On tempo tracking: Tempogram Representation and Kalman filtering

Ali Taylan Cemgil*; Bert Kappen*; Peter Desain†; Henkjan Honing†
*SNN, Dept. of Medical Physics and Biophysics, University of Nijmegen, The Netherlands
†Music, Mind and Machine, University of Nijmegen,The Netherlands
email: {taylan,bert}@mbfys.kun.nl {desain,honing}@nici.kun.nl

Abstract

We formulate tempo tracking in a Bayesian framework where a tempo tracker is modeled as a stochastic dynamical system. The tempo is modeled as a hidden state variable of the system and is estimated from a MIDI performance by Kalman filtering and smoothing. We also introduce the Tempogram representation, a wavelet-like multiscale expansion of a real performance, on which the Kalman filter operates.

1 Introduction

An important and interesting subtask in automatic music transcription is tempo tracking: how to follow the tempo in a performance that contains expressive timing and tempo variations. When these tempo fluctuations are correctly identified it becomes much easier to separate the continuous expressive timing from the discrete note categories (i.e. quantization). The sense of tempo seems to be carried by the beats and thus tempo tracking is related to the study of beat induction, the perception of beats or pulse while listening to music (see Desain and Honing (1994)). However, it is still unclear what precisely constitutes tempo and how it relates to the perception of rhythmical structure. There is a significant body of research on the psychological and computational modeling aspects of tempo tracking. Early work by Michon (1967) describes a systematic study on the modeling of human behavior in tracking tempo fluctuations in artificially constructed stimuli. Longuet-Higgins (1976) proposes a musical parser that produces a metrical interpretation of performed music while tracking tempo changes. Knowledge about meter helps the tempo tracker to quantize a performance. Desain and Honing (1991) describe a connectionist model of quantization. Here as well, a tempo tracker helps to arrive at a correct rhythmical interpretation of a performance. Both models, however, have not been systematically tested. Still, quantizers can play a important role in addressing the difficult problem of what is a correct tempo interpretation by defining it as the one which results in a simpler quantization. Large and Jones (1999) describe an empirical study on tempo tracking, interpreting the observed human behavior in terms of an oscillator model.

Another class of models makes use of prior knowledge in the form of an annotated score (Dannenberg, 1984; Vercoe, 1984). They match the known score to incoming performance data. More recently attempts are made to deal directly with the audio signal (Goto and Muraoka, 1998; Scheirer, 1998) without using any prior knowledge. However, these models assume constant tempo (albeit timing fluctuations may be present), so are in fact not tempo trackers but beat trackers. Although successful for music with a steady beat, they report problems with syncopated data. All tempo track models assume an initial tempo (or beat length) to be known to start up the tempo tracking process (e.g., Longuet-Higgins (1976); Large and Jones (1999). There is few research addressing how to arrive at a reasonable first estimate. Longuet-Higgins and Lee (1982) propose a model based on score data, Scheirer (1998) one for audio data. A complete model should incorporate both aspects.

In this paper we formulate tempo tracking in a statistical framework where a tempo tracker is modeled as a stochastic dynamical system. The tempo is modeled as a hidden state variable of the system and is estimated by Kalman filtering.

2 Dynamical Systems and the Kalman Filter

Mathematically, a dynamical system is characterized by a set of *state variables* and a set of *state transition equations* that describe how state variables evolve with time. For example, a perfect metronome can be described as a dynamical system with two state variables: a phase $\hat{\tau}$ and a period $\hat{\Delta}$. Given the values of state variables at $j-1$'th step as $\hat{\tau}_{j-1}$ and $\hat{\Delta}_{j-1}$, the next beat occurs at $\hat{\tau}_j = \hat{\tau}_{j-1} + \hat{\Delta}_{j-1}$. The period is constant so $\hat{\Delta}_j = \hat{\Delta}_{j-1}$. By using vector notation and by letting $\mathbf{s}_j = [\hat{\tau}_j, \hat{\Delta}_j]^T$ we write the state transition model as

$$\mathbf{s}_j = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{s}_{j-1} = \mathbf{A}\mathbf{s}_{j-1} \quad (1)$$

When the initial state $\mathbf{s}_0 = [\hat{\tau}_0, \hat{\Delta}_0]^T$ is given, the system is fully specified.

