On tempo tracking: Tempogram Representation and Kalman filtering

Ali Taylan Cemgil[1]; Bert Kappen[1]; Peter Desain[2]; Henkjan Honing[2]
[1]SNN, Dept. of Medical Physics and Biophysics,
[2]Music, Mind and Machine group, NICI,
University of Nijmegen,The Netherlands
email: {taylan,bert}@mbfys.kun.nl {desain,honing}@nici.kun.nl

December 7, 2000

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 by a *Kalman filter*. The Kalman filter operates on a *Tempogram*, a wavelet-like multiscale expansion of a real performance. An important advantage of our approach is that it is possible to formulate both off-line or real-time algorithms. The simulation results on a systematically collected set of MIDI piano performances of Yesterday and Michelle by the Beatles shows accurate tracking of approximately $\%90$ of the beats.

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. Tempo is a perceptual construct and cannot directly be measured in a performance.

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; a relaxation network based on the principle of steering adjacent time intervals towards integer multiples. 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 on empirical data. 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 that results in a simpler quantization (Cemgil et al., 2000).

Large and Jones (1999) describe an empirical study on tempo tracking, interpreting the observed human behavior in terms of an oscillator model. A peculiar characteristic of this model is that it is insensitive (or becomes so after enough evidence is gathered) to material in between expected beats, suggesting that the perception tempo change is indifferent to events in this interval. Toiviainen (1999) discusses some problems regarding phase adaptation.

Another class of models make use of prior knowledge in the form of an annotated score (Dannenberg, 1984; Vercoe, 1984; Vercoe and Puckette, 1985). They match the known score to incoming performance data. Vercoe and Puckette (1985) uses a statistical learning algorithm to train the system with multiple performances. Even with this information at hand tempo tracking stays a non-trivial problem.

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 (e.g., popular music), they report problems with syncopated data (e.g., reggae or jazz music).

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 a tempo tracking in a probabilistic 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. The Kalman filter operates on a multiscale representation of a real performance which we call a Tempogram. In this respect the tempogram is analogous to a wavelet transform (Rioul and Vetterli, 1991). In the context of tempo tracking, wavelet analysis and related techniques are already investigated by various researchers (Smith, 1999; Todd, 1994). A similar comb filter basis is used by Scheirer (1998). The tempogram is also related to the periodicity transform proposed by Sethares and Staley (1999), but uses a time localized basis. Kalman filters are already applied in the music domain such as polyphonic pitch tracking (Sterian, 1999) and audio restoration (Godsill and Rayner, 1998). From the modeling point of view, the framework discussed in this paper has also some resemblance to the work of Sterian (1999), who views transcription as a model based segmentation of a time-frequency image.

The outline of the paper is as follows: We first consider the problem of tapping along a “noisy” metronome and introduce the Kalman filter and its extensions. Subsequently, we introduce the Tempogram representation to extract beats from performances and discuss the probabilistic interpretation. Consequently, we discuss parameter estimation issues from data. Finally we report simulation results of the system on a systematically collected data set, solo piano performances of two Beatles songs, Yesterday and Michelle.

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 beat $\hat{\tau}$ 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 of a perfect metronome 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 can write a linear state transition model as

$$\mathbf{s}_j \;=\; \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{s}_{j-1} = \mathbf{A}\mathbf{s}_{j-1} \tag{1}$$

When the initial state $\mathbf{s}_0 = [\hat{\tau}_0, \hat{\Delta}_0]^T$ is given, the system is fully specified. For example if the metronom clicks at a tempo 60 beats per minute ($\hat{\Delta}_0 = 1$ sec.) and first click occurs at time $\hat{\tau}_0 = 0$ sec., next beats occur at $\hat{\tau}_1 = 1$, $\hat{\tau}_2 = 2$ e.t.c. Since the metronom is perfect the period stays constant.

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

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

where $\mathbf{v}$ is a Gaussian random vector with mean 0 and diagonal covariance matrix $\mathbf{Q}$, i.e. $\mathbf{v} \sim \mathcal{N}(0, \mathbf{Q})$[1]. The tempo will drift from the initial tempo quickly if the variance of $\mathbf{v}$ is large. On the other hand when $\mathbf{Q} \to 0$, we have the constant tempo case.

In a music performance, the actual beat $\hat{\tau}$ and the period $\hat{\Delta}$ can not be observed directly. By actual beat we refer to the beat iterpretation that coincides with human perception when listening to music. For example, suppose, an expert drummer is tapping along a performance at the beat level and we assume her beats as the correct tempo track. If the task would be repeated on the same piece, we would observe each time a slightly different tempo track. As an alternative, suppose we would know the score of the performance and identify onsets that coincide with the beat. However, due to small scale expressive timing deviations, these onsets will be also noisy, i.e. we can at best observe "noisy" versions of actual beats. We will denote this noisy beat by τ in contrast to the actual but unobservable beat $\hat{\tau}$. Mathematically we have

$$\tau_j \;=\; \hat{\tau}_j + \mathbf{w}_j \tag{3}$$

where $\mathbf{w}_j \sim \mathcal{N}(0, \mathbf{R})$. Here, τ_j is the beat at step j that we get from a (noisy) observation process. In this formulation, tempo tracking corresponds to the estimation of hidden variables $\hat{\tau}_j$ given observations upto j'th step. We note that in a "blind" tempo tracking task, i.e. when the score is not known, the (noisy) beat

[1] A random vector $\mathbf{x}$ is said to be Gaussian with mean μ and covariance matrix $\mathbf{P}$ if it has the probability density

$$p(\mathbf{x}) = |2\pi\mathbf{P}|^{-1/2} \exp -\frac{1}{2}(x-\mu)^T\mathbf{P}^{-1}(x-\mu)$$

In this case we write $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{P})$

τ_j can not be directly observed since there is no expert drummer who is tapping along, neither a score to guide us. The noisy-beat itself has to be *induced* from events in the music. In the next section we will present a technique to estimate both a noisy beat τ_j as well a noisy period Δ_j from a real performance.

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; Roweis and Ghahramani, 1999). The operation of the filter is illustrated in Figure 1.

2.1 Extensions

The basic model can be extended in several directions. First, the linearity constraint on the Kalman filter can be relaxed. Indeed, in tempo tracking such an 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 $\Delta \rightarrow 2\Delta$ has the same likelihood as an acceleration from $\Delta \rightarrow \Delta/2$.

