . . . . . .1. . . . . . . 2 . . . . . . .
 1400        G     x             .           .          |.           . |         .
 1500        G     x x           .           .          |.           . |         .
 1600        G     x             .           .          |.           . |         .
 1750        G     x x x   . . . . . . . . . . . . . . .|. . . .2. . . . . . . . . . .
 1850        G     x             .           .          |.      |    .           .
 1950        G     x x           .           .          |.      |    .           .
 2050        G     x             .           .          |.      |    .           .
 2150        G     x x x   . . . . . . . . . . . . . . .|. . . .2. . . . . . . . . . .
 2250        G     x             .           .          |.      |    .           .
 2350        G     x x           .           .          |.      |    .           .
 2450        G     x             .           .          |.      |    .           .
 2600       C      x x x x . . . . . . . . . . . . . . . 1 . . . . . . . 2 . . . . . .
 2700       C      x             .           .           |           .   |       .
 2800       C      x x           .           .           |           .   |       .
 2900       C      x             .           .           |           .   |       .
 3050       C      x x x   . . . . . . . . . . . . . . . | . . . . . 2 . . . . . . . .
 3150       C      x             .           .           |           |           .
 3250       C      x x           .           .           |           . 2         .
 3350       C      x             .           .           |           . |         .
 3500       C      x x x   . . . . . . . . . . . . . . . | . . . . . . . 2 . . . . . .
 3600       C      x             .           .           |           .   |       .
 3700       C      x x           .           .           |           .     2     .
 3800       C      x             .           .           |           .     |     .
 3900        G     x x x x . . . . . . . . . . . . . . .1. . . . . . . . . .2. . . . .
 4000        G     x             .           .          |.           .      |    .
 4100        G     x x           .           .          |.           .      |    .
 4200        G     x             .           .          |.           .      |    .
 4350        G     x x x   . . . . . . . . . . . . . . .|. . . .2. . . . . . . . . . .
 4450        G     x             .           .          |.      |    .           .
 4550        G     x x           .           .          |.      |    .           .
 4650        G     x             .           .          |.      |    .           .
 4800        G     x x x   . . . . . . . . . . . . . . .|. . . .2. . . . . . . . . . .
 4900        G     x             .           .          |.      |    .           .
 5000        G     x x           .           .          |.      |    .           .
 5100        G     x             .           .          |.      |    .           .
 5250        G     x x x x . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

Fig. 1. Bach, Minuet from the *Notebook for Anna Magdalena Bach*, mm. 1–4. Below the score is the output of the polyphonic analysis program. At left is the timepoint (in milliseconds) at which each lowest-level metrical segment begins. To the right of that is the harmonic structure (showing the root of each segment), the metrical grid (showing four metrical levels), and a 'piano-roll' representation of the input, with notes grouped into two streams (labelled 1 and 2).

We can maximize the expression on the left by maximizing the expression on the right; this involves identifying the probability of the signal given a note pattern, and also the probability of the note pattern itself. With regard to the latter term, in evaluating the probability of a note pattern, it stands to reason that we bring to bear musical structures of the kind discussed earlier. Roughly speaking, a probable note pattern is one that implies and adheres to a clear metrical structure, harmonic structure, and stream structure (among other things perhaps). As argued in Temperley (2007), a Bayesian transcription system really requires that any structural representations that affect $P(N)$ be integrated into a single generative model. That is to say: it makes no sense to posit separate rhythmic, harmonic, and stream segregation models that all assign probabilities to note patterns; these structures must be combined in some way to generate note patterns, and any workable model must specify how this is done.

While some studies of transcription have recognized the value of applying higher-level musical knowledge to the problem (Kashino et al., 1998; Klapuri, 2004), little concrete progress has been made in this direction. The system that I present below, as well as inferring metrical/harmonic/stream structures from a given note pattern, also calculates the probabilities of note patterns and thus could naturally be integrated into a transcription system; some efforts in this direction are underway, as I discuss at the end of the article.

Like most Bayesian probabilistic models, the current model assumes a generative process—in this case, a process that generates musical structures and then generates note patterns from those structures. For the analytical process, the usual Bayesian reasoning is then used. Assume a given note pattern $N$; then for any metrical structure $M$, harmonic structure $H$, and stream structure $S$:

$$\begin{aligned} P(M,H,S|N) &\propto P(N|M,H,S)P(M,H,S) \\ &= P(M,H,S,N). \end{aligned} \tag{2}$$

The aim of the analytical process is to maximize the expression on the left, which we do by maximizing the expression on the right.

For transcription and other surface processes, we need to know the probability of the note-pattern itself. Here again, we use the usual probabilistic reasoning, representing $P(N)$ as the sum of the joint probability of $N$ with all structures:

$$P(N) = \sum_{M,H,S} P(M,H,S,N). \tag{3}$$

We begin by describing the generative process. We then turn to the analytical process, and describe some tests of the model's analytical ability. Finally we consider the transcription problem and discuss some further issues.

Table 1. Harmonic changes at beats of different metrical levels in the Kostka–Payne corpus.

| Metrical level (2 = tactus) | % of beats with changes of harmony |
|---|---|
| 3 | 71.5 |
| 2 | 22.3 |
| 1 | 2.4 |

## 2. The generative process

The task of the generative process is to stochastically generate a metrical structure, harmonic structure, and stream structure in combination with a pattern of notes (where each note has a pitch, on-time, and off-time). The process also assigns a probability to this joint structure: $P(M, H, S, N)$. This expression is decomposed in the following way:

$$\begin{aligned} P(M,H,S,N) = P(N|M,H,S) \times P(H|M) \\ \times P(S|M) \times P(M). \end{aligned} \tag{4}$$

These dependencies are represented in Figure 2. Essentially, the model first generates a metrical structure; it then generates harmonic and stream structures (these structures are dependent on the meter but independent of each other); finally, it generates the note pattern, which is dependent on all three structural representations. We assume a discrete timeline of points spaced 50 ms apart, known as *pips*; note-onsets and offsets as well as structural events (beats, changes of harmony, and stream beginnings and endings) may only occur at pips.

The generation of the metrical structure is as shown in Figure 3. It can be seen that the structure has four levels, numbered 0 through 3 (where level 3 is the highest, i.e. sparsest, level); level 2 (L2) is assumed to be the tactus level. [This process and the resulting structure are very similar to what was proposed for the monophonic meter-finding model in Temperley (2007), the main difference being that the current structure has four levels instead of three.] The process begins by generating the tactus level. The first tactus interval (a pair of adjacent tactus beats) is generated using a distribution which favours intervals in

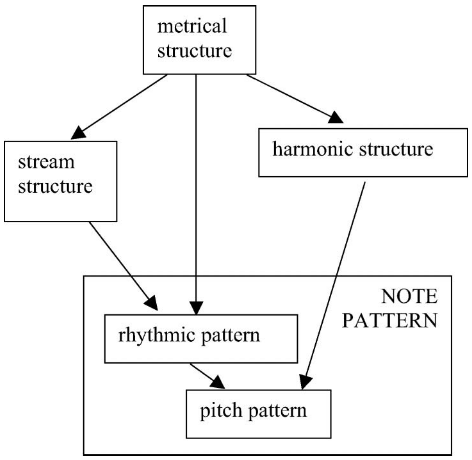


Fig. 2. The structure of the generative process.

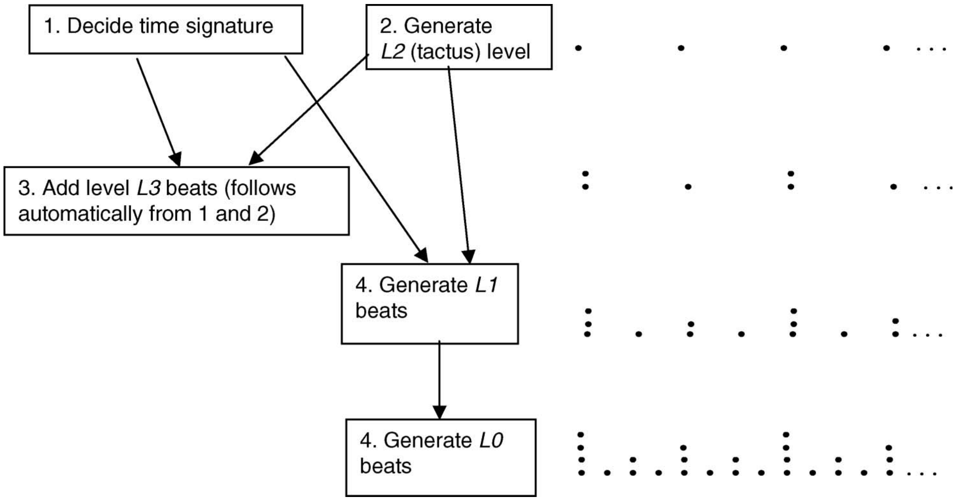

1. Decide time signature
2. Generate *L2* (tactus) level
3. Add level *L3* beats (follows automatically from 1 and 2)
4. Generate *L1* beats
4. Generate *L0* beats

Fig. 3. The generative process for the metrical grid.

the range of 600–800 ms. (The setting of specific parameter values is described below.) Subsequent tactus beats are then generated by a distribution conditional on the previous tactus interval, favouring a tactus level that is roughly regular, but allowing some fluctuation. At each tactus beat, a decision is made as to whether to generate another beat or to end the tactus level; in effect, this determines the length of the piece. Next, decisions are made as to whether level 3 should be duple (in which case every second L3 beat is an L2 beat) or triple, and whether L2 is duple or triple (determining whether tactus intervals should be divided duply or triply). L1 is assumed to be duple, given that in Western music a triple division of the sub-tactus level is extremely rare. Once these decisions are made, the exact placement of each beat at levels 1 and 0 must be determined; for this, distributions are used which favour a roughly equal division of the higher-level beat interval but allow some irregularity. Finally, the 'phase' of L3 must be chosen—whether the first L2 beat is the first, second, or third beat of a L3 beat interval. (The metrical grid is assumed to begin and end on tactus beats.)

