A Dynamical Bayesian Network for Tempo and Polyphonic Pitch Tracking

Ali Taylan Cemgil[1], David Barber[2], Bert Kappen[1]

Abstract— **We present a model for simultaneous tempo and polyphonic pitch tracking. Both these acoustic analysis tasks are difficult and, arguably, no satisfactory solution currently exists for polyphonic pitch tracking. Our model, a form of Dynamical Bayesian Network, embodies a transparent and computationally tractable approach to this acoustic analysis problem. An advantage of our approach is that it places emphasis on modelling the sound generation procedure. It provides a clear framework in which both high level (cognitive) prior information on music structure can be coupled with low level (acoustic physical) information in a principled manner to perform the analysis. The model is readily extensible to more complex sound generation processes.**

I. Introduction

Consider the following scenario: a performer is improvising freely on a musical instrument. A computer processes the performance, extracting high level information including the pitch, tempo and expressive characteristics. Our aim in this paper is to consider a computational framework to move us closer to the realisation of such a scenario[11], [1]. To interface an acoustical instrument to a computer, one needs a mechanism to sense and characterize individual events and expressive components of the sound produced by the instrumentalist. One potential solution is to use dedicated hardware and install special sensors on to the instrument body: this solution is has restricted flexibility and is applicable only to instruments designed specifically for such a purpose.

Discounting the above 'hardware' solution, we shall assume throughout the rest of this paper that we capture the sound with a microphone, so that the computer receives then no further input other than the pure acoustic information. In this paper, we desire to analyze such low level information and extract high level information including the pitch and tempo, a form of transcription[11]. Clearly, other kinds of high level information are potentially important in defining the musical expression of a performance, and a long term goal is to be able to extract these features as well. Whilst not considered here, several motivations for the subsequent use of such extracted information exist. It might be that this information is in itself the main interest; this could be used, for example, as part of a remastering process. Other uses include giving a response, based on the computer's understanding of the musical scene, such that the computer may be able, to 'jam' with a performer during a live performance[9]. The real time nature of this scenario motivates the need for fast computational procedures.

1. Nijmegen University, Nijmegen, The Netherlands, 2.Edinburgh University, EH1 2QL, U.K.

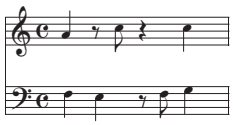
In our view, what is missing in current approaches to this problem is a consistent modelling framework and computational machinery to interface low level signal processing to high level musical knowledge. In this paper we demonstrate a methodology that focuses on the generative process underlying a musical performance. This includes modelling both the instrument (low level) as well as the instrumentalist (high level). We use as a running example the task of estimating the pitch and tempo of a musical performance but believe that the methods discussed here are generally applicable to a broad spectrum of applications for real time human-computer interaction.

II. Model

A basic starting point for music transcription from audio is the extraction of pitch and tempo information from an acoustic signal. The problem can be conveniently described in a Bayesian framework: given the audio samples, we wish to infer the onset times (times at which a 'string' is 'plucked'), note durations, tempo as well as the pitch classes of individual notes. We assume that we have essentially one microphone, so that at each time t, we have a one dimensional observed quantity y_t. Multiple microphones (such as required for stereo) would be straightforward to include in our model. We denote the audio samples $\{y_1, y_2, \ldots, y_t, \ldots, y_T\}$ by the shorthand notation $y_{1:T}$. A constant sampling frequency F_s is assumed, i.e. the actual time elapsing between two consecutive samples t and $t+1$ is $1/F_s$ seconds. We conceptualize the physical instrument (including, for example, voice) as a set of vibrating strings. Our approach considers the quantities we wish to infer, namely tempo and pitch, as 'hidden' (unobserved) quantities, whilst acoustic recording values $y_{1:T}$ are 'visible' (observed). Let us denote the unobserved quantities by $\mathcal{H}_{1:T}$ where each $\mathcal{H}_t$ is a vector. Our hidden variables will contain, in addition to the tempo and pitch, other variables required to complete the sound generation procedure. We will elucidate their meaning later. As a general inference problem, the posterior distribution is given by

$$p(\mathcal{H}_{1:T}|y_{1:T}) \quad \propto \quad p(y_{1:T}|\mathcal{H}_{1:T})p(\mathcal{H}_{1:T}) \tag{1}$$