The state space $\mathbf{s}_j$ can be extended with additional dynamic variables $\hat{\mathbf{a}}_j$. Such additional variables store information about the past states (e.g. in terms of acceleration e.t.c.) and introduce inertia to the system. Inertia reduces the random walk behavior in the state space and renders smooth state trajectories more likely. Moreover, this can result in more accurate predictions.

The observation noise $\mathbf{w}_j$ can be modeled as a mixture of gaussians. This choice has the following rationale: To follow tempo fluctuations the observation noise variance $\mathbf{R}$ should not be too "broad". A broad noise covariance indicates that observations are not very reliable, so they have less effect to the state estimates. In the extreme case when $\mathbf{R} \rightarrow \infty$, all observations are practically missing so the observations have no effect on state estimates. On the other hand, a narrow $\mathbf{R}$ makes the filter sensitive to outliers since the same noise covariance is used regardless of the distance of an observation from its prediction. Outliers can be explicitely modeled by using a mixture of Gaussians, for example one "narrow" Gaussian for normal operation, and one "broad" Gaussian for outliers. Such a switching mechanism can be implemented by using a discrete variable c_j which indicates whether the j'th observation is an outlier or not. In other words we use a different noise covariance depending upon the value of c_j. Mathematically, we write this statement as $\mathbf{w}_j | c_j \sim \mathcal{N}(0, \mathbf{R_c})$. Since c_j can not be observed, we define a prior probability $c_j \sim p(c)$ and sum over all possible settings of c_j, i.e. $p(\mathbf{w}_j) = \sum_{c_j} p(c_j) p(\mathbf{w}_j | c_j)$. In Figure 2 we compare a switching Kalman filter and a standard Kalman filter. A switch variable makes a system more robust against outliers and consequently more realistic state estimates can be obtained. For a review of more general classes of switching Kalman filters see Murphy (1998).

To summerize, the dynamical model of the tempo tracker is given by

$$\hat{\tau}_j = \hat{\tau}_{j-1} + 2^{\hat{\omega}_{j-1}} \tag{4}$$

$$\begin{pmatrix} \hat{\omega}_j \\ \hat{\mathbf{a}}_j \end{pmatrix} = \mathbf{A} \begin{pmatrix} \hat{\omega}_{j-1} \\ \hat{\mathbf{a}}_{j-1} \end{pmatrix} + \mathbf{v}_j \tag{5}$$

$$\begin{pmatrix} \tau_j \\ \omega_j \end{pmatrix} = \begin{pmatrix} \hat{\tau}_j \\ \hat{\omega}_j \end{pmatrix} + \mathbf{w}_j \tag{6}$$

where $\mathbf{v}_j \sim \mathcal{N}(0, \mathbf{Q})$, $\mathbf{w}_j | c_j \sim \mathcal{N}(0, \mathbf{R_c})$ and $c_j \sim p(c_j)$. We take c_j as a binary discrete switch variable. Note that, in Eq. 6 the observable space is two dimensional (includes both τ and ω), in contrast to one dimensional observable τ in Figure 2.

3 Tempogram Representation

In the previous section, we have assumed that the beat τ_j is observed at each step j. In a real musical situation, however, the beat can not be observed directly from performance data. The sensation of a beat emerges from a *collection* of events rather than, say, single onsets. For example, a syncopated rhythm induces beats which do not neccesarly coincide with an onset.

In this section, we will define a probability distribution which assigns probability masses to all possible beat interpretations given a performance. The Bayesian formulation of this problem is

$$p(\tau, \omega | \mathbf{t}) \quad \propto \quad p(\mathbf{t} | \tau, \omega) p(\tau, \omega) \tag{7}$$

where $\mathbf{t}$ is an onset list. In this context, a *beat interpretation* is the tuple τ (local beat) and ω (local log-period).

The first term $p(\mathbf{t}|\tau, \omega)$ in Eq.7 is the probability of the onset list $\mathbf{t}$ given the tempo track. Since $\mathbf{t}$ is actually observed, $p(\mathbf{t}|\tau, \omega)$ is a function of τ and ω and is thus called the *likelihood* of τ and ω. The second term $p(\tau, \omega)$ in Eq.7 is the *prior* distribution. The prior can be viewed as a function which weights the likelihood on the (τ, ω) space. It is reasonable to assume that the likelihood $p(\mathbf{t}|\tau, \omega)$ is high when onsets $[t_i]$ in the performance coincide with the beats of the tempo track. To construct a likelihood function having this property we propose a similarity measure between the performance and a *local* constant tempo track. First we define 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 $\psi(t; \tau, \omega) = \sum_{m=-\infty}^{\infty} \alpha_m \delta(t - \tau - m2^{\omega})$ where $\delta(t - t_0)$ is a Dirac delta function, which represents an impulse located at t_0. The coefficients α_m are positive constants such that $\sum_m \alpha_m$ is a constant. (See Figure 3). In real-time applications, where causal analysis is desirable, α_m can be set to zero for $m > 0$. When α_m is a sequence of form $\alpha_m = \alpha^m$, where $0 < \alpha < 1$, one has the infinite impulse response (IIR) comb filters used by Scheirer (1998) which we adopt here. We define the *tempogram* of $x(t)$ at each (τ, ω) as the inner product

$$\mathrm{Tg_x}(\tau, \omega) = \int dt \; x(t) \psi(t; \tau, \omega) \tag{8}$$

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 ω correspond to transition and scaling parameters (Rioul and Vetterli, 1991; Kronland-Martinet, 1988).

The tempogram parameters have simple interpretations. The filter coefficient α adjust the time locality of basis functions. When $\alpha \to 1$, basis functions ψ extend to infinity and locality is lost. For $\alpha \to 0$

the basis degenerates to a single Dirac pulse and the tempogram is effectively equal to $x(t)$ for all ω and thus gives no information about the local period.

The variance parameter σ_x corresponds to the amount of small scale expressive deviation in an onsets timing. If σ_x would be large, the tempogram gets "smeared-out" and all beat interpretations become almost equally likely. When $\sigma_x \to 0$, we get a very "spiky" tempogram, where most beat interpretations have zero probability.

In Figure 4 we show a tempogram obtained from a simple onset sequence. We define the likelihood as $p(\mathbf{t}|\tau,\omega) \propto \exp(\mathrm{Tg_x}(\tau,\omega))$. When combined with the prior, the tempogram gives an estimate of likely beat interpretations (τ,ω).

4 Model Training

In this section, we review the techniques for parameter estimation. First, we summerize the relationships among variables by using a a *graphical model*. A graphical model is a directed acyclic graph, where nodes represent variables and missing directed links represent conditional independence relations. The distributions that we have specified so far are summarized in Table 1.

