Automated Rhythm Transcription

Christopher Raphael*
Department of Mathematics and Statistics
University of Massachusetts, Amherst
raphael@math.umass.edu

May 21, 2001

Abstract

We present a technique that, given a sequence of musical note onset times, performs simultaneous identification of the notated rhythm and the variable tempo associated with the times. Our formulation is probabilistic: We develop a stochastic model for the interconnected evolution of a rhythm process, a tempo process, and an observable process. This model allows the globally optimal identification of the most likely rhythm and tempo sequence, given the observed onset times. We demonstrate applications to a sequence of times derived from a sampled audio file and to MIDI data.

1 Introduction

A central challenge of music IR is the generation of music databases in formats suitable for automated search and analysis [1], [2], [3], [4], [5], [6]. While a certain amount of information can always be compiled by hand, the thought of "typing in," for example, the complete works of Mozart seems daunting, to say the least. Given the enormity of such tasks we expect that automatic music transcription will play an important role in the construction of music databases.

We address here a component of this automatic transcription task: Given a sequence of times, we wish to identify the corresponding musical rhythm. We refer to this problem as "Rhythmic Parsing." The sequences of times that form the input to our system could come from a MIDI file or be estimated from (sampled) audio data. On output, the rhythmic parse assigns a score position, a (measure number, measure position) pair, to each time.

A trained musician's rhythmic understanding results from simultaneous identification of rhythm, tempo, pitch, voicing, instrumentation, dynamics, and other aspects of music. The advantage of posing the music recognition problem as one of simultaneous estimation is that each aspect of the music can inform the recognition of any other. For instance, the estimation of rhythm is greatly enhanced by dynamic information since, for example, strong beats are often points of dynamic emphasis. While we acknowledge that in restricting our attention to timing information we exclude many useful clues, we feel that the basic approach we present is extendible to more complex inputs.

We are aware of several applications of rhythmic parsing. Virtually every commercial score-writing program now offers the option of creating scores by directly entering MIDI data from a keyboard. Such programs must infer the rhythmic content from the time-tagged data and, hence, must address the rhythmic parsing problem. When the input data is played with anything less than mechanical precision, the transcription degrades rapidly, due to the difficulty in computing the correct rhythmic parse.

*This work is supported by NSF grant IIS-9987898.

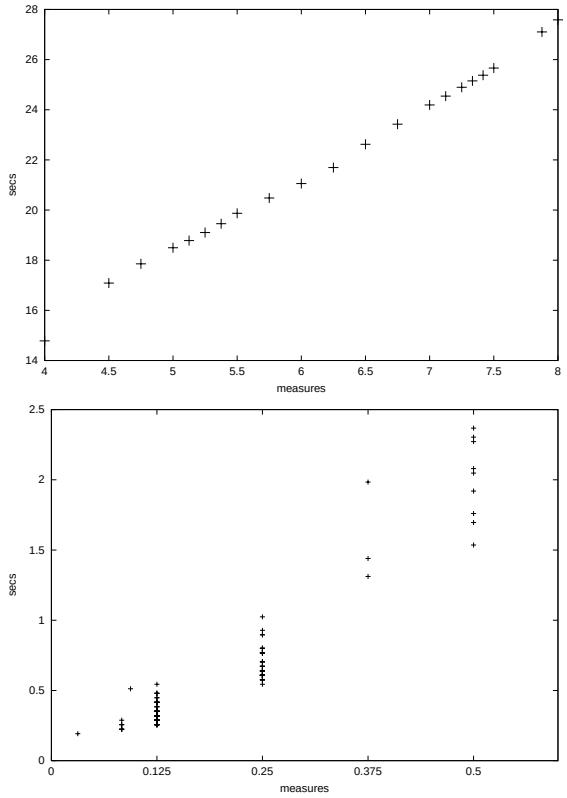

Figure 1: **Top:** Real time (seconds) vs. Musical time (measures) for a musical excerpt. **Bottom:** The actual inter onset intervals (seconds) of notes grouped by the musical duration (measures).