The harmonic structure is then generated. The task of the generative process here is to segment the piece into harmonic segments or 'chord-spans', each one labelled with a root. An important simplification here is that harmonic changes are allowed only on tactus beats. (As shown in Table 1, it appears that only a very small percentage of harmonic changes are on sub-tactus beats, so excluding this possibility results in only a small loss of accuracy.) Thus the model's task is simply to choose a root for each tactus interval. For the first tactus interval, a root is chosen out of a uniform distribution. For each subsequent interval, the model first decides whether to continue the previous root or to change to a new root; the probability of change is higher for L3 beats than L2 beats, reflecting the greater likelihood of chord changes on stronger beats (see Table 1). If a new root is chosen, there is a high probability of moving to a root that is a perfect fifth above or below the previous one, reflecting the well-known preference for root motion by fifths in Western music; all other roots are assigned the same low probability.

With regard to the stream structure, the task of the generative process is simply to generate a set of streams, each one spanning a certain portion of the piece. We limit the possible beginning and ending points of streams to tactus beats. (This constraint has no musical justification, but is made simply to limit the space of possible streams; as we will see, it does not imply that the first or last *note* of a stream must be on a tactus beat.) At each tactus beat, for each integer *n* there is a possibility of generating *n* new streams; a Poisson distribution is used here with an expected value of much less than 1. (For the initial tactus beat, a Poisson distribution with an expected value of 2 is used.) Once a stream is generated, at each subsequent tactus beat, a decision is made whether to continue the stream to the next beat or to end the stream. Notice that streams in themselves are not assigned to specific pitches or even to any pitch range.

Once the metrical, harmonic, and stream structures are formed, a pattern of notes is then generated. This process may be broken down into the generation of a rhythmic pattern—a pattern of note-onsets and offsets—and a pitch pattern, assigning a pitch to each note generated. Regarding the note-onset pattern, for each stream, at each pip within the stream, a choice is made as to whether to generate a note-onset at that point. There is an extremely low, but non-zero, probability of note-onsets occurring at non-beat pips (this allows for notes on very weak beats such as 32nd-note beats, and for 'extrametrical' notes such as grace notes). Regarding the probability of onsets at beats, I use a novel method which I call *metrical anchoring* [this was discussed but not implemented in Temperley (2007)]. In general, it is well known that note-onsets are less likely on lower-level beats than higher-level beats (Palmer & Krumhansl, 1990). But the idea of metrical anchoring is that the probability of a note-onset at a beat *depends* on the presence of notes at the surrounding higher-level beats. Consider a weak eighth-note beat with stronger beats on either side (see Figure 4). A note on such a beat is extremely unlikely if there is no note on either side (we call this an 'unanchored' note); it is only slightly more likely if there is a note only on the previous beat ('pre-anchored'), much more likely if there is a note on the following beat ('post-anchored'), and again very likely if there are notes on both beats ('both-anchored'). (Figure 4 provides some statistical evidence on this matter.) Thus the note-onset generation process proceeds in a top-down manner. First decisions are made as to the note status (onset or no onset) of beats at L2 and L3; here each decision is made independent of context, with the onset probability at L3 beats slightly higher than at L2 beats. Then note decisions are made at L1 beats, conditional on the note status of the neighbouring L2 beats; finally, note decisions are made at L0 beats conditional on neighbouring L1 beats. (It is the presence of neighbouring notes *within the same stream* that matters here, thus capturing the interaction with stream structure discussed earlier.) An attractive feature of this approach is that it allows us to indirectly incorporate the preference for longer notes on stronger beats. A long note on a weak eighth-note beat—that is, one with no note on the following quarter-note beat—will either be 'pre-anchored' (if there is a note on the previous quarter-note beat) or 'unanchored' (if there is not), and both of these situations are assigned very low probability.

After all note-onsets have been generated, an offset time must be chosen for each note. We assume, first of all, that the offset of each note must be no later than the following note-onset within the stream (if any). Beyond this, note-offsets present a difficult problem in the modelling of rhythm. In general, we assume that the correct metrical analysis of a piano-roll corresponds to music notation; with regard to the note-onset pattern, the analysis shown in Figure 1 corresponds to the correct music notation and could be converted to it quite easily. But with regard to note-offsets, notes are often played somewhat 'staccato', with the offset much earlier than is indicated by the notation, which makes the correct notation difficult to infer. For example, it is often difficult to decide whether to notate something as 'quarter-note' or 'eighth-note followed by eighth-rest'. In general, I would argue that notations of the latter type are fairly rare; to put it another way, notes are generally notated as ending on tactus beats unless another note-onset intervenes. Thus we generate note-offsets as follows. For each note-onset, at each subsequent tactus beat *T*, we make a stochastic decision as to whether to end the note at *T* or to extend it further; but if, when considering *T*, we find that the onset of the next note in the stream occurs before *T*, the note is ended at that onset with probability 1. If the stream itself ends at *T*, the note may continue past that point by the same stochastic process.

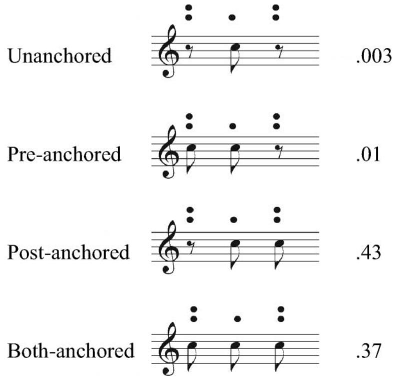


Fig. 4. Four rhythmic patterns (the third rest or note in each pattern may be of any length). The numbers to the right show the probability of a note-onset on the second (weak) beat, given the context of the first and third beats. Data is from the Essen Folksong Collection.

Finally, a pitch is chosen for each note-onset generated. This is conditional on the current harmony and the previous pitch within the stream. We create a 'proximity profile', a normal distribution centred around the previous pitch (Temperley, 2007); a 'chord-profile' is also generated, which favours notes that are chord-tones of the current root and also slightly favours notes within the major and minor scales of the current root. (For the initial note of each stream, the proximity profile is replaced by an 'initial pitch distribution'—a normal distribution across a broad pitch range.) These two profiles are multiplied, creating a distribution favouring pitches that are both chord-tones and close in pitch to the previous pitch; the distribution is normalized to sum to 1, and these probabilities are used to choose the pitch for each note.

Like any probabilistic model, the current model has a number of parameters that must be set: for example, the probability of a note-onset on a level 2 beat, the probability of a stream ending at a tactus beat, and the probability of a change of harmony. However, compared to many probabilistic models, the number of parameters is extremely small. The program contains exactly 50 probability distributions; all but 8 of these are binary distributions, i.e. variables with just two values (thus requiring only one parameter value). Where possible, the variables were set using corpus data; most of the metrical parameters were set using the Essen folksong corpus, a large corpus of over 6000 European folk songs (Schaffrath, 1995). Other parameters were set using trial-and-error testing on a miscellaneous corpus of classical pieces.

## 3. The analytical process

### 3.1 Overview

The aim of the analytical process is to find the most probable metrical, harmonic, and stream structures for a given note pattern. As noted in Equation 2 above, this can be done by maximizing $P(M, H, S, N)$. We define

$$\{M, H, S\}^* = \text{argmax}[M, H, S]P(M, H, S, N). \qquad (5)$$

Given the generative procedure outlined above, with specified parameters, we could quite easily calculate $P(M, H, S, N)$ for a given $\{M, H, S\}$. But how do we find $\{M, H, S\}^*$? The most straightforward method for this would be to consider all possible combinations, as in the pseudo-code algorithm below:

```
$best\_P = 0.0$;
for each $M${
  for each $H${
    for each $S${
      calculate $P(M, H, S, N)$;
      if $P(M, H, S, N) > best\_P$){
            $\{M, H, S\}^* = \{M, H, S\}$;
            $best\_P = P(M, H, S, N)$;
      }
    }
  }
}
```

This procedure is not remotely tractable, even for one component of the structure, let alone for all the components combined. In what follows I explain various techniques that are used for overcoming this search problem. Some of the techniques are approximate while others are exact.

Before beginning the analytical process, note-onsets and offsets are quantized to pips. This is done in a somewhat complex and context-sensitive way, to avoid certain problems such as assigning the notes of a single chord to different pips or creating notes of length zero.

### 3.2 The stream analysis process

The first stage of the search process concerns the stream structure. It was noted earlier that stream structure seems to interact with meter and harmony in the generation of notes: in particular, the probability of a note on a beat depends on the positions of other notes within the stream. However, it appears also that the most probable stream structure for a note pattern can be inferred with reasonable accuracy without consideration of meter and harmony. Roughly speaking, in choosing a stream structure, we simply want to group notes into streams such that the number of streams is fairly small, rests within streams are few and short, and pitch intervals within streams are small; none of these considerations depend heavily on meter and harmony. This point was made in Temperley (2001), where I argued that there are relatively few cases where the inference of streams seems to require knowledge of metrical and harmonic information. Another way to put this is that, given arbitrary stream structures $S_1$ and $S_2$, for any $N$,

$$\begin{aligned}&P(M, H, S_1, N) \approx \propto P(M, H, S_2, N)\\&[\approx \propto, \text{ means `approximately proportional to'}]\end{aligned} \qquad (6)$$

as $M$ and $H$ are varied. Thus if one wishes to find $\{M, H, S\}^*$, this can be done by assuming *any* metrical and harmonic structure $M_x$ and $H_y$ and finding *argmax*[$S$] $P(M_x, H_y, S, N)$; by assumption, this will also be the $S$ of $\{M, H, S\}^*$.

We thus proceed by assuming a fixed $M_x$ and $H_y$ and then finding *argmax*[$S$] $P(M_x, H_y, S, N)$. In so doing, it seemed best to assume a very neutral or 'flat' $M_x$ and $H_y$. The metrical structure we assume is one in which there is just one row of beats roughly 300 ms apart, so that note-onsets are equally likely at all beats (these pseudo-beats are adjusted to coincide with note-onsets and made more dense if necessary so that every note-onset coincides with a beat). With regard to harmony, we assume a completely 'flat' harmonic profile so that all pitch-classes are equally likely. Once the best $S$ is found, this is then assumed in searching for $\{M, H, S\}^*$; at that point, the generative process described earlier is used to calculate $P(M, H, S, N)$, factoring in probabilities for the stream structure itself.

