SEQUENTIAL INFERENCE OF RHYTHMIC STRUCTURE IN MUSICAL AUDIO

Nick Whiteley *A. Taylan Cemgil* *Simon Godsill*

University of Cambridge
Department of Engineering
Signal Processing and Communications Laboratory

ABSTRACT

This paper presents a framework for the modelling of temporal characteristics of musical signals and an approximate, sequential Monte Carlo inference scheme which yields estimates of tempo and rhythmic pattern from onset-time data. These two features are quantified through the construction of a probabilistic dynamical model of a hidden 'bar-pointer' and a Poisson observation model. The capabilities of the system are demonstrated by tracking the tempo of a 2 against 3 polyrhythm and detecting a switch in rhythm in a MIDI performance.

***Index Terms*—** Music, Statistics, Poisson distributions, Monte Carlo methods

1. INTRODUCTION

An important feature of intelligent music systems is the ability to infer attributes related to temporal structure. These attributes may include musicological constructs such as tempo and rhythmic pattern. The recognition of these characteristics forms a sub-task of automatic music transcription - the unsupervised generation of a score, or description of an audio signal in terms of musical concepts. For music categorization systems, tempo and rhythmic pattern are defining features of genre and therefore useful features for indexing of data sets.

Much work has been done on detecting the 'pulse' or foot-tapping rate of musical audio signals [1],[2]. However these approaches do not distinguish between tempo and rhythm. Goto and Muraoka detail a system which recognizes beats in terms of the 'reliability' of hypotheses for different rhythmic patterns [3]. Cemgil and Kappen model MIDI onset events in terms of a tempo process and switches between quantized score locations [4]. Raphael independently proposed a similar system [5]. Hainsworth and Macleod infer beats in a similar framework from raw audio signals [6], but rhythmic pattern is still not explicitly modelled.

Takeda et al. perform tempo and rhythm recognition from MIDI data by analogy with speech-recognition, but do not accommodate polyrhythms [7]. Klapuri et al. define metrical structure in terms of pulse sensations on different time scales, but do not explicitly discriminate between different rhythmic patterns [8].

In [9], a novel model of temporal structure in musical signals was introduced where exact inference was feasible. However, for certain extensions of the model, the exact inference scheme suffered from high computational requirements since it involved storage and manipulation of very large vectors.

In this paper we focus on the development of a practically scalable, sequential Monte Carlo inference scheme for a model of tempo and rhythmic pattern analogous to that in [9]. Development of such an inference scheme is challenging in this case due to the multi-modality of posterior probability distributions. In practical terms, this issue arises for the same reasons that human listeners can often 'explain' the same piece of music in terms of several different combinations of tempo and rhythmic pattern. Whilst the examples in this paper take as input MIDI onset data, the same framework could be used with onset times obtained from existing onset detection systems, e.g. [10].

In the Bayesian paradigm the task of joint estimation of tempo and rhythmic pattern is treated as an inference problem, where given a sequence of observations $y_{1:n} \equiv (y_1, y_2, ..., y_n)$ the aim is to compute posterior densities over the hidden state variables $\mathbf{x}_{0:n} \equiv (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_n)$. In a sequential setting we first postulate a Markovian prior density over the hidden state variables, $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$, which describes how the state variables evolve from one time index to the next. The observations are then related to the hidden state via $p(y_k|\mathbf{x}_k)$. Up to a constant of proportionality, the joint posterior density is given by:

$$p(\mathbf{x}_{0:n}|y_{1:n}) \propto p(\mathbf{x}_0) \prod_{k=1}^{n} p(y_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{x}_{k-1}) \qquad (1)$$