Model	Distribution	Parameters
State Transition (Eq. 5)	$p(\mathbf{s}_{j+1}\|\mathbf{s}_j)$	$\mathbf{A}$, $\mathbf{Q}$
(Switching) Observation (Eq. 6)	$p(\tau_j,\omega_j\|\mathbf{s}_j,c_j)$	$\mathbf{R}_c$
Switch prior (Eq. 6)	$p(c_j)$	p_c
Tempogram (Eq.8)	$p(\mathbf{t}\|\tau_j,\omega_j)$	σ_x, α

Table 1: Summary of conditional distributions and their parameters.

The resulting graphical model is shown in Figure 5. For example, the graphical model has a directed link from $\mathbf{s}_j$ to $\mathbf{s}_{j+1}$ to encode $p(\mathbf{s}_{j+1}|\mathbf{s}_j)$. Other links towards $\mathbf{s}_{j+1}$ are missing.

In principle, we could jointly optimize all model parameters. However, such an approach would be computationally very intensive. Instead, at the expense of getting a suboptimal solution, we will assume that we observe the noisy tempo track τ_j. This observation effectively "decouples" the model into two parts (See Fig. 5), (i) The Kalman Filter (State transition model and Observation (Switch) model) and (ii) Tempogram. We will train each part separately.

4.1 Estimation of τ_j from performance data

In our studies, a score is always available, so we extract τ_j from a performance $\mathbf{t}$ by matching the notes that coincide with the beat (quarter note) level and the bar (whole note). If there are more than one note on a beat, we take the median of the onset times. [2] For each performance, we compute $\omega_j = \log_2(\tau_{j+1} - \tau_j)$ from the extracted noisy beats $[\tau_j]$. We denote the resulting tempo track $\{\tau_1, \omega_1 \ldots \tau_j, \omega_j \ldots \tau_J, \omega_J\}$ as $\{\tau_{1:J}, \omega_{1:J}\}$.

[2] The scores do not have notes on each beat. We interpolate missing beats by using a switching Kalman filter with parameters $\mathbf{Q} = \mathrm{diag}([0.01^2, 0.05^2])$, $\mathbf{R}_1 = 0.01^2$, $\mathbf{R}_2 = 0.3^2$, $\mathbf{A} = 1$ and $p(c) = [0.999, 0.001]$.

4.2 Estimation of state transition parameters

We estimate the state transition model parameters $\mathbf{A}$ and $\mathbf{Q}$ by an EM algorithm (Ghahramani and Hinton, 1996) which learns a linear dynamics in the ω space. The EM algorithm monotonically increases $p(\{\tau_{1:J}, \omega_{1:J}\})$, i.e. the likelihood of the observed tempo track. Put another way, the parameters $\mathbf{A}$ and $\mathbf{Q}$ are adjusted in such a way that, at each j, the probability of the observation is maximized under the predictive distribution $p(\tau_j, \omega_j | \tau_{j-1}, \omega_{j-1}, \ldots \tau_1, \omega_1)$. The likelihood is simply the hight of the predictive distribution evaluated at the observation (See Figure 1).

4.3 Estimation of switch parameters

The observation model is a Gaussian mixture with diagonal $\mathbf{R}_c$ and prior probability p_c. We could estimate $\mathbf{R}_c$ and p_c jointly with the state transition parameters $\mathbf{A}$ and $\mathbf{Q}$. However, then the noise model would be totally independent from the tempogram representation. Instead, the observation noise model should reflect the uncertainty in the tempogram; for example the expected amount of deviations in (τ, ω) estimates due to spurious local maxima. To estimate the "tempogram noise" by standard EM methods, we sample from the tempogram around each $[\hat{\tau}_j, \hat{\omega}_j]$, i.e. we sample τ_j and ω_j from the posterior distribution $p(\tau_j, \omega_j | \hat{\tau}_j, \hat{\omega}_j, \mathbf{t}; \mathbf{Q}) \propto p(\mathbf{t} | \tau_j, \omega_j) p(\tau_j, \omega_j | \hat{\tau}_j, \hat{\omega}_j; \mathbf{Q})$. Note that $[\hat{\tau}_j, \hat{\omega}_j]$ are estimated during the E step of the EM algorithm when finding the parameters $\mathbf{A}$ and $\mathbf{Q}$.

4.4 Estimation of Tempogram parameters

We have already defined the tempogram as a likelihood $p(\mathbf{t} | \tau, \omega; \theta)$ where θ denotes the tempogram parameters (e.g. $\theta = \{\alpha, \sigma_x\}$). If we assume a uniform prior $p(\tau, \omega)$ then the posterior probability can be written as

$$p(\tau, \omega | \mathbf{t}; \theta) = \frac{p(\mathbf{t} | \tau, \omega; \theta)}{p(\mathbf{t} | \theta)} \tag{9}$$

where the normalization constant is given by $p(\mathbf{t} | \theta) = \int d\tau d\omega p(\mathbf{t} | \tau, \omega; \theta)$. Now, we can estimate tempogram parameters θ by a maximum likelihood approach. We write the log-likelihood of an observed tempo track $\{\tau_{1:J}, \omega_{1:J}\}$ as

$$\log p(\{\tau_{1:J}, \omega_{1:J}\} | \mathbf{t}; \theta) = \sum_j \log p(\tau_j, \omega_j | \mathbf{t}; \theta) \tag{10}$$

Note that the quantity in Equation 10 is a function of the parameters θ. If we have k tempo tracks in the dataset, the complete data log-likelihood is simply the sum of all individual log-likelihoods. i.e.

$$\mathcal{L} = \sum_k \log p(\{\tau_{1:J}, \omega_{1:J}\}^k | \mathbf{t}^k; \alpha, \sigma_x) \tag{11}$$

where $\mathbf{t}^k$ is the k'th performance and $\{\tau_{1:J}, \omega_{1:J}\}^k$ is the corresponding tempo track.

5 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; Dannenberg, 1993). 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 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.

5.1 Data

For the experiment 12 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 four professional jazz players (PJ), four professional classical performers (PC) and four amateur classical pianists (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 judgment of the performer). We present here the results for twelve subjects (12 subjects $\times$ 3 tempi $\times$ 3 repetitions $\times$ 2 pieces $=$ 216 performances). 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). This MIDI data, as well as related software will be made available at URL's `http://www.mbfys.kun.nl/~cemgil` and `http://www.nici.kun.nl/mmm` (under the heading Download).