Rhythmic parsing also has applications in musicology where it could be used to separate the inherently intertwined quantities of notated rhythm and expressive timing [7], [8], [9]. Either the rhythmic data or the timing information could be the focal point of further study. Finally, the musical world eagerly awaits the compilation of music databases containing virtually every style and genre of (public domain) music. The construction of such databases will likely involve several transcription efforts including optical music recognition, musical audio signal recognition, and MIDI transcription. Rhythmic parsing is an essential ingredient to the latter two efforts.

Consider the data in the top panel of Figure 1 containing estimated note times from an excerpt of Schumann's 2nd Romance for oboe and piano (oboe part only). The actual audio file can be heard at http://fafner.math.umass.edu/rhythmic_parsing. In this figure we have plotted the score position of each note, in measures, versus the actual onset time, in seconds. The points trace out a curve in which the player's tempo can be seen as the slope of the curve. The example illustrates a very common situation in music: The tempo is not a single fixed number, but rather a time-varying quantity. Clearly such time-varying tempi confound the parsing problem leading to a "chicken and egg" problem: To estimate the rhythm, one needs to know the tempo process and vice-versa.

Most commercially available programs accomplish the rhythmic parsing task by *quantizing* the observed note lengths, or more precisely inter-onset intervals (IOIs), to their closest note values (eighth note, quarter note, etc.), given a known tempo, or quantizing the observed note onset times to the closest points in a rigid grid [10]. While such quantization schemes can work reasonably well when the music is played with robotic precision (often a metronome is used), they perform poorly when faced with the more expressive and less accurate playing typically encountered. Consider the bottom panel of Figure 1 in which we have plotted the written note lengths in measures versus the actual note lengths (IOIs) in seconds from our musical excerpt. The large degree of overlap between the empirical distributions of each note length class demonstrates the futility of assigning note lengths through note-by-note quantization in this example.

We are aware of several research efforts in this direction. Some of this research addresses the problem of *beat induction*, or *tempo tracking* in which one tries to estimate a sequence of times corresponding to evenly spaced musical intervals (e.g. beats) for a given sequence of observed note onset times [11], [12]. The main issue here is trying to follow the tempo rather than transcribing the rhythm. Another direction addresses the problem of rhythmic transcription by assigning simple integer ratios to observed note lengths without any corresponding estimation of tempo [13], [14], [15]. The latter two of these approaches assume that beat induction has already been performed, whereas the former assumes that tempo variations are not significant enough to obscure the ratios of neighboring note lengths.

In many kinds of music we believe it will be exceedingly difficult to *independently* estimate tempo and rhythm, as in the cited research, since the ob-

served data is formed from a complex interplay between the two, as illustrated by the example of Figure 1. Thus, in this work we address the problem of *simultaneous* estimation of tempo and rhythm; in the following we refer to such a simultaneous estimate as a *rhythmic parse*. From a problem domain point of view, our focus on simultaneous estimation is the most significant contrast between our work and other efforts.

2 The Model

We construct a generative model that describes the simultaneous evolution of three processes: a rhythm process, a tempo process, and an observable process. The rhythm process takes on values in a finite set of possible measure positions whereas the tempo process is *continuous-valued*. In our model, these two interconnected processes are not directly observable. What we observe is the sequence of inter-onset intervals (IOIs) which depend on both unobservable quantities.

To be more specific, suppose we are given a sequence of times $o_0, o_1, \ldots, o_N$, in seconds, at which note onsets occur. These times could be estimated from audio data, as in the example in Figure 1, or could be times associated with MIDI "note-ons." Suppose we also have a finite set, $\mathcal{S}$, composed of the possible *measure positions* a note can occupy. For instance, if the music is in 6/8 time and we believe that no subdivision occurs beyond the eighth note, then

$$\mathcal{S} = \{\frac{0}{6}, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}\}$$

More complicated subdivision rules could lead to sets, $\mathcal{S}$, which are not evenly spaced multiples of some common denominator, as shown in the experiments of Section 4. We assume only that the possible onset positions of $\mathcal{S}$ are rational numbers in $[0, 1)$, decided upon in advance. Our goal, in part, is to associate each note onset o_n with a score position — a pair consisting of a measure number and an element of $\mathcal{S}$. For the sake of simplicity, assume that no two of the $\{o_n\}$ can be associated with the exact same score position as would be the case for data from a single monophonic instrument. We will drop this assumption in the second example we treat.