2. BAR-POINTER MODEL

The system is built around a dynamical model of a 'bar-pointer', a hypothetical, hidden object which maps an observed time-series to one period of a latent rhythmical pattern, i.e. one bar. At time $t_k = k\Delta$, $k \in \{1, 2, ..., n\}$ and Δ a constant, denote by $\phi_k \in [0, 1)$ the position of the bar-pointer and denote by $\dot{\phi}_k \in [\dot{\phi}_{min}, \dot{\phi}_{max}]$ its velocity, where $\dot{\phi}_{min} > 0$. The probabilistic kinematics of the bar-pointer are modelled as being a piece-wise constant velocity process:

$$\phi_{k+1} = (\phi_k + \Delta\dot{\phi}_k)\text{mod } 1 \tag{2}$$

$$p(\dot{\phi}_{k+1}|\dot{\phi}_k) \propto \mathcal{N}(\dot{\phi}_k, \sigma_{\phi}^2) \times \mathbb{I}_{\dot{\phi}_{min} \leq \dot{\phi}_{k+1} \leq \dot{\phi}_{max}} \tag{3}$$

where $\mathbb{I}_x$ is equal to 1 when x is true and zero otherwise. The velocity of the bar pointer is defined to be proportional to tempo.

A rhythmic pattern indicator, r_k, takes one value in a finite set, for example $r_k \in S = \{0, 1\}$, at each time index k. The elements of the set S correspond to different rhythmic patterns, which are described in section 3. For now we deal with the simple case in which there are only two such patterns, and switching between values of r_k is modelled as occurring if a bar line is crossed, i.e.:

if $\phi_k < \phi_{k-1}$,

$$p(r_k|r_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} p_r, & r_k \neq r_{k-1} \\ 1 - p_r, & r_k = r_{k-1} \end{cases} \tag{4}$$

otherwise, $r_k = r_{k-1}$, where p_r is the probability of a change in rhythmic pattern. In summary, $\mathbf{x}_k \equiv [\phi_k \ \dot{\phi}_k \ r_k]^T$ specifies the state of the system at time index k.

3. OBSERVATION MODEL

In this model, MIDI onset events are treated as being Poisson distributed with an intensity parameter which is conditioned on the position of the bar-pointer and the rhythm indicator variable. Defining the Poisson intensity in this fashion allows quantification of the postulate that for a given rhythm, there are regions in one bar in which onsets occur with high probability. This formalizes the onset time heuristics given in [11].

Each 'rhythmic pattern function', $\mu_r(\phi_k)$, maps the position of the bar pointer to the mean of a gamma prior distribution on an intensity parameter λ_k. For some ϕ_k, the value of $\mu_r(\phi_k)$ combined with a constant variance Q_λ, determines the shape and rate parameters of the gamma distribution:

$$a_r(\phi_k) = \mu_r(\phi_k)^2/Q_\lambda \tag{5}$$

$$b_r(\phi_k) = \mu_r(\phi_k)/Q_\lambda \tag{6}$$

For brevity, denote $a_k \equiv a_r(\phi_k)$, and $b_k \equiv b_r(\phi_k)$. Then conditional on ϕ_k and r_k, the prior density over λ_k is:

$$p(\lambda_k|\phi_k, r_k) = \begin{cases} \lambda_k^{a_k - 1} \frac{b_k^{a_k} \exp(-b_k \lambda)}{\Gamma(a_k)}, & \lambda_k \geq 0 \\ 0, & \lambda_k < 0 \end{cases} \tag{7}$$

This combination of prior distributions provides robustness against variation in the data. Examples of rhythmic pattern functions are given in figure 1.

Denote by y_k the number of onset events observed in the kth non-overlapping frame of length Δ, centred at time t_k. The number y_k is modelled as being distributed according to:

$$p(y_k|\lambda_k) = \frac{(\lambda_k \Delta)^{y_k} \exp(-\lambda_k \Delta)}{y_k!} \tag{8}$$

Inference of the intensity λ is not required so it is integrated out. This may be done analytically, yielding:

$$\begin{aligned} p(y_k|\phi_k, r_k) &= \int_0^\infty p(y_k|\lambda_k) p(\lambda_k|\phi_k, r_k) d\lambda_k \\ &= \frac{b_k^{a_k} \Gamma(a_k + y_k)}{y_k! \Gamma(a_k)(b_k + \Delta)^{a_k + y_k}} \end{aligned} \tag{9}$$