5.2 Kalman Filter Training results

We use the performances of Michelle as the training set and Yesterday as the test set. To find the appropriate filter order (Dimensionality of $\mathbf{s}$) we trained Kalman filters of several orders on two rhythmic levels: the beat (quarter note) level and the bar (whole note) level. Figure 6 shows the training and testing results as a function of filter order.

Extending the filter order, i.e. increasing the the size of the state space loosely corresponds looking more into the past. At bar level, using higher order filters merely results in overfitting as indicated by decreasing test likelihood. In contrast, on the beat level, the likelihood on the test set also increases and has a jump around order of 7. Effectively, this order corresponds to a memory which can store state information from the past two bars. In other words, tempo fluctuations at beat level have some structure that a higher dimensional state transition model can make use of to produce more accurate predictions.

5.3 Tempogram Training Results

We use a tempogram model with a first order IIR comb basis. This choice leaves two free parameters that need to be estimated from data, namely α, the coefficient of the comb filter and σ_x, the width of the Gaussian window. We obtain optimal parameter values by maximization of the log-likelihood in Equation 11 on the Michelle dataset. The resulting likelihood surface is shown in Figure 7. The optimal parameters are shown in Table 2.

	α	σ_x
Non-Causal	0.55	0.017
Causal	0.73	0.023

Table 2: Optimal tempogram parameters.

5.4 Initialization

To have a fully automated tempo tracker, the initial state $\mathbf{s}_0$ has to be estimated from data as well. In the tracking experiments, we have initialized the filter to the beat level by computing a tempogram for the first 5 seconds of each performance. By assuming a flat prior on τ and ω we compute the posterior marginal $p(\omega|\mathbf{t}) = \int d\tau p(\omega, \tau|\mathbf{t})$. Note that this is operation is just equivalent to summation along the τ dimension of the tempogram (See Figure 4). For the Beatles dataset, we have observed that for all performances of a given piece, the most likely log-period $\omega^* = \arg\max_\omega p(\omega|\mathbf{t})$ corresponds always to the same level, i.e. the ω^* estimate was always consistent. For "Michelle", this level is the beat level and for "Yesterday" the half-beat (eighth note) level. The latter piece begins with an arpeggio of eight notes; based on onset information only, and without any other prior knowledge, half-beat level is also a reasonable solution. For "Yesterday", to test the tracking performance, we corrected the estimate to the beat level.

We could estimate τ^* using a similar procedure, however since all performances in our data set started "on the beat", we have chosen $\tau^* = t_1$, the first onset of the piece. All the other state variables $\hat{\mathbf{a}}_0$ are set to zero. We have chosen a broad initial state covariance $P_0 = 9\mathbf{Q}$.

5.5 Evaluation of tempo tracking performance

We evaluated the accuracy of the tempo tracking performance of the complete model. The accuracy of tempo tracking is measured by using the following criterion:

$$\rho(\psi, \mathbf{t}) = \frac{\sum_i \max_j W(\psi_i - t_j)}{(I+J)/2} \times 100$$

where $[\psi_i] \quad i = 1 \ldots I$ is the target (true) tempo track and $[t_j] \quad j = 1 \ldots J$ is the estimated tempo track. W is a window function. In the following results we have used a Gaussian window function $W(d) = \exp(-d^2/2\sigma_e^2)$. The width of the window is chosen as $\sigma_e = 0.04$ sec which corresponds roughly

to the spread of onsets from their mechanical means during performance of short rhythms (Cemgil et al., 2000).

It can be checked that $0 \leq \rho \leq 100$ and $\rho = 100$ if and only if $\psi = \mathbf{t}$. Intuitively, this measure is similar to a normalized inner-product (as in the tempogram calculation); the difference is in the $\max$ operator which merely avoids double counting. For example, if the target is $\psi = [0,\ 1,\ 2]$ and we have $\mathbf{t} = [0,\ 0,\ 0]$, the ordinary inner product would still give $\rho = 100$ while only one beat is correct $(t = 0)$. The proposed measure gives $\rho = 33$ in this case. The tracking index ρ can be roughly interpreted as percentage of "correct" beats. For example, $\rho = 90$ effectively means that about 90 percent of estimated beats are in the near vicinity of their targets.

5.6 Results

To test the relative relevance of model components, we designed an experiment where we evaluate the tempo tracking performance under different conditions. We have varied the filter order and enabled or disabled switching. For this purpose, we trained two filters, one with a large (10) and one with a small (2) state space dimension on beat level (using the Michelle dataset). We have tested each model with both causal and non-causal tempograms. To test whether a tempogram is at all necessary, we propose a simple onset-only measurement model. In this alternative model, the next observation is taken as the nearest onset to the Kalman filter prediction. In case there are no onsets in 1σ interval of the prediction, we declare the observation as missing (Note that this is an implicit switching mechanism).

In Table 3 we show the tracking results averaged over all performances in the Yesterday dataset. The estimated tempo tracks are obtained by using a non-causal tempogram and Kalman filtering. In this case, Kalman smoothed estimates are not significantly different. The results suggest, that for the Yesterday dataset, a higher order filter or a (binary) switching mechanism does not improve the tracking performance. However, presence of a tempogram makes the tracking performance both more accurate and consistent (note the lower standard deviations). As a "base line" performance criteria, we also compute the best constant tempo track (by a linear regression to estimated tempo tracks). In this case, the average tracking index obtained from a constant tempo approximation is rather poor ($\rho = 28 \pm 18$), confirming that there is indeed a need for tempo tracking.

Filter order	Switching	tempogram	no tempogram
10	+	92 ± 7	75 ± 21
2	+	91 ± 9	75 ± 21
10	-	91 ± 6	73 ± 21
2	-	90 ± 9	73 ± 22

Table 3: Average tracking performance ρ and standard deviations on Yesterday dataset using a non-causal tempogram. $+$ denotes the case when we have the switch prior $p(c) = [0.8,\ \ 0.2]$. $-$ denotes the absence of a switching, i.e. the case when $p(c) = [1,\ \ 0]$.

We have repeated the same experiment with a causal tempogram and computed the tracking performance for predicted, filtered and smoothed estimates In Table 4 we show the results for a switching

Kalman filter. The results without switching are not significantly different. As one would expect, the tracking index with predicted estimates is lower. In contrast to a non-causal tempogram, smoothing improves the tempo tracking and results in a comparable performance as a non-causal tempogram.

	causal		
Filter order	predicted	filtered	smoothed
10	74 ± 12	86 ± 9	91 ± 8
2	73 ± 12	85 ± 8	90 ± 8

Table 4: Average tracking performance ρ on Yesterday dataset. Figures indicate tracking index ρ followed by the standard deviation. The label "non-causal" refers to a tempogram calculated using non-causal comb filters. The labels predicted, filtered and smoothed refer to state estimates obtained by the Kalman filter/smoother.