We model this situation as follows. Let $S_0, S_1, \ldots, S_N$ be the discrete measure position process, $S_n \in \mathcal{S}, n = 0, \ldots, N$. In interpreting these positions we assume that each consecutive pair of positions corresponds to a note length of at most one measure. For instance, in the 6/8 example given above $S_n = 0/6, S_{n+1} = 1/6$ would mean the nth note begins at the start of the measure and lasts for one eighth note, while $S_n = 1/6, S_{n+1} = 0/6$ would mean the nth note begins at the second eighth note of the measure and lasts until the "downbeat" of the next measure. We can then use $l(s, s')$,

$$l(s, s') = \begin{cases} s' - s & \text{if } s' > s \\ 1 + s' - s & \text{otherwise} \end{cases} \quad (1)$$

to unambiguously represent the length, in measures, of the transition from s to s'. Note that we can recover the actual score positions from the measure position process. That is, if $S_0 = s_0, S_1 = s_1, \ldots, S_N = s_N$ then score position, in measures, of the nth note is $m_n = s_0 + l(s_0, s_1) + \ldots, l(s_{n-1}, s_n)$. Extending this model to allow for notes longer than a measure complicates our notation slightly, but requires no change of our basic approach. We model the S process as a time-homogeneous Markov chain with initial distribution $p(s_0)$ and transition probability matrix

$$R(s_{n-1}, s_n) = p(s_n | s_{n-1})$$

With a suitable choice of the matrix R, the Markov model captures important information for rhythmic parsing. For instance, R could be chosen to express the notion that, in 4/4 time, the last sixteenth note of the measure will very likely be followed by the downbeat of the next measure: $R(15/16, 0/16) \approx 1$. In practice, R should be learned from actual rhythm data. When R accurately reflects the nature of the data being parsed, it serves the role of a musical expert that guides the recognition toward musically plausible interpretations.

The tempo is the most important link between the printed note lengths, $l(S_n, S_{n+1})$, and the observed note lengths, $o_{n+1} - o_n$. Let $T_1, T_2, \ldots, T_N$ be the continuously-valued tempo process, measured in *seconds per measure*, which we model as follows. We let the initial tempo be modeled by

$$T_1 \sim N(\nu, \phi^2)$$

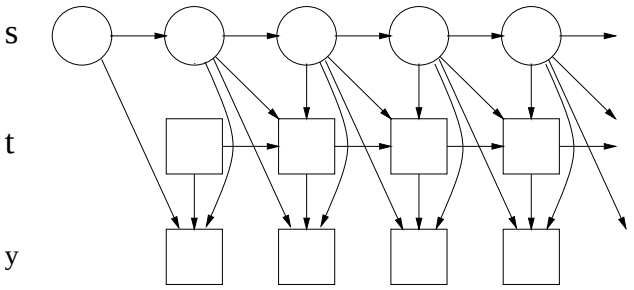

Figure 2: The DAG describing the dependency structure of the variables of our model. Circles represent discrete variables while squares represent continuous variables.

where $N(\nu, \phi^2)$ represents the normal distribution with mean ν and variance ϕ^2. With appropriate choice of ν and ϕ^2 we express both what we "expect" the starting tempo to be (ν) and how confident we are in this expectation $(1/\phi^2)$. Having established the initial tempo, the tempo evolves according to

$$T_n = T_{n-1} + \delta_n$$

for $n = 2, 3, \ldots, N$ where $\delta_n \sim N(0, \tau^2(S_{n-1}, S_n))$. When τ^2 takes on relatively small values, this "random walk" model captures the property that the tempo tends to vary smoothly. Note that our model assumes that the variance of $T_n - T_{n-1}$ depends on the transition S_{n-1}, S_n. In particular, longer notes will be associated with greater variability of tempo change.

Finally we assume that the observed note lengths $y_n = o_n - o_{n-1}$ for $n = 1, 2, \ldots, N$ are approximated by the product of the length of the note, $l(S_{n-1}, S_n)$, (measures) and local tempo, T_n, (secs. per measure). Specifically

$$Y_n = l(S_{n-1}, S_n)T_n + \epsilon_n$$

where

$$\epsilon_n \sim N(0, \rho^2(S_{n-1}, S_n)) \qquad (2)$$

Our model indicates that the observation variance depends on the note transition. In particular, longer notes should be associated with greater variance.