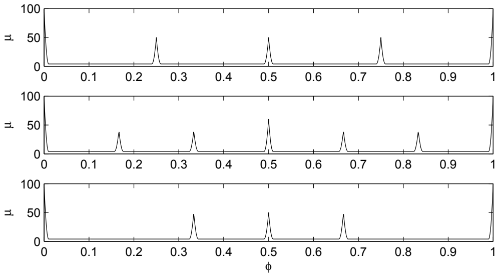

Fig. 1. Examples of rhythmic pattern functions, each corresponding to a different value of r_k. Top - a bar of duplets in 4/4 meter, middle - a bar of triplets in 4/4 meter, bottom - 2 against 3 polyrhythm. The widths of the peaks model arpeggiation of chords and expressive performance. Construction in terms of splines permits flat regions between peaks, corresponding to an onset event 'noise floor'.

4. INFERENCE SCHEME

4.1. Resample-Move Particle Filter

An analytical expression for the posterior density $p(\mathbf{x}_{0:k}|y_{1:k})$ is not available in the case of this model due to the intractability of the integral required to normalize the expression on the right of equation 1. An approximate, sampling-based inference scheme is therefore adopted.

Sequential Monte Carlo methods yield sample-based approximations to a sequence of probability distributions. The particle filter applies sequential importance sampling (SIS) to the Bayesian filtering problem [12]. The algorithm works by recursively extending and re-weighting N sampled state-trajectories ('particles') in order to construct approximations to the sequence of posterior densities:

$$p(\mathbf{x}_0), p(\mathbf{x}_{0:1}|y_1), p(\mathbf{x}_{0:2}|y_{1:2}), ..., p(\mathbf{x}_{0:n}|y_{1:n}) \tag{10}$$

Denoting by $w_k^{(i)}$ the weight of the ith particle $\mathbf{x}_{0:k}^{(i)}$ at time step k, the approximation to the posterior density is:

$$p(\mathbf{x}_{0:k}|y_{1:k}) \approx \sum_{i=1}^N w_k^{(i)} \delta_{\mathbf{x}_{0:k}^{(i)}}(\mathbf{x}_{0:k}) \tag{11}$$

From which approximations to the filtering densities $p(\mathbf{x}_k|y_{1:k})$ may be obtained.

After several iterations of an SIS algorithm, the particle system becomes *degenerate* - all but a small number of the particles have negligible weight. A resampling step is therefore employed, duplicating the heavily weighted particles and discarding the particles with small weight.

It was observed that at early time steps, the filtering distributions exhibit multiple modes corresponding to different bar pointer trajectories (for example multiples of the true tempo) which fit the observed data. By using the Metropolis Hastings (M-H) algorithm to apply Markov Chain Monte Carlo (MCMC) moves to the particles after resampling, it is possible in the case of this model to ensure that all significant modes of the posterior distribution are tracked. Technical details of resample-move schemes can be found in [13]. A mixture of velocity and position shift M-H proposals are used to ensure tempo diversity and to explore all phases of the rhythm. MCMC moves can be carried out with exponentially decreasing frequency, in order to reduce computational requirements.

The particle filtering algorithm incorporating the MCMC moves is given below.

```
for $k = 0$
    • for $i = 1$ to $N$
        – $\mathbf{x}_0^{(i)} \sim p(\mathbf{x}_0)$
        – $w_0^{(i)} = 1/N$
for $k = 1$ to $n$
    • for $i = 1$ to $N$
        – $\mathbf{x}_k^{(i)} \sim \pi(\mathbf{x}_k|\mathbf{x}_{0:k-1}^{(i)}, y_{1:k})$
        – $w_k^{(i)} \propto w_{k-1}^{(i)} \times \frac{p(y_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{\pi(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, y_{1:k})}$
    • for $i = 1$ to $N$
        – $w_k^{(i)} = \frac{\tilde{w}_k^{(i)}}{\sum_{j=1}^{N} \tilde{w}_k^{(j)}}$
    • for $i = 1$ to $N$
        – resample and set $w_k^{(i)} = 1/N$
        – if $y_k > 0$ apply velocity shift MCMC move
        – else apply position shift MCMC move
```

For this model, the optimal choice of the importance density $\pi(\mathbf{x}_k|\mathbf{x}_{0:k-1}^{(i)}, y_{1:k})$ is intractable and so the prior density $p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)})$ is used.