The problem at hand is to recover the most probable stream structure given a note pattern.[2] Given the 'pseudo-metrical structure' just described, the input to the process can be viewed as a two-dimensional array of squares $Q_{c,p}$, with pseudo-beats as columns ($c$) and pitches as rows ($p$), where $Q_{c,p} = onset$ if there is a note-onset at the square, $Q_{c,p} = continuation$ if there is a note continuation, and $Q_{c,p} = blank$ otherwise (see Figure 5). (A note-offset that occurs part way through a square is assumed for present purposes to occur at the end of the square.) We can think of each stream as occupying the squares of each note $x$ that it contains, as well as all the subsequent blank squares at the pitch of $x$ until the following note within the stream. Positing a stream at a certain pitch then simply indicates that the stream

[2] The analytical procedure presented here has much in common with the stream analysis model presented in Temperley (2001), and could (roughly speaking) be regarded as a probabilistic version of that model.

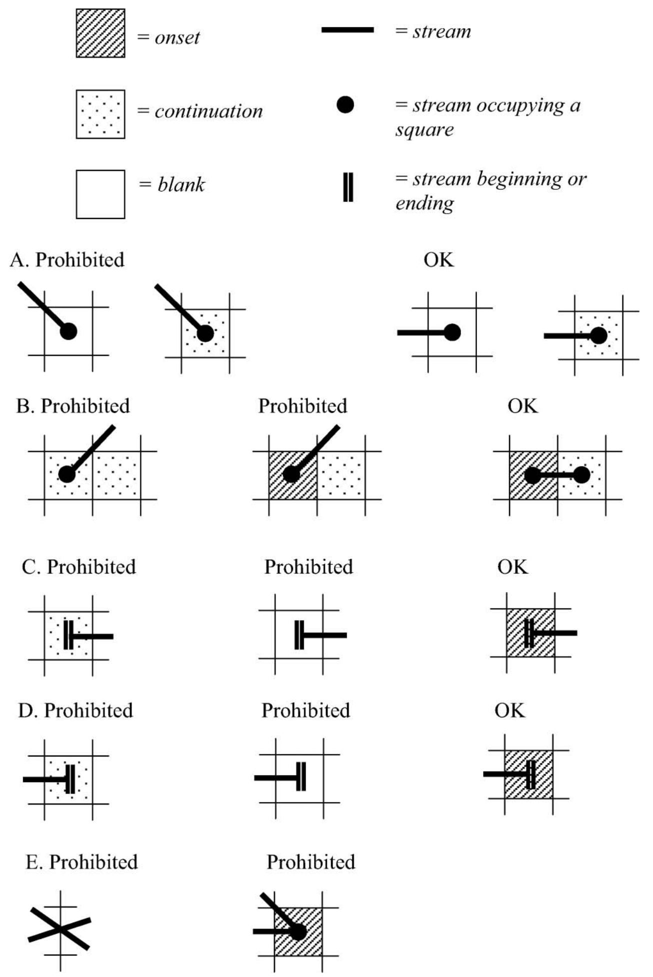


Fig. 5. Legal and illegal stream moves.

contains the most recent note at that pitch. Thus it is meaningless (and therefore illegal) for a stream to move to a pitch with no onset (i.e. moving from $Q_{c,p1}$ to $Q_{c+1,p2}$ when $p2 \neq p1$ and $Q_{c+1,p2} = \mathit{blank}$ or *continuation*; see Figure 5(a)). It is also illegal for a stream to move away from a note while that note is still in progress (i.e. moving from $Q_{c,p1}$ to $Q_{c+1,p2}$ when $p_2 \neq p_1$ and $Q_{c+1,p1} = \mathit{continuation}$; see Figure 5(b)). Every note-onset must be contained in a stream. We also assume that a stream never contains any squares before its first onset or after its last one; thus streams may only begin and end at note-onsets (Figures 5 (c) and (d)). (Since the probability of a note-onset in a square is less than 0.5, this 'minimal' analysis is always more likely than other alternatives.) Notice that, while a stream normally contains both the onset and continuation squares of a note, it contains only the onset square of its final note, not the continuation squares.

Two further constraints are added on the stream analysis process: (1) streams may never cross in pitch; (2) two streams may never occupy the same square (Figure 5(e)). These constraints are not absolute in classical music, but they are generally observed, and they greatly simplify the search process. It may be noted, however, that these constraints were not part of the generative process presented earlier. In effect, it may be assumed that stream structures with crossing and 'colliding' streams are sometimes generated but are then weeded out by some kind of filtering process. The problem with such a step is that the probabilistic model is now no longer well defined. To elaborate this point, we can imagine that the generative process begins with a space of all possible structures and then recursively subdivides a smaller and smaller region of the space as further decisions are made. By filtering out certain structures, we are in effect giving these regions a probability of zero, but other regions are not being adjusted to compensate for this, so the resulting total probability mass is less than 1. Really we should adjust for the loss of mass in some way—for example, by slightly raising the probability of all remaining structures. This does not appear to be a serious problem, however. The main goal of the model is simply to find the most probable structure, and some loss of probability mass does not interfere with this task.[3] It simply means that the probabilities assigned to structures are somewhat lower than they should be.

Let us now consider a situation where a column $C_n$ has certain onset, continuation, and blank squares, and a certain 'column analysis'; a column analysis simply chooses certain squares as being occupied by streams. There must be streams at all onset squares in $C_n$; there may also be streams at other squares. We now wish to continue this stream analysis to the next column $C_{n+1}$. This involves choosing a column analysis for $C_{n+1}$ and also deciding how the streams of $C_{n+1}$ will connect to those at $C_n$. The possibilities are in fact quite limited, due to the 'well-formedness' constraints mentioned above. Within these constraints, the probabilities of various options are defined as described earlier. Distributions dictate the probabilities of beginning and ending streams. The probability of a note-onset (or lack of one) is factored in for every square that is contained within a stream; the probability of a pitch interval is factored in at each note-onset, conditional on the previous pitch level of the stream. The probability of a continuation is factored in for every square that could be a continuation. (A square can only be a continuation if it contains no onset, the previous square within the stream is at the same pitch, and the previous square contains an onset or continuation.) In this way we indirectly discourage

[3] Let us assume that the 'correct' model is the one in which the probabilities of all structures are raised by the same proportion to make up for the loss of mass in the way just suggested. The most probable structure will be the same one with or without this adjustment, as the probabilities of all structures are adjusted equally.

streams from continuing for long periods with no onsets and discourage large pitch intervals within streams.

It can be seen that the legality and 'goodness' of a certain transition and column analysis depend only the preceding column analysis. The probability of beginning and ending streams and the probability of note-onsets and continuations within a stream can be calculated in a purely local fashion; the probability of a note being a certain pitch depends only on the interval to the pitch of the stream at the previous column. Because of the local fashion in which probabilities are calculated, a dynamic programming approach may be used. At each column $C_n$, for each possible column analysis, we find the best global analysis of the piece so far ending with that column analysis. Moving on to the next column, we need only continue each previous 'best-so-far' analysis to each possible column analysis at $C_{n+1}$, factoring in the best transition between them; this allows us to find the new set of best-so-far analyses at $C_{n+1}$, remembering the $C_n$ analysis that each one entails. At the end of the piece, we choose the best of the best-so-far analyses in the final column, and trace it back through the dynamic programming table.

### 3.3 Metrical and harmonic analysis

We now turn to the search for the optimal metrical and harmonic structure. The fact that the stream structure has already been determined somewhat reduces the search problem presented in Section 3.1, but it remains formidable. One way that we simplify the problem is by first considering only levels 0 through 2 of the metrical structure. Level 3—which simply defines every second or third level 2 beat as strong—is then determined on a second pass.

The first stage of the metrical/harmonic analysis process (the identification of the harmony and levels 0, 1, and 2 of the meter) depends on the concept of a 'tactus-root combination' (*TRC*): the combination of a hypothetical tactus interval (two adjacent tactus beats) and a root. The essential idea is that the probability of a certain *TRC* depends only on the previous *TRC*, and the probability of beats and notes within the *TRC* depends only on the *TRC*. Viewed in this way, the metrical/harmonic analysis process can be viewed as a rather complex kind of hidden Markov model.

Suppose we have a hypothetical *TRC* and a pattern of notes within it. (We include notes starting at the first tactus beat of the interval, but not the second.) The notes are also identified as belonging to certain streams. The first step is to determine the most likely locations for L0 and L1 beats. This is essentially done as an exhaustive search. For each possible L1 span, we find the best location for the L0 beat; then, for each L2 span, we find the best way of dividing it into L1 spans. The best duple and triple metrical analyses of the L2 span must both be found. A given lower-level beat analysis can be evaluated by determining the probability of all note-onsets within the span, given that beat pattern. Recall that the probability of a note at an L1 or L0 beat depends on the note status of the neighbouring higher-level beats. Since we are looking at an entire (hypothetical) tactus interval, we have this information, and can determine what the probability of (for example) a note on an L1 beat would be if this *TRC* were actually used. The probability of a note at an L2 beat is defined in a context-free fashion, so this can easily be calculated as well. The probability of note continuations is also calculated; recall that each note in a stream has the option of continuing into the next tactus interval as long as no other note in the stream intervenes.[4] As for the pitches, the probability of each pitch depends only on the current root, which we know, and the interval to the previous pitch within the stream, which we also know. (The previous pitch is not necessarily within the current *TRC*, but it does not depend on the prior metrical-harmonic analysis.)

In assigning pitch probabilities, we also assign a penalty (a reduction in probability) for any note that is not part of the harmony and is not followed by stepwise motion. This is, once again, an *ad hoc* move that is not reflected in the generative process and results in some loss of probability mass, but it seems justified by the resulting improvement in performance.