These modeling assumptions lead to a graphical model whose directed acyclic graph is given in Figure 2. In the figure each of the variables $S_0, \ldots, S_N$, $T_1, \ldots, T_n$, and $Y_1, \ldots, Y_N$ is associated with a node in the graph. The connectivity of the graph describes the dependency structure of the variables and can be interpreted as follows. The conditional distribution of a variable given all ancestors ("upstream" variables in the graph) depends only on the immediate parents of the variable. Thus the model is a particular example of a Bayesian network [16], [17], [18], [19]. Exploiting the connectivity structure of the graph is the key to successful computing in such models. Our particular model is composed of both discrete and Gaussian variables with the property that, for every configuration of discrete variables, the continuous variables have multivariate Gaussian distribution. Thus, the $S_0, \ldots, S_N$, $T_1, \ldots, T_N, Y_1, \ldots, Y_N$ collectively have a conditional Gaussian (CG) distribution [20], [21], [22], [23].

3 Finding the Optimal Rhythmic Parse

Recall that by "rhythmic parse" we mean a simultaneous estimate of the unobserved rhythm and tempo variables $S_0, \ldots, S_N$ and $T_1, \ldots, T_N$ given observed IOI data $Y_1 = y_1, \ldots, Y_n = y_N$. In view of our probabilistic formulation of the interaction between rhythm, tempo and observables, it seems natural to seek the *most likely* configuration of rhythm and tempo variables given the observed data, i.e. the *maximum a posteriori* (MAP) estimate. Thus, using the notation $a_i^j = (a_i, \ldots, a_j)$ where a is any vector, we let $f(s_0^N, t_1^N, y_1^N)$ be the joint probability density of the rhythm, tempo and observable variables. This joint density can be computed directly from the modeling assumptions of Section 2 as

$$\begin{aligned} f(s_0^N, t_1^N, y_1^N) &= p(s_0) \prod_{n=1}^{N} p(s_n|s_{n-1}) \\ &\times p(t_1) \prod_{n=2}^{N} p(t_n|s_{n-1}, s_n, t_{n-1}) \\ &\times \prod_{n=1}^{N} p(y_n|s_{n-1}, s_n, t_n) \end{aligned}$$

where $p(s_0)$ is the initial distribution for the rhythm process, $p(s_n|s_{n-1}) = R(s_{n-1}, s_n)$ is probability of moving from measure position s_{n-1} to s_n, $p(t_1)$ is the univariate normal density for the initial distribution

of the tempo process, $p(t_n|s_{n-1}, s_n, t_{n-1})$ is the conditional distribution of t_n given t_{n-1} whose parameters depend on s_{n-1}, s_n, and $p(y_n|s_{n-1}, s_n, t_n)$ is the the conditional distribution of y_n given t_n whose parameters also depend s_{n-1}, s_n. The rhythmic parse we seek is then defined by

$$\hat{s}_0^N, \hat{t}_1^N = \arg \max_{s_0^N, t_1^N} f(s_0^N, t_1^N, y_1^N)$$

where the observed IOI sequence, y_1^N, is fixed in the above maximization.

This maximization problem is ideally suited to dynamic programming due to the linear nature of the graph of Figure 2 describing the joint distribution of the model variables. Let $f_n(s_0^n, t_1^n, y_1^n)$ be the joint probability density of the variables S_0^n, T_1^n, Y_1^n (i.e. *up to* observation n) for $n = 1, 2, \ldots, N$. If we define $H_n(s_n, t_n)$ to be the density of the optimal configuration of unobservable variables ending in s_n, t_n:

$$H_n(s_n, t_n) \stackrel{\text{def}}{=} \max_{s_0^{n-1}, t_1^{n-1}} f_n(s_0^n, t_1^n, y_1^n)$$