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 a subject who tends to use quite some tempo variation[3].

We find that the tempo tracking performance is not significantly different among different groups (Table 5). However, when we consider the predictions, we see that the performances of professional classical pianists are less predictable. For different tempo conditions (Table 6) the results are also similar. As one would expect, for slower performances, the predictions are less accurate. This might have two potential reasons. First, the performance criteria ρ is independent of the absolute tempo, i.e. the window W is always fixed. Second, for slower performances there is more room for adding expression.

	non-causal	causal			
Subject Group	filtered	predicted	filtered	smoothed	Best const.
Prof. Jazz	95 ± 3	81 ± 7	92 ± 4	94 ± 3	34 ± 22
Amateur Classical	92 ± 8	74 ± 7	88 ± 5	92 ± 4	24 ± 19
Prof. Classical	89 ± 7	66 ± 14	82 ± 11	86 ± 11	27 ± 12

Table 5: Tracking Averages on subject groups. As a reference, the right most column shows the results obtained by the best constant tempo track. The label "non-causal" refers to a tempogram calculated using non-causal comb filters. The labels predicted, filtered and smoothed refer to state estimates obtained by the Kalman filter/smoother.

	non-causal	causal			
Condition	filtered	predicted	filtered	smoothed	Best const.
fast	94 ± 5	79 ± 9	90 ± 6	93 ± 6	39 ± 21
normal	92 ± 8	74 ± 9	88 ± 6	92 ± 4	25 ± 13
slow	90 ± 7	68 ± 14	84 ± 10	87 ± 11	21 ± 14

Table 6: Tracking Averages on tempo conditions. As a reference, the right most column shows the results obtained by the best constant tempo track. The label "non-causal" refers to a tempogram calculated using non-causal comb filters. The labels predicted, filtered and smoothed refer to state estimates obtained by the Kalman filter/smoother.

[3]This subject claimed to have never heard the Beatles songs before.

6 Discussion and Conclusions

In this paper, we have formulated a tempo tracking model in a probabilistic framework. The proposed model consist of a dynamical system (a Kalman Filter) and a measurement model (Tempogram). Although many of the methods proposed in the literature can be viewed as particular choices of a dynamical model and a measurement model, a Bayesian formulation exhibits several advantages in contrast to other models for tempo tracking. First, components in our model have natural probabilistic interpretations. An important and very practical consequence of such an interpretation is that uncertainties can be easily quantified and integrated into the system. Moreover, all desired quantities can be inferred consistently. For example once we quantify the distribution of tempo deviations and expressive timing, the actual behavior of the tempo tracker arises automatically from these a-priori assumptions. This is in contrast to other models where one has to invent ad-hoc methods to avoid undesired or unexpected behavior on real data.

Additionally, 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. 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. Indeed, we have demonstrated that all model parameters can be estimated from experimental data.

We have investigated several potential directions in which the basic dynamical model can be improved or simplified. We have tested the relative relevance of the filter order, switching and the tempogram representation on a systematically collected set of natural data. The dataset consists of polyphonic piano performances of two Beatles songs (Yesterday and Michelle) and contains a lot of tempo fluctuation as indicated by the poor constant tempo fits.

The test results on the Beatles dataset suggest that using a high order filter does not improve tempo tracking performance.Although beat level filters capture some structure in tempo deviations (and hence can generate more accurate predictions), this additional precision seems to be not very important in tempo tracking. This indifference may be due to the fact that training criteria (maximum likelihood) and testing criteria (tracking index), whilst related, are not identical. However, one can imagine scenarios where accurate prediction is crucial. An example would be a real-time accompaniment situation, where the application needs to generate events for the next bar.

Test results also indicate that a simple switching mechanism is not very useful. It seems that a tempogram already gives a robust local estimate of likely beat and tempo values so the correct beat can unambiguously be identified. The indifference of switching could as well be an artifact of the dataset which lacks extensive syncopations. Nevertheless, the switching noise model can further be elaborated to replace the tempogram by a rhythm quantizer (Cemgil et al., 2000).

To test the relevance of the proposed tempogram representation on tracking performance we have compared it to a simpler, onset based alternative. The results indicate that in the onset-only case, tracking performance significantly decreases, suggesting that a tempogram is an important component of the

system.

It must be noted that the choice of a comb basis set for tempogram calculation is rather arbitrary. In principle, one could formulate a "richer" tempogram model, for example by including parameters that control the shape of basis functions. The parameters of such a model can similarly be optimized by likelihood maximization on target tempo tracks. Unfortunately, such an optimization (e.g. with a generic technique such as gradient descent) requires the computation of a tempogram at each step and is thus computationally quite expensive. Moreover, a model with many adjustable parameters might eventually overfit.

We have also demonstrated that the model can be used both online (filtering) and offline (smoothing). Online processing is necessary for real time applications such as automatic accompaniment and offline processing is desirable for transcription applications.

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 to Belinda Thom for her comments to the earlier versions of the manuscript, Ric Ashley and Paul Trilsbeek (MMM Group) for their contribution in the design and running of the experiment and we gratefully acknowledge the pianists from Northwestern University and Nijmegen University for their excellent performances.

References

Cemgil, A. T., Desain, P., and Kappen, H. 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.

Dannenberg, R.B. 1993. "Music understanding by computer". In *Proceedings of the International Workshop on Knowledge Technology in the Arts*.

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.

Godsill, Simon J. and Rayner, Peter J. W. 1998. *Digital Audio Restoration - A Statistical Model-Based Approach*. Springer-Verlag.

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.

Kronland-Martinet, R. 1988. "The wavelet transform for analysis, synthesis and processing of speech and music sounds". *Computer Music Journal*, 12(4):11–17.

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*. Soesterberg: RVO TNO.

Murphy, Kevin. "Switching kalman filters". Technical report, Dept. of Computer Science, University of California, Berkeley, 1998.

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

Roweis, Sam and Ghahramani, Zoubin. 1999. "A unifying review of linear gaussian models". *Neural Computation*, 11(2):305–345.

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

Sethares, W. A. and Staley, T. W. 1999. "Periodicity transforms". *IEEE Transactions on Signal Processing*, 47 (11):2953–2964.