Such a deterministic model is not realistic for natural music performance and can not be used for tracking the tempo in presence of tempo fluctuations and expressive timing deviations. Tempo fluctuations may be modeled by introducing a noise term that “corrupts” the state vector

$$\mathbf{s}_j \quad = \quad \mathbf{A}\mathbf{s}_{j-1} + \varepsilon \tag{2}$$

where ε is a Gaussian random vector with mean 0 and diagonal covariance matrix $\mathbf{Q}$, i.e. $\varepsilon \sim \mathcal{N}(0, \mathbf{Q})$. In addition, expressive timing deviations can be modeled by introducing a noise term

$$\tau_j \quad = \quad \hat{\tau}_j + \nu = \mathbf{C}\mathbf{s}_j + \nu \tag{3}$$

where $\nu \sim \mathcal{N}(0, \mathbf{R})$. Here, τ_j is the observed “noisy” beats. In this formulation, tempo tracking corresponds to estimation of $\mathbf{s}_j$ given observations upto j'th step. We note that we do not observe the (noisy) beat τ_j directly but *induce* it from events in music. This will be the topic of the next section.

Equations 2 and 3 define a *linear dynamical system*, because all noises are assumed to be Gaussian and all relationships between variables are linear. Hence, all state vectors $\mathbf{s}_j$ have Gaussian distributions. A Gaussian distribution is fully characterized by its mean and covariance matrix and in the context of linear dynamical systems, these quantities can be estimated very efficiently by a *Kalman filter*(Kalman, 1960). The operation of the filter is illustrated in Figure 1. The basic model can be extended in

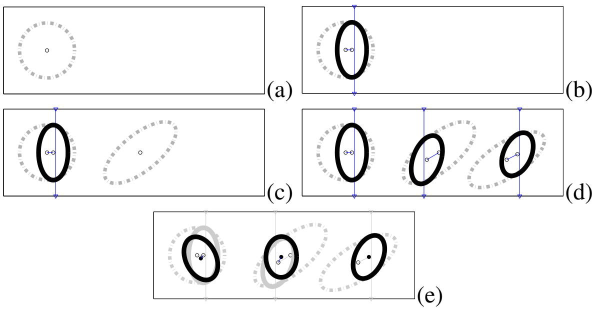

Figure 1: Operation of the Kalman Filter and Smoother. The horizontal axis represents the time and the vertical axis represent the period of the tracker. The system is given by $\mathbf{Q} = 0.01\mathbf{I}$ and $\mathbf{R} = 0.02$. $\mu_{j|j-1}$ and $P_{j|j-1}$ denote the mean (center) and covariance (ellipses) of the hidden state $\mathbf{s}_j$ given observations $\tau_1 \ldots \tau_{j-1}$. (a) The algorithm starts with the initial state estimate $(\mu_{1|0}, P_{1|0})$ at $\tau = 0$ and period $\Delta = 1$. in presence of no evidence, (b) [The beat is observed to be at τ_1, The state is updated to $(\mu_{1|1}, P_{1|1})$ according to the new evidence. Note that the uncertainty “shrinks”, (c) On the basis of current state a new prediction $(\mu_{2|1}, P_{2|1})$ is made, (d) Steps are repeated until all evidence is processed to obtain filtered estimates $(\mu_{j|j}, P_{j|j})$, $j = 1 \ldots N$. In this case $N = 3$. (e) Filtered estimates are updated by backtracking to obtain smoothed estimates $(\mu_{i|N}, P_{i|N})$ (Kalman smoothing).

several directions. First, the state space can be extended to include additional variables. Additional variables reduce

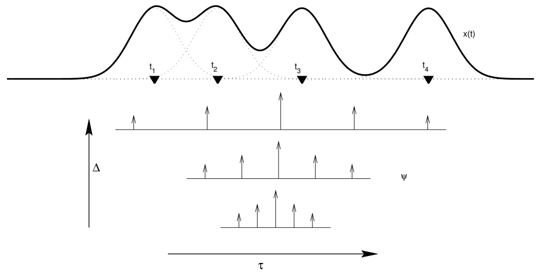

Figure 2: Tempogram Calculation. The continuous signal $x(t)$ is obtained from the onset list by convolution with a Gaussian function. Below, three different basis functions ψ are shown. All are localized at the same τ and different Δ. The tempogram at (τ, Δ) is calculated by taking the inner product of $x(t)$ and $\psi(t; \tau, \Delta)$. Due to the sparse nature of the basis functions, the inner product operation can be implemented very efficiently.