then $H_n(s_n, t_n)$ can be computed through the recursion

$$\begin{aligned} H_1(s_1, t_1) &= \max_{s_0} p(s_0)p(s_1|s_0)p(t_1)p(y_1|s_0, s_1, t_1) \\ H_n(s_n, t_n) &= \max_{s_{n-1}, t_{n-1}} H_{n-1}(s_{n-1}, t_{n-1}) \\ &\times\ p(s_n|s_{n-1}) \\ &\times\ p(t_n|t_{n-1}, s_{n-1}, s_n) \\ &\times\ p(y_n|s_{n-1}, s_n, t_n) \end{aligned}$$

for $n = 2, \ldots, N$. Having computed H_n for $n = 1, \ldots, N$ we see that

$$\max_{s_N, t_N} H_N(s_N, t_N) = \max_{s_0^N, t_1^N} f(s_0^N, t_1^N, y_1^N)$$

is the most likely value we seek.

When all variables involved are discrete, it is a simple matter to perform this dynamic programming recursion and to traceback the optimal value value to recover the *globally* optimal sequence $\hat{s}_0^N, \hat{t}_1^N$. However, the situation is complicated in our case due to the fact that the tempo variables are continuous. We have developed methodology specifically to handle this important case, however a presentation of this methodology takes us too far afield. A general description of a strategy for computing the global MAP estimate of unobserved variables, given observed variables, in conditional Gaussian distributions (such as our rhythmic parsing example), can be found in [24].

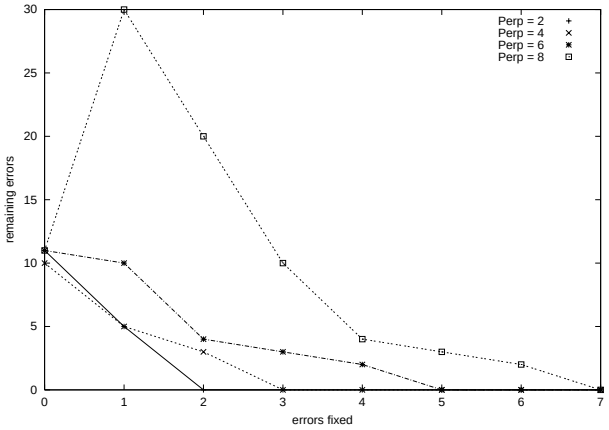

Figure 3: The number of errors produced by our system at different perplexities and with different numbers of errors already corrected.

4 Experiments

We performed several experiments using two different data sets. The first data set is a performance of the first section of Schumann's *2nd Romance for Oboe and Piano* (oboe part only), an excerpt of which is depicted in Figure 1. The original data, which can be heard at http://fafner.math.umass.edu/rhythmic_parsing, is a sampled audio signal, hence inappropriate for our experiments. Instead, we extracted a sequence of 129 note onset times from the data using the HMM methodology described in [25]. These data are also available at the above web page. In the performance of this excerpt, the tempo changes quite freely, thereby necessitating simultaneous estimation of rhythm and tempo.

Since the musical score for this excerpt was available, we extracted the complete set of possible measure positions,

$$\mathcal{S} = \left\{ \frac{0}{1}, \frac{1}{8}, \frac{1}{4}, \frac{1}{3}, \frac{3}{8}, \frac{5}{12}, \frac{15}{32}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8} \right\}$$

(The position 15/32 corresponds to a grace note which we have modeled as a 32nd note coming before the 3rd beat in 4/4 time). The most crucial parameters in our model are those that compose the transition probability matrix R. The two most extreme choices for R are the uniform transition probability matrix

$$R^{\text{unif}}(s_i, s_j) = 1/|\mathcal{S}|$$

and the matrix ideally suited to our particular recognition experiment

$$R^{\text{ideal}}(s_i, s_j) = \frac{|\{n : S_n = s_i, S_{n+1} = s_j\}|}{|\{n : S_n = s_i\}|}$$