Smith, Leigh. 1999. *A Multiresolution Time-Frequency Analysis and Interpretation of Musical Rhythm*. PhD thesis, University of Western Australia.

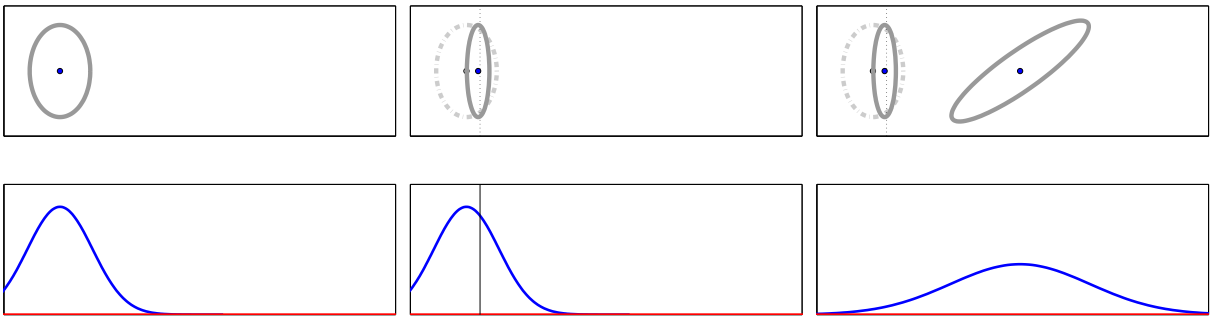
Sterian, A. 1999. *Model-Based Segmentation of Time-Frequency Images for Musical Transcription*. PhD thesis, University of Michigan, Ann Arbor.

Todd, Neil P. McAngus. 1994. "The auditory "primal sketch": A multiscale model of rhythmic grouping.". *Journal of new music Research*.

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

Vercoe, B. 1984. "The synthetic performer in the context of live performance". In *Proceedings of ICMC*, San Francisco. International Computer Music Association, pages 199–200.

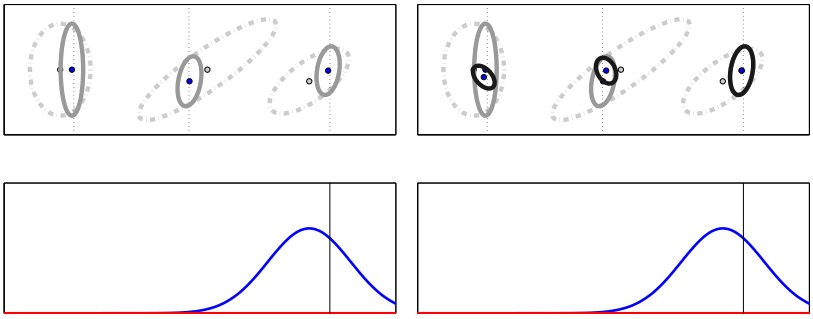
Vercoe, B and Puckette, M. 1985. "The synthetic rehearsal: Training the synthetic performer". In *Proceedings of ICMC*, San Francisco. International Computer Music Association, pages 275–278.



(a) The algorithm starts with the initial state estimate $\mathcal{N}(\mu_{1|0}, P_{1|0})$. In presence of no evidence this state estimate gives rise to a prediction in the observable τ space,

(b) The beat is observed at τ_1, The state is updated to $\mathcal{N}(\mu_{1|1}, P_{1|1})$ according to the new evidence. Note that the uncertainty "shrinks",

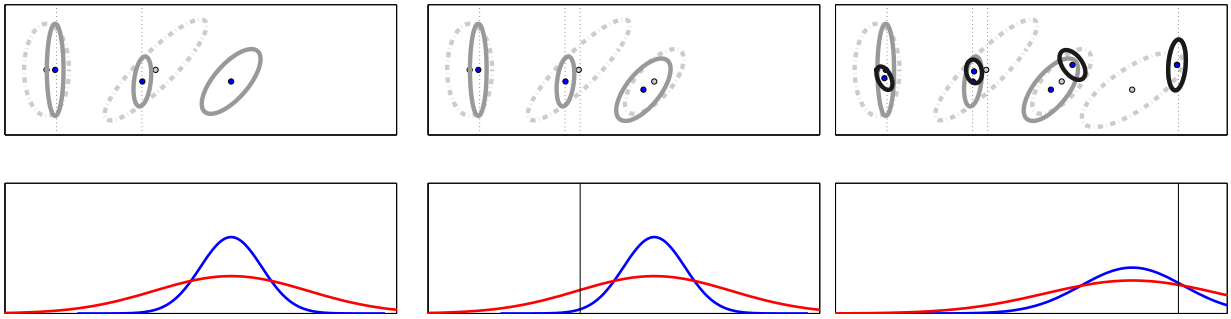
(c) On the basis of current state a new prediction $\mathcal{N}(\mu_{2|1}, P_{2|1})$ is made,



(d) Steps are repeated until all evidence is processed to obtain filtered estimates $\mathcal{N}(\mu_{j|j}, P_{j|j})$, $j = 1 \ldots N$. In this case $N = 3$.

(e) Filtered estimates are updated by backtracking to obtain smoothed estimates $\mathcal{N}(\mu_{i|N}, P_{i|N})$ (Kalman smoothing).

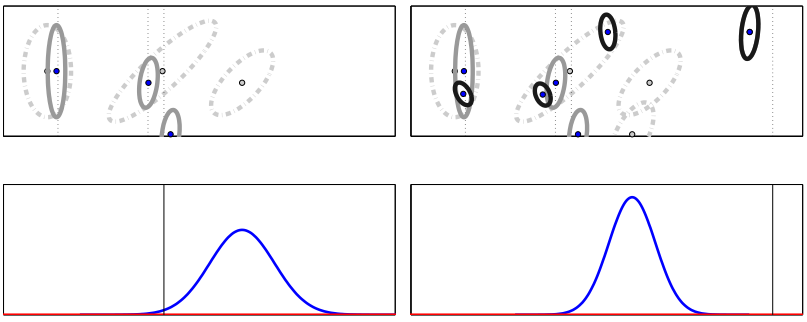
Figure 1: Operation of the Kalman Filter and Smoother. The system is given by Equations 2 and 3. In each subfigure, the above coordinate system represents the hidden state space $[\hat{\tau}, \hat{\Delta}]^T$ and the below coordinate system represent the observable space τ. In the hidden space, the x and y axes represent the phase $\hat{\tau}$ period $\hat{\Delta}$ of the tracker. The ellipse and its center correspond to the covariance and the mean of the hidden state estimate $p(\mathbf{s}_j|\tau_1 \ldots \tau_k) = \mathcal{N}(\mu_{j|k}, P_{j|k})$ where $\mu_{j|k}$ and $P_{j|k}$ denote the estimated mean and covariance given observations $\tau_1 \ldots \tau_k$. In the observable space, the vertical axis represents the predictive probability distribution $p(\tau_j|\tau_{j-1} \ldots \tau_1)$.



