Integrating Tempo Tracking and Quantization using Particle Filtering

Ali Taylan Cemgil, Bert Kappen
SNN, Dept. of Biophysics, University of Nijmegen, The Netherlands
email:cemgil@mbfys.kun.nl

Abstract

We present a probabilistic switching state space model for timing deviations in expressive music performance. We formulate tempo tracking and automatic transcription (rhythm quantization) as filtering and maximum a posteriori (MAP) state estimation tasks. The resulting model is suitable for real-time tempo tracking and transcription and hence useful in a number of music applications such as adaptive automatic accompaniment and score typesetting.

1 Introduction

Simultaneous estimation of the score *and* the tempo from onset times of an expressive performance is a mathematical "chicken-and-egg" problem. If the tempo is known, quantization, i.e. the association of onset times with discrete score locations is simpler. Similarly, if the score is given, the tempo can be estimated more easily.

However, if both tempo and the score are unknown, the problem becomes computationally intractable; there are simply too many alternative score-tempo pairs that may have given rise to an observed onset sequence. Due to this conceptual difficulty, most of research in the past has focused on quantization and tempo tracking separately.

Several models have been proposed to solve the rhythm quantization problem, e.g. (Longuet-Higgins 1987), (Desain and Honing 1991) (Cambouropoulos 2000), (Hamanaka et al. 2001). Researchers have also demonstrated sophisticated implementations (Pressing and Lawrence 1993), (Agon et al. 1994).

There is a significant body of research on the psychological and computational modeling aspects of tempo tracking. The work of (Large and Jones 1999), and its extensions by (Toiviainen 1999) describe a nonlinear adaptive oscillator model to human behavior in tracking the tempo. Attempts are also made to deal directly with the audio signal (Goto and Muraoka 1998), (Scheirer 1998), (Dixon and Cambouropoulos 2000).

Another class of tempo tracking models are developed in the context of interactive performance systems and score following. These models make use of prior knowledge in the form of an annotated score (Dannenberg 1984),(Vercoe and Puckette 1985). More recently, (Raphael 1999) has demonstrated an interactive real-time system that follows a solo player and schedules accompaniment events according to the players tempo interpretation. The system is based on a probabilistic model, hence can be trained to learn the soloists interpretation.

Our approach to transcription and tempo tracking is also from a probabilistic, i.e. Bayesian modeling perspective. In (Cemgil et al. 2000), we introduced a probabilistic approach to perceptually realistic quantization. This work assumed that the tempo was known or was estimated by an external procedure. For tempo tracking, we introduced a Kalman filter model (Cemgil et al. 2001). In this approach, we modeled the tempo as a smoothly varying hidden state variable of a stochastic dynamical system.

In the current paper, we integrate quantization and tempo tracking. Basically, our model balances score complexity versus smoothness in tempo deviations. The correct tempo interpretation results in a simple quantization and the correct quantization results in a smooth tempo fluctuation. Here, we give an outline of the main ideas, the theory is described in more detail in (Cemgil and Kappen 2002). A similar approach is proposed recently by (Raphael 2001) using a different and somewhat less flexible inference technique.

2 Bayes Theorem

The joint probability $p(X, Y)$ of two random variables Y and X defined over the respective state spaces S_Y and S_X can be factorized in two ways:

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X) \qquad (1)$$

where $p(Y|X)$ denotes the conditional probability of Y given X: for each value of X, this is a probability distribution over Y. The marginal distribution of a variable can be found from the joint distribution by summing over all states of the other variable, e.g.:

$$p(Y) = \sum_{X \in S_X} p(Y, X) = \sum_{X \in S_X} p(Y|X)p(X) \qquad (2)$$

It is understood that summation is to be replaced by integration if the state space is continuous. Bayes theorem results from Eq. 1 and Eq. 2 as:

$$p(X|Y) = \frac{p(Y|X)p(X)}{\sum_{X \in S_X} p(Y|X)p(X)} \qquad (3)$$

This rather simple looking "formula" has surprisingly far reaching consequences and can be directly applied to rhythm transcription and tempo tracking. To be more concrete, we introduce some mathematical notation. Let $Y = y_{1:K}$ denote onset times observed in an expressive performance. We use the abbreviation $y_{1:K}$ to denote the sequence $y_1, y_2 \ldots y_K$, i.e. there are K onsets. Let $X = \{c_{1:K}, z_{1:K}\}$ where c_k denotes the score position and z_k denotes the tempo at the k'th onset. Then Eq. 3 can be written as

$$p(c_{1:K}, z_{1:K}|y_{1:K}) = \frac{p(y_{1:K}|c_{1:K}, z_{1:K})p(c_{1:K}, z_{1:K})}{p(y_{1:K})}$$