the random walk behavior since they introduce inertia to the system. The linearity constraint on the Kalman filter can also be relaxed. Indeed, in tempo tracking such a extension is necessary to ensure that the period $\hat{\Delta}$ is always positive. Therefore we define the state transition model in a warped space defined by the mapping $\omega = \log_2 \Delta$. This warping also ensures the perceptually more plausible assumption that tempo changes are relative rather than absolute. For example, under this warping, a deceleration from $\Delta \to 2\Delta$ has the same likelihood as an acceleration from $\Delta \to \Delta/2$.

3 Tempogram Representation

In this section, we propose a method to extract the noisy estimate τ_j from the performance. We demonstrate how a phase τ and period Δ can be inferred locally, i.e. from an short segment of an onset list $\mathbf{t} = [t_i]$. The Bayesian formulation of this problem is

$$p(\tau, \Delta|\mathbf{t}) \quad \propto \quad p(\mathbf{t}|\tau, \Delta)p(\tau, \Delta) \tag{4}$$

The likelihood term $p(\mathbf{t}|\tau, \Delta)$ is interpreted as the probability of the performance given the tempo track. $p(\tau, \Delta)$ is the prior probability of τ and Δ given by the Kalman filter. It is reasonable to assume that the likelihood $p(\mathbf{t}|\tau, \Delta)$ is high when onsets $[t_i]$ in the performance coincide with the beats of the tempo track. To construct a likelihood distribution having this property we propose a similarity measure between the performance and a local constant tempo track. First we define a a continuous time signal $x(t) = \sum_{i=1}^{I} G(t - t_i)$ where we take $G(t) = \exp(-t^2/2\sigma_x^2)$, a Gaussian function with variance σ_x^2. We represent a local tempo track as a pulse train $\psi(t; \tau, \Delta) = \sum_m \alpha_m \delta(t - \tau - m\Delta)$ where $\delta(t - t_0)$ is a translated Dirac delta function, which represents an impulse located at t_0. The coefficients α_m are positive constants such that $\sum_m \alpha_m = 1$ (See Figure 2). If a causal analysis is desired, α_m can be set to zero for $m > 0$. When α_m is a sequence decaying to zero exponentially, i.e. $\alpha_m = \alpha_0^m$, one has the infinite impulse response (IIR) comb filters employed by Scheirer (1998). We define the *tempogram* of $x(t)$ at (τ, Δ) as the inner product

$$\mathrm{Tg_x}(\tau, \Delta) = \int dt\, x(t)\psi(t; \tau, \Delta) \tag{5}$$

The tempogram representation can be interpreted as the response of a comb filter bank and is analogous to a multiscale representation (e.g. the wavelet transform), where τ and Δ correspond to transition and scaling parameters (Rioul and Vetterli, 1991). In Figure 3 we show a tempogram obtained from a simple onset sequence. We define the likelihood as $p(\mathbf{t}|\tau, \Delta) \propto \exp(\mathrm{Tg_x}(\tau, \Delta))$. The tempogram gives a local estimate of likely (τ, Δ) values.

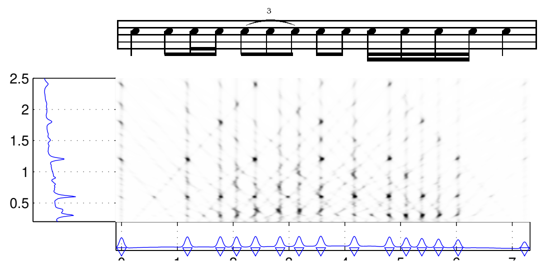

Figure 3: A simple rhythm and its Tempogram. x and y axes correspond to τ and Δ respectively. The bottom figure shows the onset sequence (triangles). Assuming flat priors on τ and Δ, the curve along the Δ axis is the marginal $p(\Delta|\mathbf{t}) \propto \int d\tau \exp(\mathrm{Tg_x}(\tau, \Delta))$. We note that $p(\Delta|\mathbf{t})$ has peaks at Δ, which correspond to quarter, eight and sixteenth note level as well as dotted quarter and half note levels of the original notation. This distribution can be used to estimate a reasonable initial state.