The likelihood term $p(y_{1:T}|\mathcal{H}_{1:T})$ in (1) requires us to specify a generative process that gives rise to the observed audio samples, i.e., a sound synthesis model for the musical instrument as well as a performance model that describes timing characteristics of the performer. The prior term $p(\mathcal{H}_{1:T})$ reflects our general knowledge about the nature of these hidden quantities, e.g., range of pitch classes or amount of reasonable tempo fluctuations. We will frame



k	c	l	n	class
1	0	1	69	A
2	0	1	53	F
3	1	1	52	E
4	1.5	0.5	72	C
5	2.5	0.5	53	F
6	3	1	72	C
7	3	1	55	G

t	1	2	3	4	...	465	466	467	...
θ	-1	-0.99	-0.99	-0.97	...	-1.23	-1.23	-1.22	...
v	0	0.002	0.004	0.006	...	0.998	1.001	1.003	...
k	0	0	1	2	...	2	2	3	...
h_o	0	0	0	1	...	1	1	1.5	...
κ	0	1	1	0	...	0	1	0	...

Fig. 1. (Top) Simple polyphonic score and the sequence of note events it represents. The k'th note has three attributes: the score position c_k, duration l_k and the pitch index n_k. (Bottom) A possible realization from the generative model. Variables are described in the text.

our model as a Dynamical Bayesian Network which are extensions of Kalman Filters[6]; references to inference and learning issues in such networks are given in [4].

III. Model

Musical signals have a very rich temporal structure, both on physical (signal) and cognitive (symbolic) level. From a statistical modeling point of view, such a hierarchical structure induces very long range correlations, that are difficult to capture with conventional signal models. Moreover, in many music applications, such as transcription or score following, we are usually interested into a symbolic representation (such as a score) and not so much into the "details" of the actual waveform. To abstract away from the signal details, we define a set of intermediate variables (a sequence of indicators and pitches), somewhat analogous to a "piano roll" representation. This piano roll representation will form an "interface" between a symbolic representation and the actual signal process. We will first introduce a *Score* and a *Timer* model to induce a prior on piano rolls. Conditioned on the piano roll, we will define a *Signal* model; a sinusoidal model that we will formulate as a conditionally Gaussian process (a Kalman filter model). Roughly, the score model describes how a piece is composed, a timer model describes how it is performed, and a signal model describes how the actual waveform is synthesized.

A. Timer and Score Models

Our timer model, when viewed as a probabilistic generative model, is analogous to a MIDI sequencer, a program that schedules note events and generates control signals that drive a sound generating device. We imagine that each performance is a realization from a score. The score itself is generated by a score model and is "performed" by an "expressive" sequencer. An expressive sequencer, like a human performer, can fluctuate the tempo or introduce timing deviations (plays scheduled notes a little bit earlier or later). The generated control signals, when viewed as functions of actual time, constitute an intermediate representation analogous to a piano roll. In Figure 1, we show a simple polyphonic score and the corresponding note sequence.

We implement the timer mechanism as follows: At each time step, a continuous variable, v, the *score position pointer*, is increased monotonically with a rate proportional to the tempo. Each time the pointer v reaches the next note in the score, an interrupt is generated. We represent the tempo in log-period by θ_t. For example, a tempo of 120 beats per minute corresponds to $\theta = \log_2 60/120 = -1$. At each new sample, we allow the tempo to change by a small amount $\epsilon_\theta \sim \mathcal{N}(0, \Sigma_\theta)$.

$$\begin{aligned} \theta_t &= \theta_{t-1} + \epsilon_\theta \\ v_t &= v_{t-1} + 2^{-\theta_t}/F_s \end{aligned}$$

When θ becomes large, the score pointer v is incremented less so the tempo gets effectively slower.

To represent the score, we define a counter variable k_t that counts the number of notes we have generated so far. We also define $h_{o,t}$, the *onset threshold*, that specifies the score position of the *next* note c_{new}

$$\begin{aligned} k_t &= k_{t-1} + [\kappa_{t-1} = \text{onset}] \\ c_{\text{new}} &\sim f(c|h_{o,t-1}, k_t) \\ h_{o,t} &= h_{o,t-1}[\kappa_{t-1} \neq \text{onset}] + c_{\text{new}}[\kappa_{t-1} = \text{onset}] \end{aligned}$$