R^{ideal} is unrealistically favorable to our experiments since this choice of R is optimal for recognition purposes and incorporates information normally unavailable; R^{unif} is unrealistically pessimistic in employing no prior information whatsoever. The actual transition probability matrices used in our experiments were convex combinations of these two extremes

$$R = \alpha R^{\text{ideal}} + (1 - \alpha) R^{\text{unif}}$$

for various constants $0 < \alpha < 1$. A more intuitive description of the effect of a particular α value is the *perplexity* of the matrix it produces: $\text{Perp}(R) = 2^{H(R)}$ where $H(R)$ is the $\log_2$ entropy of the corresponding Markov chain. Roughly speaking, if a transition probability matrix has perplexity M, the corresponding Markov chain has the same amount of "indeterminacy" as one that chooses randomly from M equally likely possible successors for each state. The extreme transition probability matrices have

$$\begin{aligned} \text{Perp}(R^{\text{ideal}}) &= 1.92 \\ \text{Perp}(R^{\text{unif}}) &= 11 = |\mathcal{S}| \end{aligned}$$

In all experiments we chose our initial distribution, $p(s_0)$, to be uniform, thereby assuming that all starting measure positions are equally likely. The remaining constants, $\nu, \phi^2, \tau^2, \rho^2$ were chosen to be values that seemed "reasonable."

The rhythmic parsing problem we pose here is based solely on timing information. Even with the aid of pitch and interpretive nuance, trained musicians occasionally have difficulty parsing rhythms. For this reason, it is not terribly surprising that our parses contained errors. However, a virtue of our approach is that the parses can be incrementally improved by allowing the user to correct individual errors. These corrections are treated as constrained variables in subsequent passes through the recognition algorithm. Due to the global nature of our recognition strategy, correcting a single error often fixes others parse errors automatically. Such a technique may well be useful in a more sophisticated music recognition system in which it is unrealistic to hope to achieve the necessary degree of accuracy without the aid of a human guide. In Figure 3 we show the number of errors produced under various experimental conditions. The four traces in the plot correspond to perplexities $2, 4, 6, 8$, while each individual trace gives the number of errors produced by the recognition after correcting $0, \ldots, 7$ errors. In each pass the first error found from the previous pass was corrected. In each case we were able to achieve a perfect parse after correcting 7 or fewer errors. Figure 3 also demonstrates that recognition accuracy improves with decreasing perplexity, thus showing that significant benefit results from using a transition probability matrix well-suited to the actual test data.

In our next, and considerably more ambitious, example we parsed a MIDI performance of the Chopin Mazurka Op. 6, no. 3. for solo piano. Unlike the monophonic instrument of the previous example, the piano can play several notes at a single score position. This situation can be handled with a very simple modification of the approach we have described above. Recall from Section 2 that $l(s, s')$ describes the note length associated with the transition from state s to state s'. We modify the definition of Eqn. 1 to be

$$l(s, s') = \begin{cases} s' - s & \text{if } s' \geq s \\ 1 + s' - s & \text{otherwise} \end{cases}$$

where we have simply replaced the $>$ in Eqn. 1 by $\geq$. The effect is that a "self-transition" (from state s to state s) is interpreted having 0 length, i.e. corresponding to two notes having the same score position.

For this example, in 3/4 time, we took the possible measure positions from the actual score, giving

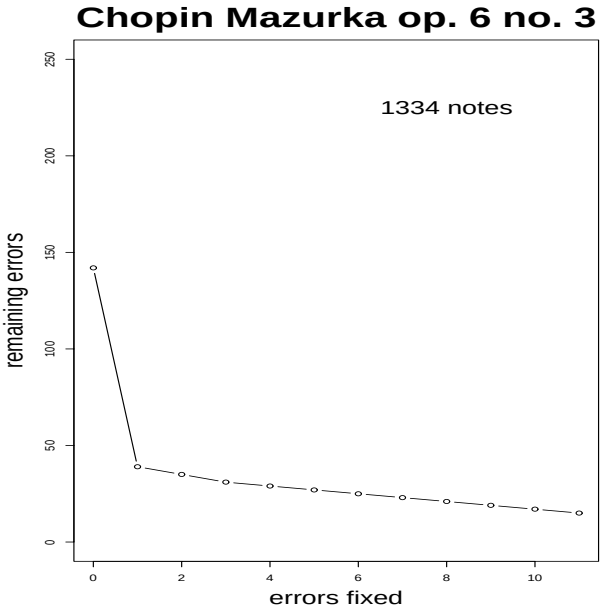

Figure 4: Results of rhythmic parses of Chopin Mazurka Op. 6, No. 3.