(a) Based on the state estimate $\mathcal{N}(\mu_{2|2}, P_{2|2})$ the next state is predicted as $\mathcal{N}(\mu_{3|2}, P_{3|2})$. When propagated through the measurement model, we obtain $p(\tau_3|\tau_2, \tau_1)$, which is a mixture of Gaussians where the mixing coeffi cients are given by $p(c)$,

(b) The observation τ_3 is way off the mean of the prediction, i.e. it is highly likely an outlier. Only the broad Gaussian is active, which reflects the fact that the observations are expected to be very noisy. Consequently, the updated state estimate $\mathcal{N}(\mu_{3|3}, P_{3|3})$ is not much different than its prediction $\mathcal{N}(\mu_{3|3}, P_{3|3})$. However, the uncertainty in the next prediction $\mathcal{N}(\mu_{4|3}, P_{4|3})$ will be higher,

(c) After all observations are obtained, the smoothed estimates $\mathcal{N}(\mu_{j|4}, P_{j|4})$ are obtained. The estimated state trajectory shows that the observation τ_3 is correctly interpreted as an outlier.



(d) In contrast to the switching Kalman fi lter, the ordinary Kalman fi lter is sensitive against outliers. In contrast to (b), the updated state estimate $\mathcal{N}(\mu_{3|3}, P_{3|3})$ is way off the prediction.

(e) Consequently a very "jumpy" state trajectory is estimated. This is simply due to the fact that the observation model does not account for presence of outliers.

Figure 2: Comparison of a standard Kalman filter with a switching Kalman filter.

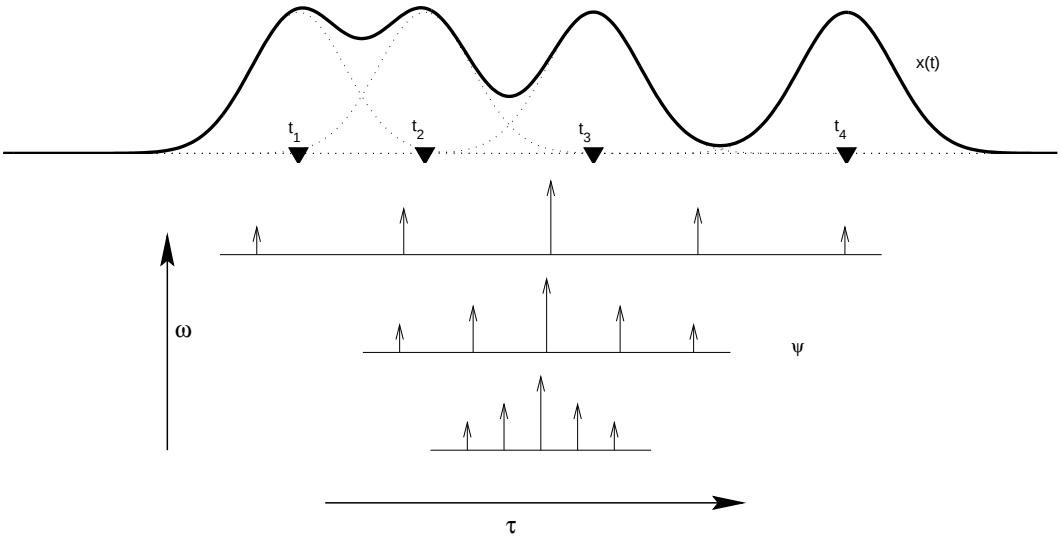


Figure 3: Tempogram Calculation. The continuous signal $x(t)$ is obtained from the onset list by convolution with a Gaussian function. Below, three different basis functions ψ are shown. All are localized at the same τ and different ω. The tempogram at (τ, ω) is calculated by taking the inner product of $x(t)$ and $\psi(t; \tau, \omega)$. Due to the sparse nature of the basis functions, the inner product operation can be implemented very efficiently.

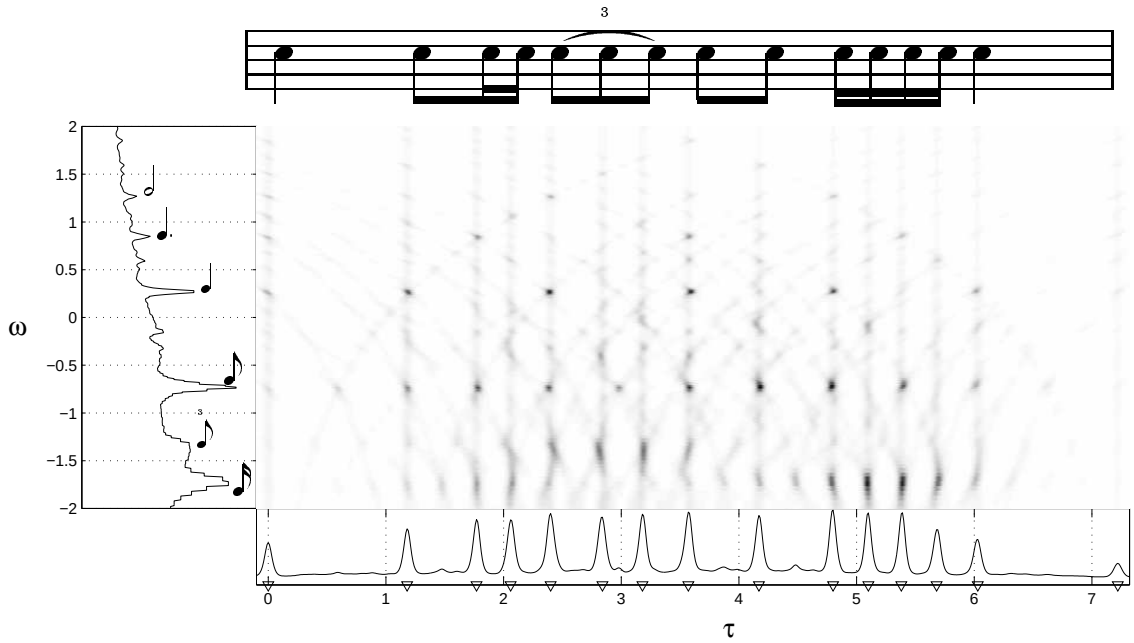


Figure 4: A simple rhythm and its Tempogram. x and y axes correspond to τ and ω respectively. The bottom figure shows the onset sequence (triangles). Assuming flat priors on τ and ω, the curve along the ω axis is the marginal $p(\omega|\mathbf{t}) \propto \int d\tau \exp(\mathrm{Tg}_{\mathrm{x}}(\tau, \omega))$. We note that $p(\omega|\mathbf{t})$ has peaks at ω, 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.