The quantities in the numerator are called *likelihood* and *prior* respectively. Roughly, the prior distribution specifies our knowledge about the score simplicity and tempo smoothness. The likelihood term defines a model for short scale expressive timing deviations. In the following section we will define these models.

3 Model

We will consider the following generative model for a sequence of onset times obtain from an expressive music performance

$$c_k = c_{k-1} + \gamma_{k-1} \qquad (4)$$

$$\omega_k = \omega_{k-1} + \zeta_k \qquad (5)$$

$$\tau_k = \tau_{k-1} + 2^{\omega_k}(c_k - c_{k-1}) \qquad (6)$$

$$y_k = \tau_k + \epsilon_k \qquad (7)$$

In Eq. 4, c_k denotes the discrete grid location of k'th onset in a score. The interval between two consecutive note onsets is denoted by γ_{k-1}. For example consider the conventional music notation ♩ ♫ which encodes $\gamma_{1:3} = [1 \;\; 0.5 \;\; 0.5]$. The corresponding note onset sequence is $c_{1:4} = [0 \;\; 1 \;\; 1.5 \;\; 2]$. We assign a prior of form $p(c_k) \propto \exp(-\lambda d(c_k))$ where $d(c_k)$ is the number of significant digits in the binary expansion of the fraction of c_k (Cemgil et al. 2000) and λ is a positive parameter. One can check that such a prior prefers simpler notations, e.g. $p($ ♫♬ $) < p($ ♩ ♫ $)$.

Eq. 5 defines a distribution over possible tempo trajectories. We represent the tempo by the logarithm of its inverse (log-period) that we denote by ω as in (Cemgil et al. 2001). For example a tempo of 120 beats per minute (bpm) corresponds to $\omega = \log 60/120 \;\; \text{sec} = -1$. The tempo appears as a positive scale variable hence a representation in the logarithmic scale is quite natural. More precisely, we take the unknown tempo change ζ_k to be a Gaussian with $\mathcal{N}(0, \gamma_k Q)$[1]. Depending upon the interval between consecutive onsets, we scale the variance; longer jumps in the score allow for more tempo fluctuation.

Given the log-period sequence ω, Eq. 6 defines a model for "idealized" onset times that are only subject to tempo fluctuations. We can interpret τ_k as the ideal timing of an onset without any expressive timing or motor error. To simplify our notation, we will sometimes denote the pair (τ_k, ω_k) by z_k.

Eq. 7 defines the observation model. Here y_k is the actual observed onset time of the k'th onset in the performance. The noise term ϵ_k models small scale expressive deviations in timing of individual notes. In this paper we will assume that ϵ_k has a Gaussian distribution parameterized by $\mathcal{N}(0, R)$. A perceptually more plausible model for quantization is described in (Cemgil et al. 2000). The graphical model is shown in Figure 1.

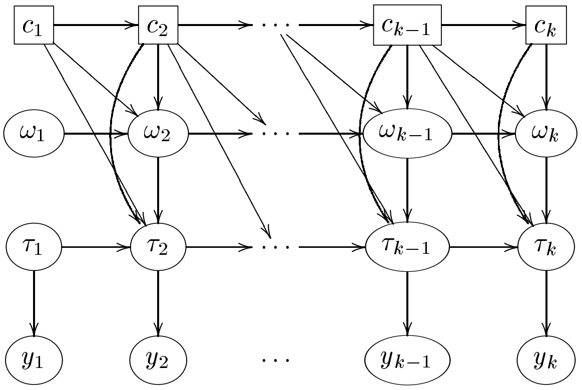

Figure 1: Graphical Model. The pair of continuous hidden variables (τ_k, ω_k) is denoted by z_k. Both c and z are hidden; only the onsets y are observed.