the set

$$\mathcal{S} = \left\{ \frac{0}{1}, \frac{1}{3}, \frac{2}{3}, \frac{1}{6}, \frac{11}{12}, \frac{23}{24}, \frac{1}{4}, \frac{1}{9}, \frac{2}{9}, \frac{1}{2}, \frac{5}{6}, \frac{1}{12}, \frac{13}{24}, \frac{7}{12}, \frac{1}{24} \right\}$$

Again, several of the measure positions correspond to grace notes. Rather than fixing the parameters of our model by hand, we instead estimated them from actual data. The transition probability matrix, R, was estimated from scores of several different Chopin Mazurka extracted from MIDI files. The result was a transition probability matrix having $\text{Perp}(R) = 2.02$, thereby providing a model that has enormously improved predictive power over the uniform transition model having perplexity $\text{Perp}(R) = |\mathcal{S}| = 15$. We also learned the variances of our model, $\tau^2(S_{n-1}, S_n)$ and $\rho^2(S_{n-1}, S_n)$ by applying the EM algorithm to a MIDI Mazurka using a known score.

We then iterated the procedure of parsing the data and then fixing the error beginning the longest run of consecutive errors. The results of our experiments with this data set are shown in Figure 4. The example contained 1334 notes. The MIDI file can be heard at http://fafner.math.umass.edu/rhythmic_parsing.

5 Discussion

We have presented a method for simultaneous estimation of rhythm and tempo, given a sequence of note onset times. Our method assumes that the collection of possible measure positions is given in advance. We believe this assumption is a relatively simple way of limiting the complexity of the recognized rhythm produced by the algorithm. When arbitrary rhythmic complexity is allowed without penalty, one can always find a rhythm with an arbitrarily accurate match to the observed time sequence. Thus, we expect that any approach to rhythm recognition will need some form of information that limits or penalizes this complexity. Other than this assumption, all parameters of our model can, and should, be learned from actual data, as in our second example. Such estimation requires a set of training data that "matches" the test data to be recognized in terms of rhythmic content and rhythmic interpretation. For example, we would not expect successful results if we trained our model on Igor Stravinsky's *Le Sacre du Printemps* and recognized on Hank Williams' *Your Cheatin' Heart*. In our experiments with the Chopin Mazurka in Section 4, we used different Chopin Mazurkas for training; however, it is likely that a less precise match between training and test would still prove workable.

We believe that the basic ideas we have presented can be extended significantly beyond what we have described. We are currently experimenting with a model that represents simultaneous evolution of rhythm *and pitch*. Since these quantities are intimately intertwined, one would expect better recognition of rhythm when pitch is given, as in MIDI data. For instance, consider the commonly encountered situation in which downbeats are often marked by low notes as in the Chopin example.

The experiments presented here deal with estimating the *composite* rhythm obtained by superimposing the various parts on one another. A disadvantage of this approach is that composite rhythms can be quite complicated even when the individual voices have simple repetitive rhythmic structure. For instance, consider a case in which one voice uses triple subdivisions while another use duple subdivisions. A more sophisticated project we are exploring is the simultaneous estimation of rhythm, tempo

and voicing. Our hope is that rhythmic structure becomes simpler and easier to recognize when one models and recognizes rhythm as the superposition of several rhythmic sources. Rhythm and voicing collective constitute the "lion's share" of what one needs for for automatic transcription of MIDI data.

While the Schumann example was much simpler than the Chopin example, it illustrates another direction we will pursue. Rhythmic parsing can play an important roll in interpreting the results of a preliminary analysis of audio data that converts a sampled acoustic signal into a "piano roll" type of representation. As discussed, we favor simultaneous estimation over "staged" estimation whenever possible, but we feel that an effort to simultaneously recover all parameters of interest from an acoustic signal is extremely ambitious, to say the least. We feel that the two problems of "signal-to-piano-roll" and rhythmic parsing together constitute a reasonable partition of the problem into manageable pieces. We intend to consider the transcription of audio data for considerably more complex data than those discussed here.