We are now once again in a position to use dynamic programming—an approach that has long been standard in metrical and harmonic analysis models (Temperley, 1997, 2001; Cemgil et al., 2000b; Raphael & Stoddard, 2004). Suppose we are at a given $TRC_{j,i,R}$ (where $R$ is the root and $j$ and $i$ are the initial and final pips of the interval, respectively) and we wish to find the 'best-so-far' analysis of the piece ending with that *TRC*. More precisely, if $\{M, H, S\}_i$ is a combination of partial structures up to pip $i$, and $P_i\ (M, H, S, N)$ is their joint probability with the note pattern up to pip $i$, we wish to find the $\{M, H, S\}_i$ entailing $TRC_{j,i,R}$ that maximizes $P_i(M, H, S, N)$. We try adding $TRC_{j,i,R}$ on to each previous $TRC_{k,j,R'}$ (where $R'$ is a previous root), whose best-so-far analyses have already been determined. The probability of the tactus interval ($j$, $i$) depends only on the previous tactus interval; the probability of the root $R$ depends only on $R'$; and the probability of the note pattern within $TRC_{j,i,R}$ given that *TRC*, along with its best lower-level metrical analysis, has already been computed. (Regarding the division of L2, we must consider both duple and triple analysis; we add the best duple analysis of $TRC_{j,i,R}$ on to the best duple analysis of

[4] For this purpose, any offset that occurs after tactus beat $T_n$ and before or at beat $T_{n+1}$ is assumed to occur at $T_{n+1}$; following the logic of the generative model, we do not recognize offsets at non-tactus positions, unless the offset coincides with a note-onset.

$TRC_{k,j,R'}$, and similarly for triple.) It is then a simple matter to calculate $P_i(M, H, S, N)$ for each $TRC_{k,j,R'}$ and choose the one yielding the highest value. We proceed in this way in a left-to-right manner through the piece; at each pip $i$, we consider each $TRC_{j,i,R}$ ending at that pip.[5] At the end of the piece, the usual 'traceback' process then applies to find the overall best analysis.

All that remains now is to find level 3 of the metrical structure. Since the tactus level has already been computed, there are just five possibilities: L3 could be duple (with an L3 beat at the first or second L2 beat) or triple (with an L3 beat at the first, second, or third L2 beat). We consider each of these possibilities, factoring in a somewhat higher probability for harmonic changes and note-onsets at L3 beats than L2 beats. (We also factor in 'phase scores' for each of the five possibilities. For example, if L3 is triple, the first L3 beat is quite likely to be the first or second tactus beat; it is unlikely to be the third.) We also redo the harmonic analysis at this stage, considering each possible root for each tactus interval, on the reasoning that the addition of L3 may affect the most optimal points for harmonic change. Since the search process at this stage is not at all expensive, a natural further step would be to expand the harmonic possibilities, e.g. using key-specific names for chords (I/C, V/F, etc.); this would yield a richer harmonic analysis and might improve the metrical analysis as well. This has not been attempted yet, however.

## 4. Testing the analytical model

Research in modelling metrical, harmonic, and stream analysis has been hindered by the absence of agreed methods and materials for testing. Obviously one person can do little to address this problem, but I will attempt to provide some basis for comparison by using materials that were also used for testing the harmonic and metrical models presented in Temperley (2001) (part of the Melisma system). The corpus used there was the Kostka–Payne (K-P) corpus, a set of 46 excerpts from the common-practice repertoire from the workbook accompanying Kostka and Payne's (1995) theory textbook, with harmonic analysis (showing keys and Roman-numeral chord symbols) done by the authors. The excerpts were converted into midifiles as described in Temperley (2001); the harmonic analyses were encoded by Bryan Pardo.

With regard to metrical analysis, a fairly simple method of evaluation will be used here. Any metrical analysis of a piece can be represented with five integers: $TL$ = average tactus length (in milliseconds): $TD$ = division of the tactus level (2 if duple, 3 if triple); $UD$ = division of the level above the tactus (2 if duple, 3 if triple); $TP$ = number of 'pickup' notes, i.e. notes preceding the first tactus beat; $UPh$ = the phase of the upper level, i.e. whether the first tactus beat is the first (1), second (2) or third (3) beat of an upper-level span (the third option is only possible if $UD = 3$). $TL$ is considered correct if it is within 10% of the correct value; all other values must match exactly to be correct. It is important to note that, if the tactus level of an analysis is substantially wrong, then the other statistics reported above are of little interest. (For example, suppose an analysis incorrectly chooses the dotted-quarter note rather than the quarter-note as the tactus, and suppose that $TD = 2$ is correct; $TD = 2$ reported by the model would imply something quite different—a dotted-eighth-note level—and should be considered wrong.) Thus, figures for $TD$, $UD$, $TP$, and $UPh$ are only given for cases where $TL$ is correct.

The probabilistic model and the Melisma model were both tested on the K-P corpus using the evaluation system described above. Table 2 shows the results. Of most interest is the fact that the probabilistic model achieves substantially better results on the tactus level, obtaining a correct result on 37 cases versus 32 for the Melisma model. Regarding other aspects of metrical structure, the two models are very similar in their level of performance. Inspection of the output suggests that the probabilistic model's consideration of harmony is a crucial factor in its superior performance. Figure 6 shows

Table 2. The performance of the Melisma and probabilistic meter-finding models on the Kostka–Payne corpus. $TL$ = tactus length; $TD$ = upper-level division (duple or triple); $UD$ = tactus division (duple or triple); $TP$ = number of pickup notes (before first tactus beat); $UPh$ = phase of upper level.

| | Percentage correct | |
|---|---|---|
| | Melisma model | Probabilistic model |
| Quantized corpus (46 excerpts) | | |
| $TL$ | 32/46 (69.6%) | 37/46 (80.4%) |
| $TD$ | 32/32 (100.0%) | 37/37 (100.0%) |
| $UD$ | 31/32 (96.9%) | 34/37 (91.9%) |
| $TP$ | 29/32 (90.6%) | 36/37 (97.3%) |
| $UPh$ | 27/31 (87.1%) | 30/34 (88.2%) |
| Performed piano corpus (19 excerpts) | | |
| $TL$ | 14/19 (73.7%) | 14/19 (73.7%) |
| $TD$ | 13/14 (92.9%) | 12/14 (85.7%) |
| $UD$ | 10/14 (71.4%) | 12/14 (85.7%) |
| $TP$ | 14/14 (100.0%) | 14/14 (100.0%) |
| $UPh$ | 10/10 (100.0%) | 12/12 (100.0%) |

[5] At the beginning of the piece, we allow the first tactus beat to be anywhere within a range of pips before and including the first note-onset, thus allowing that the first onset may not be on a tactus beat; at the end of the piece, similarly, we consider a range of possible positions for the last tactus.

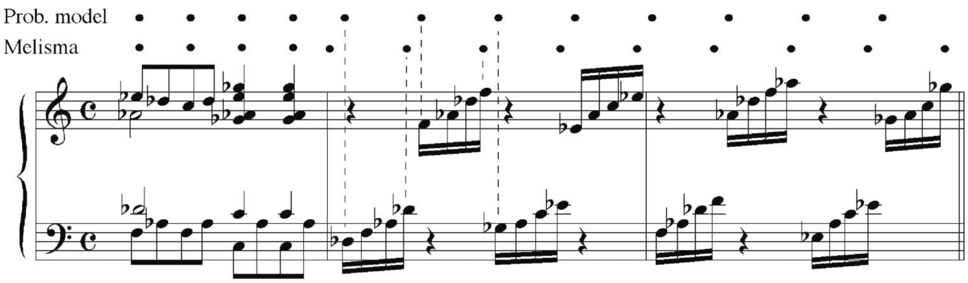


Fig. 6. Beethoven, Rondo Op. 51 No. 1, mm. 109-11 (from the Kostka–Payne corpus), showing the tactus analyses of the Melisma model and the probabilistic model.

a passage from one excerpt from the corpus. The Melisma model places tactus beats one 16th-note too early, as shown; it favours these positions because the coinciding notes are 'long' (by the Melisma model's definition). But the probabilistic model considers harmonic information and thus favours beat locations that correspond with the changes of harmony, as is in fact correct.

The midifiles of the K-P corpus are generated from musical notation, and are therefore 'quantized'—with perfectly regular timing. However, the 19 excerpts from the corpus for solo piano were also performed by a semi-professional pianist and midifiles were generated from these. This allows the model to be tested using the more irregular and complex timing characteristic of human performance. The results for both the probabilistic model and the Melisma model are shown in Table 2; it can be seen that the two models are very close in performance, though the probabilistic model achieves slightly better results on the upper level.

For testing the harmonic model, we again use the K-P corpus. In this case, we simply measure the proportion of time in the entire corpus that the model assigns the correct root. Again, we test both the probabilistic model and the Melisma model. In this case, however, the two models are not quite comparable. The reason is that the Melisma harmonic model requires a metrical analysis as part of the input. In the current test (as in Temperley, 2001), the Melisma model was given the correct metrical structure (as indicated by the score). By contrast, the polyphonic model must infer the metrical structure on its own, and is not always correct, as indicated by the test results reported above. Thus the Melisma model has a significant advantage. Even so, the Melisma model performs only slightly better than the polyphonic model (see Table 3). As another comparison, the performance of the harmonic analysis model of Pardo and Birmingham (2002) is also shown in Table 3; this comparison too is not entirely fair, however, as Pardo and Birmingham's model also identifies chord quality (major/minor/diminished), and they required correct chord quality for a correct answer.

Table 3. Performance of harmonic analysis models on the Kostka–Payne corpus.

| Model | Score |
|---|---|
| Melisma model (% of total time correctly labelled) | 80.8% |
| Pardo and Birmingham 2002 (% of minimal segments correctly labelled)* | 76.5% |
| Probabilistic model (% of total time correctly labelled) | 78.7% |

*A 'minimal segment' is a time segment between successive onsets or offsets. While the Melisma and probabilistic models only produce root judgments, Pardo and Birmingham's model also identifies chord quality (major/minor/diminished), and they required correct chord quality for a correct answer.