We define tempo tracking as a filtering problem

$$z_k^* = \underset{z_k}{\text{argmax}} \sum_{c_k} p(c_k, z_k|y_{1:k}) \qquad (8)$$

and rhythm transcription as a MAP state estimation problem

$$c_{1:K}^* = \underset{c_{1:K}}{\text{argmax}} p(c_{1:K}|y_{1:K}) \qquad (9)$$

$$p(c_{1:K}|y_{1:K}) = \int dz_{1:K} p(c_{1:K}, z_{1:K}|y_{1:K}) \qquad (10)$$

The quantities in Eq. 8 and Eq. 9 are intractable due to the explosion in the number of mixture components required to represent the exact posterior at each step k (See Figure 3). Consequently, we will reside to a numerical approximation technique called particle filtering.

[1] We denote a (scalar or multivariate) Gaussian distribution $p(\mathbf{x})$ with mean vector μ and covariance matrix P by $\mathcal{N}(\mu, P) \hat{=} |2\pi P|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^T P^{-1}(\mathbf{x} - \mu))$.

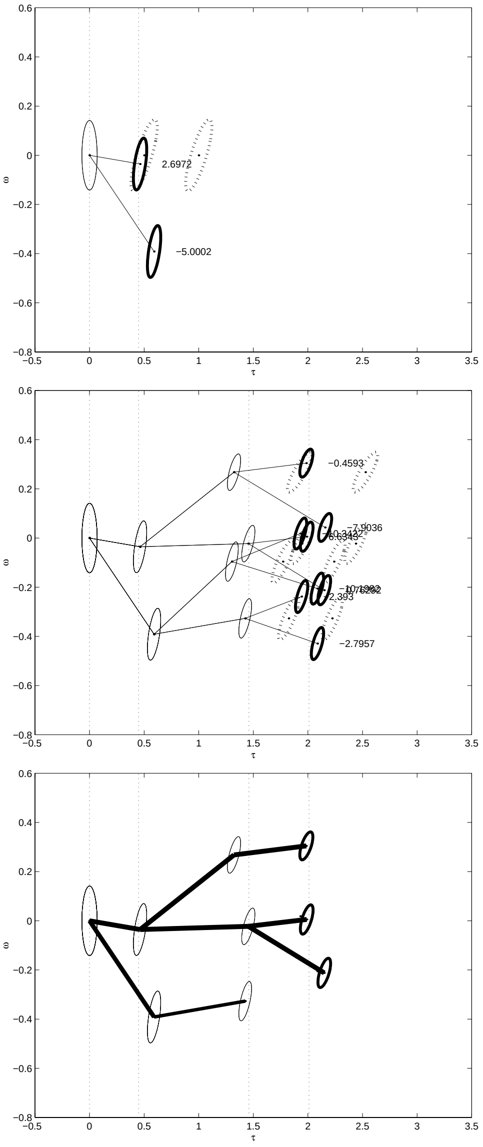

Figure 2: Example demonstrating the explosion of the number of components to represent the exact posterior. Ellipses denote the conditional marginals $p(\omega_k, \tau_k | c_{1:k}, y_{1:k})$. For clarity, we assume that a score consists only of notes of length ♪ and ♩, i.e. γ_k can be either $1/2$ or 1. (Above) We start with a unimodal posterior $p(\omega_1, \tau_1 | c_1, y_1)$, e.g. a Gaussian centered at $(\tau, \omega) = (0, 0)$. Since we assume that a score can only consist of eight- and quarter notes, the predictive distribution $p(\omega_2, \tau_2 | c_{1:2}, y_1)$ is bimodal where the modes are centered at $(0.5, 0)$ and $(1, 0)$ respectively (shown with a dashed contour line). Once the next observation y_2 is observed (shown with a dashed vertical line around $\tau = 0.5$), the predictive distribution is updated to yield $p(\omega_2, \tau_2 | c_{1:2}, y_{1:2})$. The numbers denote the respective log-posterior weight of each mixture component. (Middle) The number number of components to represent the exact posterior grows exponentially with k. (Bottom) Basic idea of particle. At each step k, particles with low likelihood are discarded. Surviving particles are linked to their parents. The method can be interpreted as a (stochastic) breadth first tree search procedure in the score-tempo space.