4.2. Monte-Carlo Smoothing

Backward simulation can be used to obtain approximate smoothed samples from $p(\mathbf{x}_{l:m}|y_{1:n})$, where $l \leq m \leq n$, using the weighted sample approximations to the filtering densities, $p(\mathbf{x}_k|y_{1:k})$ [14]. Smoothing is important in the case of this model because it yields correct alignment of changes in rhythm and corrects otherwise apparent deviations in tempo. The algorithm for backwards simulation is given below.

```
• choose $\tilde{\mathbf{x}}_n = \mathbf{x}_n^{(i)}$ with probability $w_n^{(i)}$
• for $k = n - 1$ to 0
    – for $i = 1$ to $N$,  calculate $w_{k|k+1}^{(i)} \propto w_k^{(i)} p(\tilde{\mathbf{x}}_{k+1}|\mathbf{x}_k^{(i)})$
    – choose $\tilde{\mathbf{x}}_k = \mathbf{x}_k^{(i)}$ with probability $w_{k|k+1}^{(i)}$
• $\tilde{\mathbf{x}}_{0:n}$ is an approximate realization from $p(\mathbf{x}_{0:n}|y_{1:n})$
```

5. RESULTS

5.1. Tracking a Polyrhythm

The ‘2 against 3’ polyrhythm simultaneously exhibits periodicity at two frequencies. This kind of rhythm could cause problems for simple beat trackers which are liable to ‘lock-on’ to one of these frequencies and ignore the other. A tempo-modulated performance was simulated and the frame-wise event counts - the observed data - can be seen at the top of figure 2. The particle filter was run on this data with the single rhythmic pattern function at the bottom of figure 1 and $N = 200$ particles. An initial prior distribution, $p(\mathbf{x}_0)$, was set to be uniform over all $(\phi, \dot{\phi}) \in [0, 1) \times [0.1, 2]$. The following parameter settings were used: $\Delta = 0.02$s, $\sigma_\phi^2 = 0.0005$, and $Q_\lambda = 10$. Figure 2 shows maximum a-posteriori (MAP) estimates for the bar-pointer position and tempo.

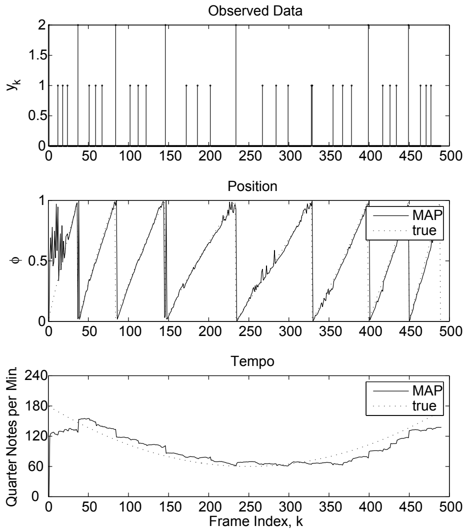

Fig. 2. Filtered position and tempo estimates for a simulated polyrhythm.

5.2. Recognizing a change in Rhythm

Figure 3 shows results using Monte Carlo smoothing for an excerpt of a MIDI performance of ‘Michelle’ by the Beatles. The performance, by a professional pianist, was recorded using a Yamaha Disklavier C3 Pro Grand Piano. The two top-most rhythmic patterns in figure 1 were used and a uniform initial prior distributions were set on ϕ_k, $\dot{\phi}_k$ and r_k, with $N = 600$ particles. The following parameter settings were used: $\Delta = 0.02$s, $\sigma_{\phi}^{2} = 0.0001$, $p_r = 0.5$ and $Q_{\lambda} = 10$.