On balance, the probabilistic model's metrical analysis is significantly better than Melisma's, and its harmonic performance is nearly as good despite the disadvantage of having imperfect metrical information. Perhaps further refinement of the parameters could yield further improvement. As for the stream component of the model, this is difficult to test; it is frequently unclear in common-practice music what the 'correct' analysis would be, and no annotated corpora are available.

## 5. Transcription

While the performance of the probabilistic model on metrical and harmonic analysis is quite respectable in comparison to the Melisma model, a much more significant advantage of the probabilistic model is its ability to contribute to 'surface-level' processes such as transcription—inferring a note pattern from an auditory signal. As noted earlier, an analytical model could contribute to transcription by assessing the probabilities of note patterns ($N$), which can then serve as the 'prior' for a Bayesian transcription process:

$$P(N|Signal) \propto P(Signal|N)P(N). \qquad (7)$$

In the current case, the probability of a note pattern can be represented as

$$P(N) = \sum_{M,H,S} P(M, H, S, N). \tag{8}$$

A method for computing this arises quite naturally out of the analytical process described earlier. As with the analytical method itself, however, it is somewhat approximate. In the first place, rather than summing the quantity over all stream structures, we consider only the most probable stream structure $S$*, derived in the 'first-pass' stream analysis. Thus we use the estimate

$$P(N) \approx \sum_{M,H} P(M, H, S^*, N). \tag{9}$$

The calculation of this quantity takes place in the metrical-harmonic search described earlier. In choosing a lower-level metrical analysis for each *TRC*, we sum the probabilities of all possible structures being considered. Then, in doing the dynamic-programming search to find the best sequence of *TRC*s, we also compute—for each $TRC_x$—the total probability of all analyses up to that point ending in $TRC_x$. At the end of the piece, the sum of this quantity for all final *TRC*s gives us the quantity in Equation 9.

In order for a note pattern to yield a high probability, there must be some combined metrical/harmonic/stream structure (or perhaps more than one) with which it achieves reasonably high joint probability. This means that the notes should be—for the most part—aligned with regularly spaced beats; they should be grouped into harmonic segments such that most notes within each segment are chord-tones of the same root; and they should be organized into a small number of melodic streams with relatively few long rests and large leaps within streams. If a note pattern does not meet these requirements, its probability will presumably be low, because for every possible combined structure, either $P(N|structure)$ or $P(structure)$ will be low (or perhaps both). Thus, the probabilities assigned by the model reflect a kind of typicality or 'grammaticality' of note patterns within the language of common-practice music.

We can illustrate this point with a simple example. Figure 7(a) shows the Bach minuet passage from Figure 1; Figures 7(b)–(d) show altered versions of the phrase. Beside each example is the probability of the note pattern assigned by the model. It can be seen that the model assigns higher probability to the original pattern than to any of the variants. And indeed, each variant is in some way 'ungrammatical' in relation to the common-practice style, or at least less normative than the original. In Figure 7(b), one of the notes has been displaced by an octave, forcing the model to create a new stream just for that note (which incurs a low probability). In Figure 7(c), one of the right-hand notes has been shifted by one eighth-note, placing it on a weak eighth-note beat rather than a quarter-note beat. And in Figure 7(d), the second D of the melody has been replaced by C#, which makes it a non-chord-tone in relation to the apparent root of G (and one that does not resolve by step).

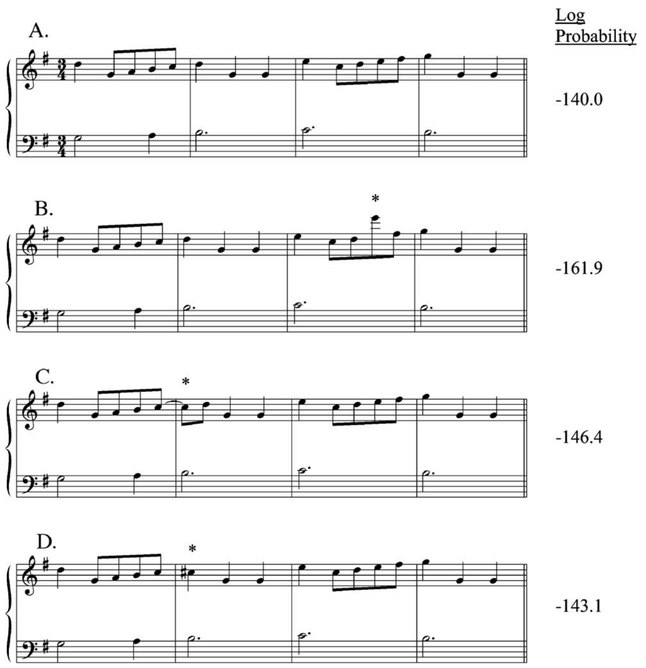


Fig. 7. (a) The Bach passage shown in Figure 1; (b), (c), and (d) show three altered versions of the passage (changes are marked with asterisks). Beside each passage is the log probability assigned to the note pattern by the model.

In this way, the analytical model presented here can be used to evaluate the probability of a note pattern. Combined with a probabilistic signal-processing model which could calculate the 'likelihood' of the signal, $P(Signal|N)$, this would yield a proportional estimate of $P(N|Signal)$. However, this is only one part of the transcription problem. As with higher-level analytical problems, there remains a formidable *search* problem of finding the note pattern that yields the maximal value of $P(N|Signal)P(N)$. In recent work in collaboration with Taylan Cemgil, I have begun to explore a solution to this problem. As in the framework just described, the system uses the analytical model to evaluate the probabilities of note patterns (we will henceforth call this the 'prior' model), as well as a signal-processing model that analyses the resulting likelihood of the signal (the 'signal' model), but the two interact in a rather complex way.

Our solution relies on the concept of a *pitch × pip (PP) array*—a two-dimensional array with pips on one

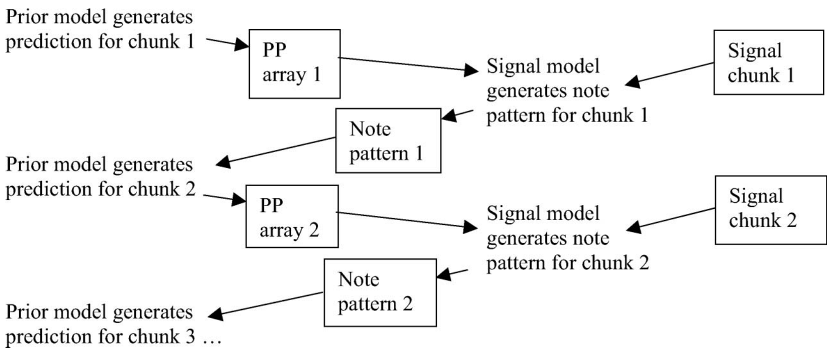

Prior model generates prediction for chunk 1
PP array 1
Signal model generates note pattern for chunk 1
Signal chunk 1
Note pattern 1
Prior model generates prediction for chunk 2
PP array 2
Signal model generates note pattern for chunk 2
Signal chunk 2
Note pattern 2
Prior model generates prediction for chunk 3 …

Fig. 8. The transcription process.

axis and pitch categories on the other. The number in each cell of the array represents the probability of a note-onset at that pip and pitch. The idea is that the prior model generates a PP array which represents its predictions as to the likely locations of note-onsets, both in pitch and time, based on the prior musical context. But of course this requires input from the signal-processing model indicating what notes have already occurred. Thus we use an iterative back-and-forth process. The piece is divided into 'chunks' of one second in length. For each chunk, the prior model produces a PP array; the signal model then uses this information in analysing the signal for the chunk, and returns a determinate note pattern for the chunk. The prior model adds this note pattern on to its note representation, analyses it, and generates a prediction for the next chunk, and so on through the piece (see Figure 8).

Some explanation is needed for how the prior model generates PP arrays. Once again, the stream analysis is handled somewhat separately. At each chunk $K_n$, we assume that the note pattern of chunk $K_{n-1}$ has already been generated; we assume that a stream analysis has already been generated as well (the handling of the first chunk is an exception and will be discussed below). The metrical and harmonic structures of previous chunks are in an indeterminate state, represented by the dynamic programming table. Let us assume just one stream has been found in chunk $K_{n-1}$. Now, for each combination of a pip $x$ and a pitch $y$ in chunk $K_n$, we wish to know

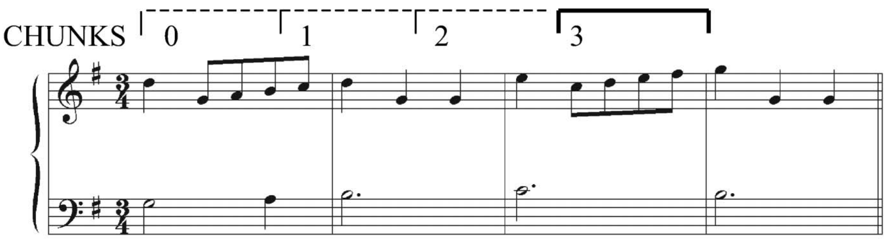
$$\begin{aligned} &P(x = onset, pitch_x = y|PNP) \\ &\quad = P(x = onset|PNP) \times P(pitch_x = y|PNP), \end{aligned} \tag{10}$$

where $pitch_x$ refers to the pitch of the note at pip $x$, if any, and $PNP$ is the note pattern of all prior chunks. Like the analytical process, our calculation of this depends crucially on the concept of a $TRC$ (tactus-root combination). We expand the expression above as follows ($X$ is the set of $TRC$s containing pip $x$):

$$\begin{aligned} &P(x{=}onset|PNP) \times P(pitch_x{=}y|PNP) \\ &= \left(\sum_{TRC\in X} P(x{=}onset, TRC|PNP)\right) \\ &\times \left(\sum_{TRC\in X} P(pitch_x{=}y, TRC|PNP)\right) \\ &= \left(\sum_{TRC\in X} P(x{=}onset|TRC) \times P(TRC,PNP)/P(PNP)\right) \\ &\times \left(\sum_{TRC\in X} P(pitch_x{=}y|TRC) \times P(TRC,PNP)/P(PNP)\right). \end{aligned} \tag{11}$$

Given a $TRC$, the probabilities of notes at different pitches and times can be calculated in a purely local fashion. We consider all possible lower-level metrical analyses of the $TRC$, and for each one we calculate the probability of a note-onset at each pip, adding the appropriate probability mass. (The 'anchoring' method cannot really be used here—since the note status of neighbouring higher-level beats is not known—so we simply define the probability of onsets at each beat level in a context-free fashion.) As for $P(pitch_x{=}y)$, this depends only on the root of the $TRC$ and the interval to the previous note in the stream. $P(TRC, PNP)$ can be calculated as the summed joint probability of the $TRC$ with all prior metrical and harmonic structures, which can easily be found from the dynamic programming table. And $P(PNP)$ can be calculated using the method described above for calculating the probability of complete note patterns. Once the quantity in Equation 11 has been calculated for each active stream, the PP array value for a (pip, pitch) combination simply sums its values for all active streams.