4 Particle Filtering

Particle filtering (a.k.a. Sequential Monte Carlo sampling) is an integration method especially powerful for inference in dynamical systems. Recently, it has been applied very successfully in a broad spectrum of applications in applied science, ranging from analysis of financial data to aircraft tracking and real time robotics. See (Doucet, de Freitas, and Gordon 2001) for a detailed review of state of the art.

A particle refers to a configuration of unobserved variables. In our model, each particle corresponds to a score and tempo level hypothesis that may have generated the data. More precisely, each particle is a marginal posterior distribution $\phi_k^{(i)} = p(z_k | c_{1:k}^{(i)})$, i.e. a score and corresponding mean and variance estimate of current tempo. The basic idea in particle filtering is to construct the particles at step k from particles at step $k-1$. The outline of the algorithm is as follows

1. *Generation*
   for each particle $\phi_{k-1}^{(i)}$ $i = 1, 2, \ldots, N$

   Find L candidate quantization locations for the observed onset y_k. Denote each candidate by $\hat{c}_k^{(l|i)}$ where $l = 1 \ldots L$.

2. *Evaluation*
   for each candidate $\hat{c}_k^{(l|i)}$ Evaluate the likelihood $w^{(l|i)} = p(y_k | \phi_{k-1}^{(i)})$. This is equivalent to one step Kalman filtering.

3. *Selection*
   Select N candidates from all candidates $\hat{c}_k^{(l|i)}$ generated according to their likelihood $w^{(l|i)}$. Update the tempo and score of surviving particles.

5 Simulation Results

We demonstrate tempo tracking and quantization performance of the model on two different examples. The first example is a repeating “son-clave” pattern

𝄆 𝄽 ♩ ♩ 𝄽 | ♩. ♩. ♩ 𝄇 ($c = [1 \quad 2 \quad 4 \quad 5.5 \quad 7 \ldots]$) with fluctuating tempo [2]. Such syncopated rhythms are usually hard to transcribe and make it difficult to track the tempo even for experienced human listeners. Moreover, since onsets are absent at prominent beat locations, standard beat tracking algorithms usually loose track.

We observe that for various realistic tempo fluctuations and observation noise level, the particle filter is able to identify the correct tempo trajectory and the corresponding quantization (Figure 3, above).

The second example is a piano arrangement of the Beatles song (Yesterday) performed by a professional classical pianist on a MIDI grand piano. Since the original arrangement is known, we estimate the true tempo

[2] We modulate the tempo deterministically according to $\omega_k = 0.3 \sin(2\pi c_k / 32)$. The observation noise variance is $R = 0.0005$.

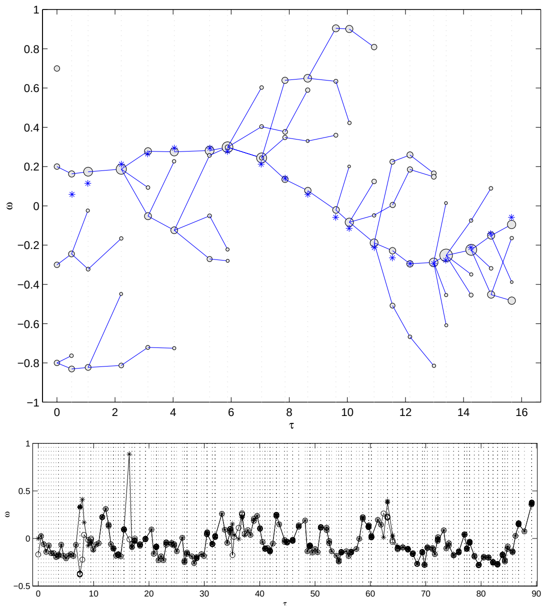

Figure 3: Above: Tempo tracking results for the clave pattern with 4 particles. Each circle denotes the mean $(\tau_k^{(i)}, \omega_k^{(i)})$. The diameter of each particle is proportional to the normalized importance weight at each generation. '*' denote the true (τ, ω) pairs. Below: Tracking results for "Yesterday". '*' denote the mean of the filtered $z_{1:K}$ after clamping to true $c_{1:K}$. Small circles denote the mean $z_{1:K}$ corresponding to the estimated MAP trajectory $c^*_{1:K}$ using 10 particles.

trajectory by Kalman filtering after clamping $c_{1:K}$. As shown in Figure 3, the particle filter estimate and the true tempo trajectory are almost identical.