References

[1] Hewlett W., (1992), "A Base-40 Number-Line Representation of Musical Pitch Notation," *Musikometrika* Vol. 4, 1–14, 1992.

[2] Hewlett W., (1987), "The Representation of Musical Information in Machine-Readable Format," *Directory of Computer Assisted Research in Musicology*, Vol. 3, 1–22 1987.

[3] Selfridge-Field E., (1994), "The MuseData Universe: A System of Musical Information," *Computing in Musicology*, Vol. 9, 9–30, 1994.

[4] McNab R., Smith L., Bainbridge D., Witten I., (1997) "The New Zealand Digital Library MELody inDEX," D-Lib Magazine, http://www.dlib.org/dlib/may97/meldex/05witten.html May 1997.

[5] Bainbridge D. (1998), "MELDEX: A Web-based Melodic Index Search Service," *Computing in Musicology* Vol. 11 223–230, 1998.

[6] Schaffrath, H., (1992), "The EsAC Databases and MAPPET Software," *Computing and Musicology* vol. 8, 1992, 66.

[7] Desain P, Honing H., (1991) "Towards a calculus for expressive timing in music," *Computers in Music Research*, Vol. 3,43–120, 1991.

[8] Repp B., (1990), "Patterns of Expressive Timing In Performances of a Beethoven Minuet by Nineteen Famous Pianists," *Journal of the Acoustical Society of America* Vol. 88, 622–641, 1990.

[9] Bilmes J., (1993), "Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive music," S.M. thesis, Massachusetts Institute of Technology Media Lab, Cambridge, 1993.

[10] Trilsbeek P., van Thienen H., (1999), "Quantization for Notation: Methods used in Commercial Music Software," handout at *106th Audio Engineering Society conference*, May 1999, Munich.

[11] Cemgil A. T., Kappen B., Desain P., Honing, H. (2000), "On Tempo Tracking: Tempogram Representation and Kalman Filtering" *Proceedings of the International Computer Music Conference*, Berlin, 2000.

[12] Desain P., Honing H. (1994), "A Brief Introduction to Beat Induction," *Proceedings of the International Computer Music Conference*, San Francisco, 1994.

[13] Desain P., Honing H. (1989), "The Quantization of Musical Time: A Connectionist Approach," *Computer Music Journal*, Vol 13, no. 3.

[14] Desain P., Aarts R., Cemgil A. T., Kappen B., van Thienen H, Trilsbeek P. (1999), "Robust Time-Quantization for Music from Performance to Score," *Proceedings of 106th Audio Engineering Society conference*, May 1999, Munich.

[15] Cemgil A. T., Desain P., Kappen B. (1999), "Rhythm Quantization for Transcription," *Computer Music Journal*, 60-76.

[16] Lauritzen S. L., (1996), "Graphical Models," Oxford University Press, New York.

[17] Spiegelhalter D., Dawid A. P., Lauritzen S., Cowell R. (1993), "Bayesian Analysis in Expert Systems," *Statistical Science,* Vol. 8, No. 3, pp. 219–283.

[18] Jensen F., (1996), "An Introduction to Bayesian Networks," Springer-Verlag, New York.

[19] Cowell R., Dawid A. P., Lauritzen S., Spiegelhalter D. (1999), "Probabilistic Networks and Expert Systems," Springer, New York.

[20] Lauritzen S. L. and Wermuth N (1984), "Mixed Interaction Models," *Technical Report R-84-8*, Institute for Electronic Systems, Aalborg University.

[21] Lauritzen S. L. and Wermuth N (1989), "Graphical Models for Associations Between Variables, some of which are Qualitative and some Quantitative," *Annals of Statistics*, 17, 31-57.

[22] Lauritzen S. (1992), "Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models," *Journal of the American Statistical Association*, Vol. 87, No. 420, (Theory and Methods), pp. 1098–1108.

[23] Lauritzen S. L., Jensen F. (1999), "Stable Local Computation with Conditional Gaussian Distributions," *Technical Report R-99-2014*, Department of Mathematic Sciences, Aalborg University.

[24] Raphael C., (2001), "A Mixed Graphical Model for Rhythmic Parsing," *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence*, Seattle, 2001

[25] Raphael C., (1999), "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no 4, 360 – 370, 1999.