Once the prior model has produced a PP array for an unseen chunk, it then receives a determinate note pattern for that chunk from the signal model, which it must analyse. This required some modifications of the



| | Pip times (msec) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pitches | 3000 | 3150 | 3100 | 3150 | 3200 | 3250 | 3300 | 3350 | 3400 | 3450 | 3500 | 3550 | 3600 | 3650 | 3700 | 3750 | 3800 | 3850 | 3900 | 3950 |
| 30 | -50.8 | -51.8 | -52.4 | -53.5 | -52.2 | -52.4 | -53.1 | -53.6 | -51.8 | -51.6 | -51.7 | -52.3 | -52.3 | -52.4 | -52.7 | -52.9 | -52.3 | -52.0 | -52.2 | -52.2 |
| 31 | -47.4 | -48.5 | -49.2 | -50.3 | -48.9 | -49.2 | -49.8 | -50.4 | -48.6 | -48.5 | -48.6 | -49.2 | -49.2 | -49.3 | -49.7 | -49.8 | -49.2 | -48.9 | -49.2 | -49.1 |
| 32 | -45.0 | -46.0 | -46.6 | -47.7 | -46.4 | -46.6 | -47.3 | -47.8 | -46.0 | -45.8 | -45.9 | -46.5 | -46.5 | -46.6 | -47.0 | -47.1 | -46.5 | -46.2 | -46.4 | -46.4 |
| 33 | -42.2 | -43.2 | -43.8 | -44.9 | -43.5 | -43.8 | -44.4 | -45.0 | -43.2 | -43.0 | -43.1 | -43.7 | -43.7 | -43.8 | -44.1 | -44.3 | -43.7 | -43.4 | -43.6 | -43.6 |
| 34 | -39.5 | -40.5 | -41.1 | -42.3 | -40.9 | -41.1 | -41.8 | -42.4 | -40.6 | -40.4 | -40.5 | -41.1 | -41.1 | -41.2 | -41.5 | -41.7 | -41.1 | -40.7 | -41.0 | -41.0 |
| 35 | -36.8 | -37.8 | -38.5 | -39.6 | -38.2 | -38.5 | -39.1 | -39.7 | -37.9 | -37.7 | -37.9 | -38.5 | -38.5 | -38.6 | -38.9 | -39.1 | -38.5 | -38.1 | -38.4 | -38.4 |
| 36 | -34.4 | -35.4 | -36.0 | -37.2 | -35.8 | -36.0 | -36.7 | -37.3 | -35.5 | -35.3 | -35.4 | -36.0 | -36.0 | -36.1 | -36.5 | -36.6 | -36.0 | -35.7 | -36.0 | -35.9 |
| 37 | -32.4 | -33.3 | -33.9 | -35.1 | -33.7 | -33.9 | -34.6 | -35.1 | -33.3 | -33.1 | -33.2 | -33.8 | -33.8 | -33.9 | -34.2 | -34.4 | -33.8 | -33.4 | -33.7 | -33.6 |
| 38 | -29.8 | -30.8 | -31.4 | -32.6 | -31.2 | -31.4 | -32.1 | -32.6 | -30.9 | -30.7 | -30.8 | -31.4 | -31.4 | -31.5 | -31.9 | -32.0 | -31.4 | -31.1 | -31.3 | -31.3 |
| 39 | -27.7 | -28.8 | -29.4 | -30.5 | -29.1 | -29.4 | -30.0 | -30.6 | -28.8 | -28.6 | -28.8 | -29.3 | -29.3 | -29.4 | -29.8 | -29.9 | -29.3 | -29.0 | -29.3 | -29.2 |
| 40 | -25.5 | -26.6 | -27.2 | -28.4 | -27.0 | -27.2 | -27.9 | -28.4 | -26.7 | -26.5 | -26.6 | -27.2 | -27.2 | -27.3 | -27.7 | -27.8 | -27.2 | -26.9 | -27.2 | -27.1 |
| 41 | -23.9 | -24.8 | -25.4 | -26.6 | -25.2 | -25.4 | -26.1 | -26.7 | -24.8 | -24.7 | -24.8 | -25.3 | -25.3 | -25.4 | -25.8 | -25.9 | -25.3 | -25.0 | -25.3 | -25.2 |
| 42 | -22.0 | -23.0 | -23.6 | -24.7 | -23.4 | -23.6 | -24.3 | -24.8 | -23.0 | -22.8 | -22.9 | -23.5 | -23.5 | -23.6 | -23.9 | -24.1 | -23.5 | -23.2 | -23.4 | -23.4 |
| 43 | -19.8 | -20.9 | -21.6 | -22.7 | -21.3 | -21.6 | -22.2 | -22.8 | -21.0 | -20.9 | -21.0 | -21.6 | -21.6 | -21.7 | -22.1 | -22.2 | -21.6 | -21.3 | -21.6 | -21.5 |
| 44 | -18.6 | -19.6 | -20.2 | -21.3 | -20.0 | -20.2 | -20.9 | -21.4 | -19.6 | -19.4 | -19.5 | -20.1 | -20.1 | -20.2 | -20.6 | -20.7 | -20.1 | -19.8 | -20.0 | -20.0 |
| 45 | -17.0 | -18.0 | -18.6 | -19.7 | -18.3 | -18.6 | -19.2 | -19.8 | -18.0 | -17.8 | -17.9 | -18.5 | -18.5 | -18.6 | -18.9 | -19.1 | -18.5 | -18.2 | -18.4 | -18.4 |
| 46 | -15.5 | -16.5 | -17.1 | -18.3 | -16.9 | -17.1 | -17.8 | -18.4 | -16.6 | -16.4 | -16.5 | -17.1 | -17.1 | -17.2 | -17.5 | -17.7 | -17.1 | -16.7 | -17.0 | -17.0 |
| 47 | -14.0 | -15.0 | -15.7 | -16.8 | -15.4 | -15.7 | -16.3 | -16.9 | -15.1 | -14.9 | -15.1 | -15.7 | -15.7 | -15.8 | -16.1 | -16.3 | -15.7 | -15.3 | -15.6 | -15.6 |
| 48 | -12.8 | -13.8 | -14.4 | -15.6 | -14.2 | -14.4 | -15.1 | -15.7 | -13.9 | -13.7 | -13.8 | -14.4 | -14.4 | -14.5 | -14.9 | -15.0 | -14.4 | -14.1 | -14.4 | -14.3 |
| 49 | -12.0 | -12.9 | -13.5 | -14.7 | -13.3 | -13.5 | -14.2 | -14.7 | -12.9 | -12.7 | -12.8 | -13.4 | -13.4 | -13.5 | -13.8 | -14.0 | -13.4 | -13.0 | -13.3 | -13.2 |
| 50 | -10.6 | -11.6 | -12.2 | -13.4 | -12.0 | -12.2 | -12.9 | -13.4 | -11.7 | -11.5 | -11.6 | -12.2 | -12.2 | -12.3 | -12.7 | -12.8 | -12.2 | -11.9 | -12.1 | -12.1 |
| 51 | -9.7 | -10.8 | -11.4 | -12.5 | -11.1 | -11.4 | -12.0 | -12.6 | -10.8 | -10.6 | -10.8 | -11.3 | -11.3 | -11.4 | -11.8 | -11.9 | -11.3 | -11.0 | -11.3 | -11.2 |
| 52 | -8.7 | -9.8 | -10.4 | -11.6 | -10.2 | -10.4 | -11.1 | -11.6 | -9.9 | -9.7 | -9.8 | -10.4 | -10.4 | -10.5 | -10.9 | -11.0 | -10.4 | -10.1 | -10.4 | -10.3 |
| 53 | -8.3 | -9.2 | -9.8 | -11.0 | -9.6 | -9.8 | -10.5 | -11.1 | -9.2 | -9.1 | -9.2 | -9.7 | -9.7 | -9.8 | -10.2 | -10.3 | -9.7 | -9.4 | -9.7 | -9.6 |
| 54 | -7.6 | -8.6 | -9.2 | -10.3 | -9.0 | -9.2 | -9.9 | -10.4 | -8.6 | -8.4 | -8.5 | -9.1 | -9.1 | -9.2 | -9.5 | -9.7 | -9.1 | -8.8 | -9.0 | -9.0 |
| 55 | -6.6 | -7.7 | -8.4 | -9.5 | -8.1 | -8.4 | -9.0 | -9.6 | -7.8 | -7.7 | -7.8 | -8.4 | -8.4 | -8.5 | -8.9 | -9.0 | -8.4 | -8.1 | -8.4 | -8.3 |
| 56 | -6.6 | -7.6 | -8.2 | -9.3 | -8.0 | -8.2 | -8.9 | -9.4 | -7.6 | -7.4 | -7.5 | -8.1 | -8.1 | -8.2 | -8.6 | -8.7 | -8.1 | -7.8 | -8.0 | -8.0 |
| 57 | -6.2 | -7.2 | -7.8 | -8.9 | -7.5 | -7.8 | -8.4 | -9.0 | -7.2 | -7.0 | -7.1 | -7.7 | -7.7 | -7.8 | -8.1 | -8.3 | -7.7 | -7.4 | -7.6 | -7.6 |
| 58 | -5.9 | -6.9 | -7.5 | -8.7 | -7.3 | -7.5 | -8.2 | -8.8 | -7.0 | -6.8 | -6.9 | -7.5 | -7.5 | -7.6 | -7.9 | -8.1 | -7.5 | -7.1 | -7.4 | -7.4 |
| 59 | -5.6 | -6.6 | -7.3 | -8.4 | -7.0 | -7.3 | -7.9 | -8.5 | -6.7 | -6.5 | -6.7 | -7.3 | -7.3 | -7.4 | -7.7 | -7.9 | -7.3 | -6.9 | -7.2 | -7.2 |
| 60 | -5.6 | -6.6 | -7.2 | -8.4 | -7.0 | -7.2 | -7.9 | -8.5 | -6.7 | -6.5 | -6.6 | -7.2 | -7.2 | -7.3 | -7.7 | -7.8 | -7.2 | -6.9 | -7.2 | -7.1 |