6 Discussion and Conclusion

There are several advantages offered by particle filtering approach. The algorithm is suitable for real time implementation. Since the implementation is easy, this provides an important flexibility in the models one can employ. Although we have not addressed issues such as learning and online adaptation in this paper, parameters of the model can also treated as hidden variables.

Especially in real time music applications fine tuning and careful allocation of computational resources is of primary importance. Particle filtering is suitable since one can simply reduce the number of particles when computational resources become overloaded.

Motivated by the advantages of the particle filtering approach, we have implemented a prototype of our system that operates in real time. Consequently, the music is quantized such that it can be typeset in a notation program. We will eventually provide a short demonstration during the conference.

Acknowledgments: This research is supported by the Dutch Technology Foundation STW and the applied science division of NWO.

References

Agon, C., G. Assayag, J. Fineberg, and C. Rueda (1994). Kant: A critique of pure quantification. In *Proceedings of the ICMC*, Aarhus, Denmark, pp. 52–9.

Cambouropoulos, E. (2000). From midi to traditional musical notation. In *Proceedings of the AAAI Workshop on AI and Music*, Austin, Texas.

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

Cemgil, A. T. and H. Kappen (2002). Rhythm quantization and tempo tracking by sequential monte carlo. In *Advances in Neural Information Processing Systems 14*. MIT Press.

Cemgil, A. T., H. Kappen, P. Desain, and H. Honing (2001). On tempo tracking: Tempogram representation and kalman filtering. *Journal of New Music Research*.

Dannenberg, R. (1984). An on-line algorithm for real-time accompaniment. In *Proceedings of ICMC*, San Francisco, pp. 193–198.

Desain, P. and H. Honing (1991). The quantization of musical time: A connectionist approach. In P. M. Todd and D. G. Loy (Eds.), *Music and Connectionism*, pp. 150–67. Cambridge University Press: MIT Press.

Dixon, S. and E. Cambouropoulos (2000). Beat tracking with musical knowledge. In W. Horn (Ed.), *Proceedings of ECAI 2000*, Amsterdam.

Doucet, A., N. de Freitas, and N. J. Gordon (Eds.) (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Goto, M. and Y. Muraoka (1998). Music understanding at the beat level: Real-time beat tracking for audio signals. In *Computational Auditory Scene Analysis*.

Hamanaka, M., M. Goto, H. Asoh, and N. Otsu (2001). A learning-based quantization: Estimation of onset times in a musical score. In *Proceedings of (SCI 2001)*, Volume X, pp. 374–379.

Large, E. W. and M. R. Jones (1999). The dynamics of attending: How we track time-varying events. *Psychological Review 106*, 119–159.

Longuet-Higgins, H. C. (1987). *Mental Processes: Studies in Cognitive Science*. Cambridge: MIT Press. 424p.

Pressing, J. and P. Lawrence (1993). Transcribe: A comprehensive autotranscription program. In *Proceedings of the International Computer Music Conference*, Tokyo, pp. 343–345. Computer Music Association.

Raphael, C. (1999). A probabilistic expert system for automatic musical accompaniment. *Journal of Computational and Graphical Statistics*.

Raphael, C. (2001). A mixed graphical model for rhythmic parsing. In *Proc. of Uncertainty in AI*.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of Acoustical Society of America 103:1*, 588–601.

Toiviainen, P. (1999). An interactive midi accompanist. *Computer Music Journal 22:4*, 63–75.

Vercoe, B. and M. Puckette (1985). The synthetic rehearsal: Training the synthetic performer. In *Proceedings of ICMC*, San Francisco, pp. 275–278.