This section of 'Michelle' is potentially problematic for tempo trackers because of the triplets, each of which by definition has a duration of 2/3 quarter notes. A performance of this excerpt could be wrongly interpreted as having a local change in tempo in the second bar, when really the rate of quarter notes remains constant; the bar of triplets is just a change in rhythm. Further results will later be made available on-line at http://www-sigproc.eng.cam.ac.uk/∼npw24/.

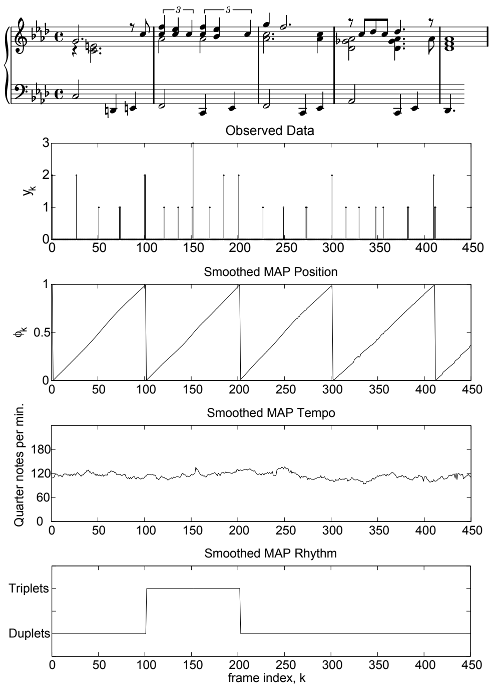

Fig. 3. Results of smoothing by backward simulation.

6. CONCLUSIONS

A model of temporal characteristics of music has been presented, along with an approximate inference scheme which yields filtered and smoothed estimates of tempo and rhythmic pattern. The inference scheme is scalable because it avoids handling large matrices. Demonstrations of the capabilities of the system were presented for two pieces, one involving a modulated polyrhythm and the other a switch in rhythm. The results show that the system handles such temporal variations which could defeat simple tempo trackers. Future work will address joint statistical modelling of high level temporal structure and raw audio signals.

7. REFERENCES

[1] E. Scheirer, "Tempo and beat analysis of acoustic music signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, 1998.

[2] W. A. Sethares, R. D. Morris, and J. C. Sethares, "Beat tracking of musical performances using low-level audio features," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, 2005.

[3] M. Goto and Y. Muraoka, "Music understanding at the beat level - real-time beat tracking of audio signals," in *Proc. of IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 1995.

[4] A. T. Cemgil and H. J. Kappen, "Monte carlo methods for tempo tracking and rhythm quantization," *Journal of Artificial Intelligence Research*, vol. 18, 2003.

[5] C. Raphael, "Automated rhythm transcription," in *Proc. of the 2nd Ann. Int. Symp. on Music Info. Retrieval.*, 2001.

[6] S. W. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 15, 2004.

[7] H. Takeda, T. Nishimoto, and S. Sagayama, "Rhythm and tempo recognition of music performance from a probabilistic approach," in *Proc. of the 5th Ann. Int. Symp. on Music Info. Retrieval*, 2004.

[8] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.

[9] N. Whiteley, A.T. Cemgil, and S. Godsill, "Bayesian modelling of temporal structure in musical audio," in *Proc. of the 7th International Conference on Music Information Retrieval*, 2006.

[10] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, 2005.

[11] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, 2001.

[12] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, 2000.

[13] W.R. Gilks and C. Berzuini, "Following a moving target - monte carlo inference for dynamic bayesian models," *J. R. Statist. Soc. B*, vol. 63, no. 1, 2001.

[14] S. Godsill A. Doucet and M. West, "Monte carlo smoothing for nonlinear timeseries," *Journal of the American Statistical Association*, vol. 99, no. 465, 2004.