| 61 | -6.0 | -6.9 | -7.5 | -8.7 | -7.3 | -7.5 | -8.2 | -8.7 | -6.9 | -6.7 | -6.8 | -7.4 | -7.4 | -7.5 | -7.8 | -8.0 | -7.4 | -7.0 | -7.3 | -7.2 |
| 62 | -5.8 | -6.8 | -7.4 | -8.6 | -7.2 | -7.4 | -8.1 | -8.6 | -6.9 | -6.7 | -6.8 | -7.4 | -7.4 | -7.5 | -7.9 | -8.0 | -7.4 | -7.1 | -7.3 | -7.3 |
| 63 | -6.1 | -7.2 | -7.8 | -8.9 | -7.5 | -7.8 | -8.4 | -9.0 | -7.2 | -7.0 | -7.2 | -7.7 | -7.7 | -7.8 | -8.2 | -8.3 | -7.7 | -7.4 | -7.7 | -7.6 |
| 64 | -6.3 | -7.4 | -8.0 | -9.2 | -7.8 | -8.0 | -8.7 | -9.2 | -7.4 | -7.3 | -7.4 | -8.0 | -8.0 | -8.1 | -8.5 | -8.6 | -8.0 | -7.7 | -8.0 | -7.9 |
| 65 | -7.0 | -8.0 | -8.6 | -9.8 | -8.4 | -8.6 | -9.3 | -9.9 | -8.0 | -7.8 | -8.0 | -8.5 | -8.5 | -8.6 | -9.0 | -9.1 | -8.5 | -8.2 | -8.5 | -8.4 |
| 66 | -7.6 | -8.6 | -9.2 | -10.3 | -8.9 | -9.1 | -9.8 | -10.4 | -8.6 | -8.4 | -8.5 | -9.1 | -9.0 | -9.2 | -9.5 | -9.7 | -9.1 | -8.7 | -9.0 | -8.9 |
| 67 | -7.7 | -8.7 | -9.4 | -10.5 | -9.1 | -9.4 | -10.0 | -10.6 | -8.8 | -8.7 | -8.8 | -9.4 | -9.4 | -9.5 | -9.9 | -10.0 | -9.4 | -9.1 | -9.4 | -9.3 |
| 68 | -8.3 | -9.3 | -9.9 | -11.0 | -9.7 | -9.9 | -10.6 | -11.1 | -9.3 | -9.1 | -9.3 | -9.8 | -9.8 | -9.9 | -10.3 | -10.4 | -9.8 | -9.5 | -9.7 | -9.7 |
| 69 | -8.0 | -9.0 | -9.6 | -10.7 | -9.4 | -9.6 | -10.3 | -10.8 | -9.0 | -8.8 | -8.9 | -9.5 | -9.5 | -9.6 | -10.0 | -10.1 | -9.5 | -9.2 | -9.4 | -9.4 |
| 70 | -7.5 | -8.5 | -9.1 | -10.2 | -8.9 | -9.1 | -9.8 | -10.3 | -8.5 | -8.3 | -8.5 | -9.1 | -9.0 | -9.1 | -9.5 | -9.6 | -9.1 | -8.7 | -9.0 | -8.9 |
| 71 | -6.8 | -7.8 | -8.5 | -9.6 | -8.2 | -8.5 | -9.1 | -9.7 | -7.9 | -7.7 | -7.9 | -8.5 | -8.4 | -8.6 | -8.9 | -9.1 | -8.5 | -8.1 | -8.4 | -8.4 |
| 72 | -6.4 | -7.4 | -8.0 | -9.2 | -7.8 | -8.0 | -8.7 | -9.3 | -7.5 | -7.3 | -7.4 | -8.0 | -8.0 | -8.1 | -8.5 | -8.6 | -8.0 | -7.7 | -8.0 | -7.9 |
| 73 | -6.4 | -7.3 | -7.9 | -9.1 | -7.7 | -7.9 | -8.6 | -9.1 | -7.3 | -7.1 | -7.2 | -7.8 | -7.8 | -7.9 | -8.2 | -8.4 | -7.8 | -7.4 | -7.7 | -7.6 |
| 74 | -5.8 | -6.8 | -7.4 | -8.6 | -7.2 | -7.4 | -8.1 | -8.6 | -6.9 | -6.7 | -6.8 | -7.4 | -7.4 | -7.5 | -7.9 | -8.0 | -7.4 | -7.1 | -7.3 | -7.3 |
| 75 | -5.7 | -6.8 | -7.4 | -8.5 | -7.1 | -7.4 | -8.0 | -8.6 | -6.8 | -6.6 | -6.8 | -7.3 | -7.3 | -7.4 | -7.8 | -7.9 | -7.3 | -7.0 | -7.3 | -7.2 |
| 76 | -5.5 | -6.6 | -7.2 | -8.4 | -7.0 | -7.2 | -7.9 | -8.4 | -6.7 | -6.5 | -6.6 | -7.2 | -7.2 | -7.3 | -7.7 | -7.8 | -7.2 | -6.9 | -7.2 | -7.1 |
| 77 | -5.9 | -6.8 | -7.4 | -8.6 | -7.2 | -7.4 | -8.1 | -8.7 | -6.8 | -6.7 | -6.8 | -7.3 | -7.3 | -7.4 | -7.8 | -7.9 | -7.3 | -7.0 | -7.3 | -7.2 |
| 78 | -6.0 | -7.0 | -7.6 | -8.7 | -7.4 | -7.6 | -8.3 | -8.8 | -7.0 | -6.8 | -6.9 | -7.5 | -7.5 | -7.6 | -7.9 | -8.1 | -7.5 | -7.2 | -7.4 | -7.4 |
| 79 | -5.8 | -6.9 | -7.6 | -8.7 | -7.3 | -7.6 | -8.2 | -8.8 | -7.0 | -6.9 | -7.0 | -7.6 | -7.6 | -7.7 | -8.1 | -8.2 | -7.6 | -7.3 | -7.6 | -7.5 |
| 80 | -6.6 | -7.6 | -8.2 | -9.3 | -8.0 | -8.2 | -8.9 | -9.4 | -7.6 | -7.4 | -7.5 | -8.1 | -8.1 | -8.2 | -8.6 | -8.7 | -8.1 | -7.8 | -8.0 | -8.0 |
| 81 | -7.0 | -8.0 | -8.6 | -9.7 | -8.3 | -8.6 | -9.2 | -9.8 | -8.0 | -7.8 | -7.9 | -8.5 | -8.5 | -8.6 | -8.9 | -9.1 | -8.5 | -8.2 | -8.4 | -8.4 |
| 82 | -7.5 | -8.5 | -9.1 | -10.3 | -8.9 | -9.1 | -9.8 | -10.4 | -8.6 | -8.4 | -8.5 | -9.1 | -9.1 | -9.2 | -9.5 | -9.7 | -9.1 | -8.7 | -9.0 | -9.0 |
| 83 | -8.0 | -9.0 | -9.7 | -10.8 | -9.4 | -9.7 | -10.3 | -10.9 | -9.1 | -8.9 | -9.1 | -9.7 | -9.7 | -9.8 | -10.1 | -10.3 | -9.7 | -9.3 | -9.6 | -9.6 |
| 84 | -8.8 | -9.8 | -10.4 | -11.6 | -10.2 | -10.4 | -11.1 | -11.7 | -9.9 | -9.7 | -9.8 | -10.4 | -10.4 | -10.5 | -10.9 | -11.0 | -10.4 | -10.1 | -10.4 | -10.3 |
| 85 | -10.0 | -10.9 | -11.5 | -12.7 | -11.3 | -11.5 | -12.2 | -12.7 | -10.9 | -10.7 | -10.8 | -11.4 | -11.4 | -11.5 | -11.8 | -12.0 | -11.4 | -11.0 | -11.3 | -11.2 |
| 86 | -10.6 | -11.6 | -12.2 | -13.4 | -12.0 | -12.2 | -12.9 | -13.4 | -11.7 | -11.5 | -11.6 | -12.2 | -12.2 | -12.3 | -12.7 | -12.8 | -12.2 | -11.9 | -12.1 | -12.1 |
| 87 | -11.7 | -12.8 | -13.4 | -14.5 | -13.1 | -13.4 | -14.0 | -14.6 | -12.8 | -12.6 | -12.8 | -13.3 | -13.3 | -13.4 | -13.8 | -13.9 | -13.3 | -13.0 | -13.3 | -13.2 |
| 88 | -12.7 | -13.8 | -14.4 | -15.6 | -14.2 | -14.4 | -15.1 | -15.6 | -13.9 | -13.7 | -13.8 | -14.4 | -14.4 | -14.5 | -14.9 | -15.0 | -14.4 | -14.1 | -14.4 | -14.3 |
| 89 | -14.3 | -15.2 | -15.8 | -17.0 | -15.6 | -15.8 | -16.5 | -17.1 | -15.2 | -15.1 | -15.2 | -15.7 | -15.7 | -15.8 | -16.2 | -16.3 | -15.7 | -15.4 | -15.7 | -15.6 |
| 90 | -15.6 | -16.6 | -17.2 | -18.3 | -17.0 | -17.2 | -17.9 | -18.4 | -16.6 | -16.4 | -16.5 | -17.1 | -17.1 | -17.2 | -17.5 | -17.7 | -17.1 | -16.8 | -17.0 | -17.0 |

Fig. 9. A PP array for Chunk 3 of the Bach passage.

program to operate in a more left-to-right ('causal') manner. Let us assume we have produced a PP array for an unseen chunk $K_n$, extending from pip $i$ to pip $j$, and have now received a note pattern for it. As before, the stream analysis occurs first; the first-pass stream analysis procedure, which has already been used to find the stream structure of the previous note pattern, is simply continued from pips $i$ through $j$. The metrical-harmonic analysis for the chunk is then performed, by continuing the dynamic-programming table for the region $(i, j)$; as noted earlier, however, no determinate metrical-harmonic analysis is chosen (though this can easily be done at any time if desired, e.g. at the end of the piece). The 'second-pass' metrical-harmonic analysis, in which level 3 of the meter is found and the harmony is reanalysed, is simply skipped. Given this analysis of chunk $K_n$, the program is then ready to generate the prediction for chunk $K_{n+1}$.