4 Evaluation

Many tempo trackers described in the introduction are often tested with ad hoc examples. However, to validate tempo tracking models, more systematic data and rigorous testing is necessary. A tempo tracker can be evaluated by systematically modulating the tempo of the data, for instance by applying instantaneous or gradual tempo changes and comparing the models responses to human behavior (Michon, 1967). Another approach is to evaluate tempo trackers on a systematically collected set of natural data, monitoring piano performances in which the use of expressive tempo change is free. This type of data has, next to being ecologically valid, the advantage of reflecting the type of data one expects automated music transcription systems to deal with. The latter approach was adopted in this study. For the experiment six pianists were invited to play arrangements of two Beatles songs, Michelle and Yesterday. Both pieces have a relatively simple rhythmic structure with ample opportunity to add expressiveness by fluctuating the tempo. The subjects consisted of one professional jazz player (PJ), four professional classical performers (PC) and one amateur classical pianist (AC). Each arrangement had to be played in three tempo conditions, three repetitions per tempo condition. The tempo conditions were normal, slow and fast tempo (all in a musically realistic range and all according to the judgement of the performer). We present here the results for these six subjects (6 subjects x 3 tempi x 3 repetitions x 2 pieces - 2 performances = 106 performances). The final data set will contain four pianists for each category (PJ, PC and AC). The performances were recorded on a Yamaha Disklavier Pro MIDI grand piano using Opcode Vision. To be able to derive tempo measurements related to the musical structure (e.g., beat, bar) the performances were matched with the MIDI scores using the structure matcher of Heijink et al. (2000) available in POCO (Honing, 1990). Tempo measurements were extracted for the notes that coincide with the beat (quarter note) level and the bar (whole note). In other words, we extract the (noisy) τ_j from the performance guided by the score.

4.1 Training the Kalman Filter

There are free parameters in the model, namely $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{C}$ and $\mathbf{R}$. In principle, all of these parameters can be estimated from data. Here, however, we restrict ourselves to the estimation of $\mathbf{A}$ and $\mathbf{Q}$ and set $\mathbf{C}$ and $\mathbf{R}$ to appropriate values. We divided the data set in a training set and a test set. We compute $\omega_j = \log_2(\tau_{j+1} - \tau_j)$ from the extracted tempotrack $[\tau_j]$ and learn a linear dynamics in the ω space by an EM algorithm (Ghahramani and Hinton, 1996). To find the appropriate filter order (Dimensionality of $\mathbf{s}$) we trained Kalman filters of orders from 1 to 6. We observed that a filter of order roughly between 1 and 4 is sufficient both in bar and beat levels. In any case, there is no large difference between models of different order.

4.2 Evaluation of tempo tracking performance

We evaluated the accuracy of the tempo tracking performance of the complete model with a Kalman filter of order one and a non-causal comb-filter tempogram with $\sigma_x = 0.04$ and $\alpha_0 = 0.4$. In the tracking experiments, we have initialized the filter to a reasonable estimate at beat level. For each performance in the data set, we obtain smoothed estimates of the beat $\hat{\tau}_j$. We compare $\hat{\tau}_j$ to the assumed true tempotrack τ_j as follows: for all j we check whether our beat estimate $\hat{\tau}_j$ is contained in the time window $\tau_j \pm 0.075$ sec. For each performance we calculate the percentage of correct beats. Figure 4 shows the results for the whole data set. Note that this is a quite “pessimistic” measure; if the tracker misses just one beat in the beginning but otherwise tracks the beat correctly, the correct beat percentage score would still be very low. Many of the poor quality tempo tracks are due to problems of this nature.

Naturally, the performance of the tracker depends on the amount of tempo variations introduced by the performer. For example, the tempo tracker fails consistently for subject PC2 who tends to use quite some tempo variation (Table 1). The performance is not very different among tempo conditions but somewhat better for normal tempo (Table 2).

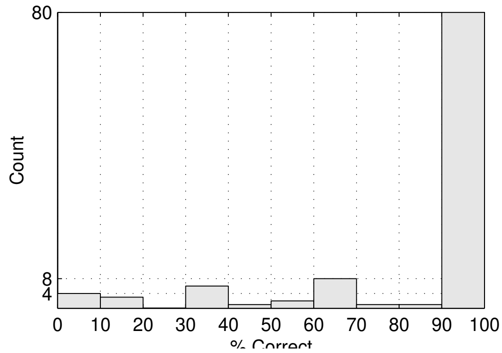

Figure 4: Histogram of correct beat percentage. 80 performances (of a total of 106) are tracked with an accuracy between %90-100.