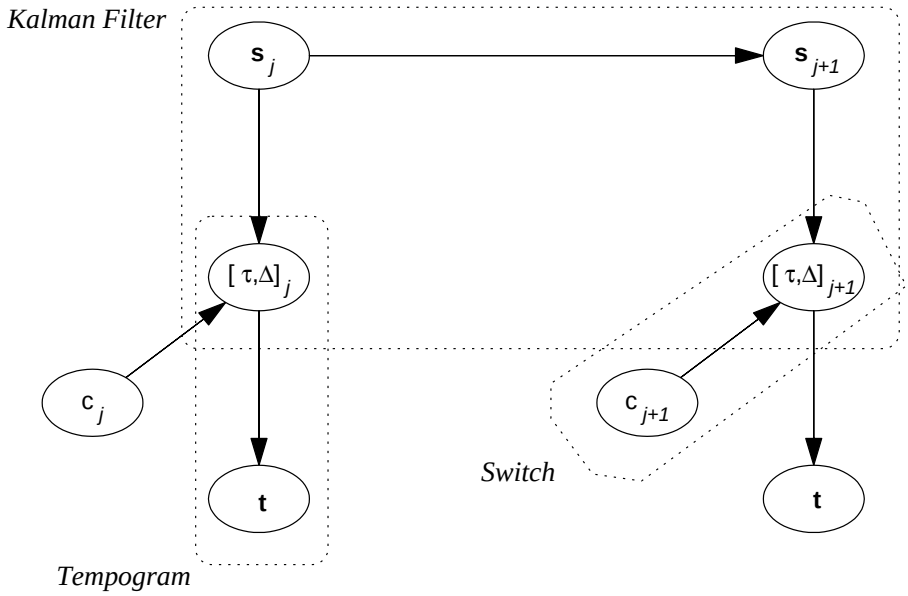


Figure 5: The Graphical Model

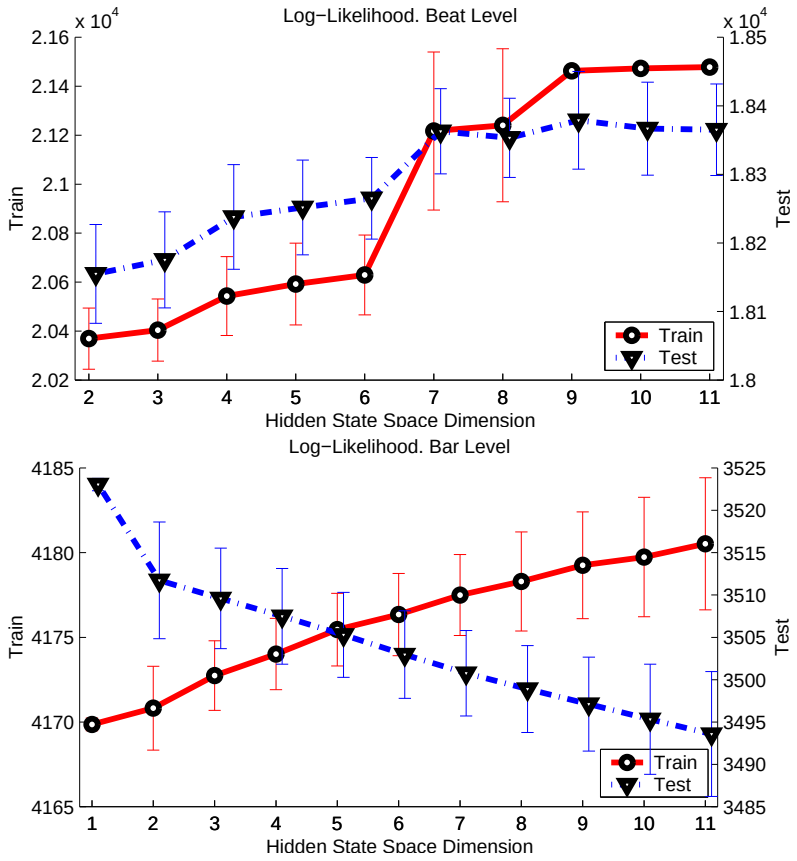


Figure 6: Kalman Filter training. Training Set: Michelle, Test Set: Yesterday.

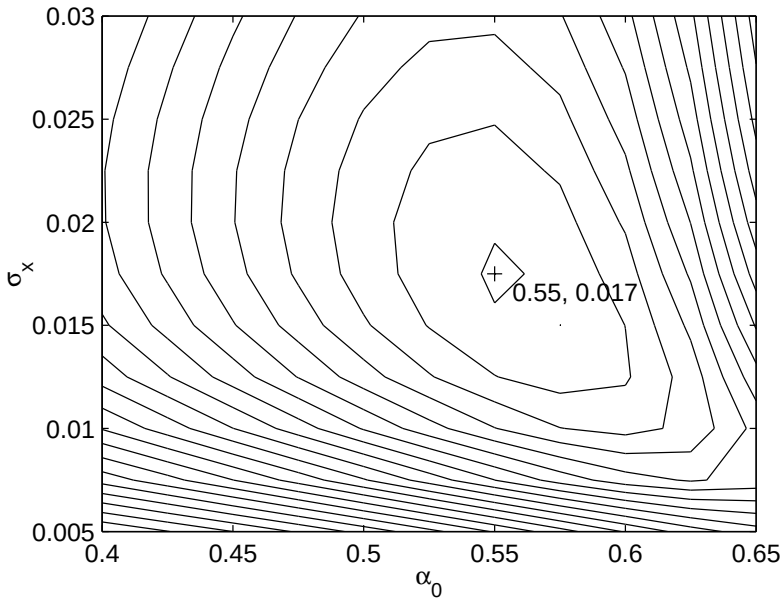


Figure 7: Log-likelihood surface of tempogram parameters α and σ_x on Michelle dataset.