While the transcription system described above is still under development, the 'prior model' component is complete. A sample PP array generated by the prior model is shown in Figure 9. This is for a chunk of the Bach minuet shown in Figure 1 (shown again in Figure 9): chunk 3, marked with a solid bracket. The PP array is the prior model's prediction for this 'unseen' chunk, given the note pattern of previous chunks. (Rather than using note patterns produced by the signal model, here we simply use the correct note patterns for chunks 0 through 2.) Each value in the array represents the log probability of a note-onset at that pip and pitch. Several features of the PP array deserve mention. In the time dimension, one can see a peak in the probability values at the very beginning of the chunk (time 3000), followed by a decline, and subsequent peaks at around times 3450 and 3850. This reflects the influence of the metrical structure; the previous context has suggested a tactus level of around 450 ms, with a previous beat at time 2600. The chunk also reflects smaller peaks halfway between the tactus peaks, representing the expected locations for L1 events. In the pitch dimension, if we consider just the first pip (or any other pip for that matter), peaks can be seen at pitches 60 (C4) and 76 (E5); this is because there are two active streams in the previous context, and the most recent pitches in the two streams are C4 and E5. Probabilities decrease gradually as we move away from these pitches, reflecting the fact that the most likely next pitch in each voice is likely to be close to the previous one. However, the curve is somewhat irregular. For example, pitch 62 (D4, two half-steps away from C4) has a higher value than 61 (C#4, one half-step away). No doubt this is due to the effect of harmony. Under the most probable analysis of the previous context, the previous tactus interval had a root of C; the most probable harmonic continuations from this root are C (involving no harmonic change) and G and F (involving root motion by fifth). D is a scale-tone in relation to C and a chord-tone in relation to G, whereas C# is not a chord-tone of any of these roots and a scale-tone only in relation to F (minor). In this example, then, we see the effect of meter, harmony, and stream structure on the prediction of future notes.

A special situation occurs at the initial chunk of a piece; in this case, the prior model must make a prediction for the chunk without any previous note pattern to guide it. In particular, there are no active streams that can be used to set pitch probabilities. Thus the model assigns pitch probabilities using an initial pitch distribution similar to that described in the analytical model. Beyond this, the model's procedure for the first chunk is essentially the same as for non-initial chunks. The result is a very 'flat' PP array which assigns more or less equal probability to all pips and pitches (except for a increase towards the middle of the pitch range). Clearly, this is of limited value to the signal model, and it can be expected that the resulting note pattern returned by the signal model will be of poorer quality than that of subsequent chunks.

As mentioned earlier, the transcription system described above is still under construction. Work is ongoing to refine the signal model and the interaction of the signal model and prior model.

## 6. Conclusions

While the current model leaves room for improvement in many ways, it demonstrates the overall viability of a unified approach to the modelling of polyphonic music, which identifies multiple kinds of structure and at the same time estimates the probabilities of note patterns. There are many other kinds of musical knowledge that could be incorporated into the model—for example, more detailed knowledge about harmony (knowledge of functional harmony and of stylistic progressions such as cadences), knowledge of conventional phrase structures (the norm of 4-bar phrases), and awareness of repeated melodic patterns or 'parallelisms' (which play an important role in meter, among other things). The challenges will be, first, to find logical ways of integrating these kinds of musical knowledge with the existing generative process, and second, to keep the computational complexity of the inference problem at a tractable level, either with exact methods such as dynamic programming or with reasonable approximations.

With regard to the transcription problem, the system sketched above appears to be the first concerted attempt to bring to bear higher-level musical knowledge on the transcription process. It remains to be seen how much benefit this knowledge will yield. The fairly simple interactive procedure outlined in Figure 8 could of course be expanded in various ways. It is possible, for example, that the 'left-to-right' process of generating PP arrays could be combined with a 'right-to-left' process, so that information from the signal model could lead to improved predictions for previous chunks, though that would not fit so naturally with the model's left-to-right analytical process. It is also possible that, once a note pattern was chosen, it could be adjusted by the analytical model in an incremental 'hill-climbing' fashion, by deleting or adding notes or shifting them in pitch or time, so as to improve the overall probability of the note pattern. (One can imagine, for example, that such a system might be able to adjust Figures 7(b)–(d) to produce Figure 7(a).)

The applications and implications of the current project are numerous. The many possible uses of an accurate polyphonic transcription system are apparent and well known. Structural analysis also has important applications. For example, in classifying music by style, judging the similarity of two pieces, or extracting salient melodic and thematic material from a piece, the availability of structural information such as meter, harmony and stream structure would certainly be helpful. The model may also have important implications for the modelling of cognition. Experimental work has made clear that metrical, harmonic and stream structure have broad psychological reality among listeners (for a survey of the evidence, see Temperley, 2001). It seems clear, also, that listeners have at least some 'transcription' ability, in that they can recover some note information from polyphonic audio input; for example, most listeners can identify at least the melody in a classical polyphonic piece of moderate complexity. As argued in Temperley (2007), the ability to assign probabilities to note patterns may also be involved in cognitive processes such as expectation and error-detection. Most previous experimental and computational work has treated such 'surface-level' processes as separate and independent from structural processes such as metrical and harmonic analysis. But the current model suggests that they may in fact be closely intertwined, and offers a proposal as to how they may be accomplished within a single unified framework.

## References

Cambouropoulos, E. (2008). Voice and stream: Perceptual and compuational modeling of voice separation. *Music Perception*, *26*, 75–94.

Cemgil, A.T., Desain, P. & Kappen, B. (2000a). Rhythm quantization for transcription. *Computer Music Journal*, *24*(2), 60–76.

Cemgil, A.T. & Kappen, B. (2003). Monte Carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, *18*, 45–81.

Cemgil, A.T., Kappen, B. & Barber, D. 2005. A generative model for music transcription. *IEEE Transactions on Speech and Audio Processing*, *14*(2), 679–694.

Cemgil, A.T., Kappen, B., Desain, P. & Honing, H. (2000b). On tempo tracking: Tempopgram representation and Kalman filtering. *Journal of New Music Research*, *29*, 259–273.

Chai, W. & Vercoe, B. (2001). Folk music classification using hidden Markov models. In *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas, USA. Las Vegas: CSREA Press.

Davy, M. (2006). Multiple fundamental frequency estimation based on generative models. In A.P. Klapuri & M. Davy (Eds.), *Signal Processing Methods for Music Transcription* (pp. 203–227). New York: Springer.

de la Higuera, C., Piat, F. & Tantini, F. (2005). Learning stochastic finite automata for musical style recognition. In *Proceedings of CIAA 2005*, Sophia Antipolis, France (pp. 345–346). Berlin: Springer Verlag.

Jones, M.R., Moynihan, H., MacKenzie, N. & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *13*, 313–319.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1998). Application of Bayesian probability networks to musical scene analysis. In D.F. Rosenthal & H.G. Okuno (Eds.), *Computational Auditory Scene Analysis* (pp. 115–137). Mahwah, NJ: Lawrence Erlbaum.

Kirlin, P. & Utgoff, P. (2005). VOISE: Learning to segregate voices in explicit and implicit polyphony. In J. Reiss & G. Wiggins (Eds.), *Proceedings of the Sixth International Conference on Music Information Retrieval* (pp. 552–557). London, UK: University of London.

Kostka, S. & Payne, D. (1995). *Workbook for Tonal Harmony*. New York: McGraw Hill.

Klapuri A.P. 2004. Automatic music transcription as we know it today. *Journal of New Music Research*, *33*, 269–282.

Klapuri, A.P. & Davy, M. (2006). *Signal Processing Methods for Music Transcription*. New York: Springer.

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

Maxwell, H.J. (1992). An expert system for harmonic analysis of tonal music. In M. Balaban, K. Ebcioglu, & O. Laske (Eds.), *Understanding Music with AI* (pp. 335–353). Cambridge, MA: MIT Press.

Palmer, C. & Krumhansl, C. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 728–741.

Pardo, B. & Birmingham, W. (2002). Algorithms for chordal analysis. *Computer Music Journal*, *26*(2), 27–49.

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

Povel, D.-J. & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*, 411–440.

Raphael, C. (2002). A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, *137*, 217–238.

Raphael, C. & Stoddard, J. (2004). Functional harmonic analysis using probabilistic models. *Computer Music Journal*, *28*(3), 45–52.

Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal*, *16*(1), 64–76.

Schaffrath, H. (1995). In D. Huron (Ed.), *The Essen Folksong Collection*. Stanford, CA: Center for Computer-Assisted Research in the Humanities.

Schellenberg, E.G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception*, *14*, 295–318.

Temperley, D. (1997). An algorithm for harmonic analysis. *Music Perception*, *15*, 31–68.

Temperley, D. (2001). *The Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press.

Temperley, D. (2004). Bayesian models of musical structure and cognition. *Musicae Scientiae*, *8*, 175–205.

Temperley, D. (2007). *Music and Probability*. Cambridge, MA: MIT Press.