5 Discussion and Future Research

In this paper, we have formulated a tempo tracking model in a Bayesian framework that incorporates a dynamical system and a measurement model. We employed a Kalman filter based dynamical system and a Tempogram based measurement model. In our view, many of the existing methods can be viewed as particular choices of a dynamical model and a measurement model. Bayesian formulation has several advantages: First, uncertainties can be integrated into the system in a natural way and desired quantities can be inferred in a consistent way. Moreover, prior knowledge (such as smoothness constraints in the state transition model and the particular choice of measurement model) are explicit and can be changed when needed. For example, the same state transition model can be used for both audio and MIDI; only the measurement model needs to be elaborated. For MIDI data, the Tempogram can also be replaced by a rhythm quantizer (Cemgil et al., 2000). Another advantage is that, for a large class of related models efficient inference and learning algorithms are well understood (Ghahramani and Hinton, 1996). This is appealing since we can train tempo trackers with different properties automatically from data. Online (filtering) or offline (smoothing) formulation is also possible. Online processing is necessary for real time applications such as automatic accompaniment and offline processing is desirable for transcription applications. The evaluation of the model on a systematicly collected set of natural data shows a high overall correctness. The next step will be an analysis of the local tempo behavior of the model (e.g., to test for its robustness once an error occurred) and characterize it in more qualitative terms (making use of the different musical conditions present in the full data set).

Subject:	PJ	AC	PC1	PC2	PC3	PC4
Yesterday	95.8	68.0	92.6	62.7	97.6	83.7
Michelle	96.6	98.5	98.5	44.9	75.5	93.5

Table 1: Correct beat percentage for subjects and pieces.

Condition:	fast	normal	slow
% Correct	82.7	88.1	80.5

Table 2: Correct beat percentage for tempo conditions.

Acknowledgments: This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs. We would like to thank Ric Ashley and Paul Trilsbeek for their contribution in the design and running of the experiment and we gratefully acknowledge the pianists from Northwestern University for their excellent performances.

References

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

Dannenberg, R.B. 1984. “An on-line algorithm for real-time accompaniment”. In *Proceedings of ICMC*, San Francisco. pages 193–198.

Desain, P. and Honing, H. 1991. “Quantization of musical time: a connectionist approach”. In Todd, P. M. and Loy, D. G., editors, *Music and Connectionism.*, pages 150–167. MIT Press., Cambridge, Mass.

Desain, P. and Honing, H. 1994. “A brief introduction to beat induction”. In *Proceedings of ICMC*, San Francisco.

Ghahramani, Zoubin and Hinton, Goeffrey E. “Parameter estimation for linear dynamical systems. (crg-tr-96-2)”. Technical report, University of Totronto. Dept. of Computer Science., 1996.

Goto, M. and Muraoka, Y. 1998. “Music understanding at the beat level: Real-time beat tracking for audio signals”. In Rosenthal, David F. and Okuno, Hiroshi G., editors, *Computational Auditory Scene Analysis*.

Heijink, H., Desain, P., and Honing, H. 2000. “Make me a match: An evaluation of different approaches to score-performance matching”. *Computer Music Journal*, 24(1):43–56.

Honing, H. 1990. “Poco: An environment for analysing, modifying, and generating expression in music.”. In *Proceedings of ICMC*, San Francisco. pages 364–368.

Kalman, R. E. 1960. “A new approach to linear filtering and prediction problems”. *Transaction of the ASME-Journal of Basic Engineering*, pages 35–45.

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

Longuet-Higgins, H. C. and Lee, C.S. 1982. “Perception of musical rhythms”. *Perception*.

Longuet-Higgins, H.C. 1976. “The perception of melodies”. *Nature*, 263:646–653.

Michon, J.A. 1967. “Timing in temporal tracking”. In *Soesterberg: RVO TNO*.

Rioul, Oliver and Vetterli, Martin. 1991. “Wavelets and signal processing”. *IEEE Signal Processing Magazine*, October:14–38.

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

Vercoe, B. 1984. “The synthetic performer in the context of live performance”. In *Proceedings of ICMC*, San Francisco. pages 199–200.