Above $f(c|h_{o,t-1}, k_t)$ is a distribution on score positions of notes, that reflects the statistics of scores that we expect to generate. If the score would be given, then $c_{\text{new}} = c_{k_t+1}$ and f would be a deterministic (degenerate) distribution. Here, [Q] is an indicator that evaluates to 1 (0) when the Boolean proposition Q is true (false). We generate an interrupt if $v_t \geq h_{o,t}$,i.e., when the score pointer has reached the onset threshold; this decision is made "softer" by using a sigmoid $\sigma(x) \equiv 1/(1 + \exp(-ax))$ where we define the probability of an onset as

$$p(\kappa_t = \text{onset}|v_t, h_{o,t}) = \sigma(v_t - h_{o,t})$$

The sigmoid parameter a adjusts the timing accuracy: a smaller a allows for more deviation from the value specified by the threshold $h_{o,t}$. A numerical example of this onset generation is given in Figure 1. The graphical submodel of the timer and score process are shown in the top section of Figure 2. At any time t, we assume that our idealized polyphonic instrument can produce at most M independent voices or notes, i.e. has M sound generators. A loose analogy would be a guitar with M strings or a piano with M keys. When an onset is generated by the timer process, the index of a sound generator is drawn $m_{\text{new}} \sim f(m|k_t)$. If the score would be known and each generator would be assigned to a unique note (e.g. as in a piano) then f would be a deterministic mapping. We denote the label of the selected sound generator by m_t. We reserve $m_t = 0$ for the case when no onset is to be generated at time t. Thus :

$$m_t = 0 \cdot [\kappa_{t-1} \neq \text{onset}] + m_{\text{new}}[\kappa_{t-1} = \text{onset}]$$

With each sound generator $j = 1 \ldots M$, we associate a sequence of threshold variables $h_{j,t}$ that denote the score

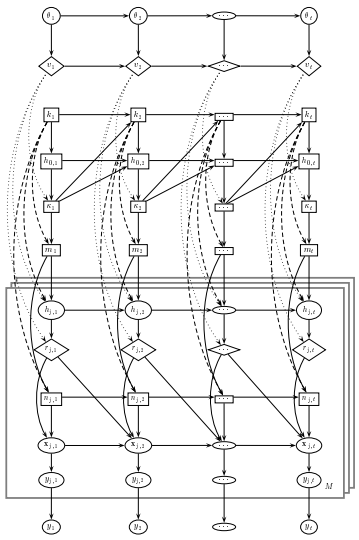


Fig. 2. Graphical Model. Signal model parameters ω_t, ρ_t, transient noise process z_t and periodic process s_t are not explicitly shown, but are summarized as $\mathbf{x}$. The rectangle box denotes "plates", M replications of the nodes inside. Some links are plotted dotted only to improve readability.

position of the next note offset

$$\begin{aligned} d_{\text{new}} &\sim f(d|k_t) \quad n_{\text{new}} \sim f(n|k_t) \\ h_{\text{new}} &= v_i + d_{\text{new}} \\ & \quad j = 1 \dots M \\ h_{j,t} &= h_{j,t}[j \neq m_t] + h_{\text{new}}[j = m_t] \\ n_{j,t} &= n_{j,t}[j \neq m_t] + n_{\text{new}}[j = m_t] \end{aligned}$$

The distribution $f(d|k_t)$ specifies how the current note is articulated, possibly depending upon its length l_{k_t} as notated in the score. Similarly, $f(n|k_t)$ specifies the pitch of current note. Each indicator $r_{j,t}$ is binary, with values "sound" or "mute". Given $h_{j,t}$ and v_t, the state of the indicator $r_{j,t}$ is deterministic:

$$r_{j,t} = \text{sound}[v_t \leq h_{j,t}] + \text{mute}[v_t > h_{j,t}]$$

The collection of variables $r_{1:M,1:T}$ and $n_{1:M,1:T}$ represent the piano roll.

B. Signal Model

Musical instruments tend to create oscillations with modes that are roughly related by integer ratios, albeit with strong damping effects and transient attack characteristics [3]. It is convenient to model such signals as the sum of a periodic component and a transient component [10], [8]. The sinusoidal model is often a good approximation that provides a compact representation for the periodic component. The transient component can be modeled as a correlated Gaussian noise process [5], [2]. Our signal model is also in the same spirit, but we will define it in state space form, because this provides a natural way to couple the signal model with the onset generation process. Consider a Gaussian process where typical realizations $y_{1:T}$ are damped "noisy" sinusoidals with (possibly variable) angular frequency ω:

$$\begin{aligned} s_t &= \rho_t B(\omega_t) s_{t-1} + \epsilon_s \quad (2) \\ y_t &= C s_t \quad (3) \end{aligned}$$

Here $B(\omega) = \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}$ is a Givens rotation matrix that rotates a two dimensional vector by ω degrees counterclockwise. C is a projection matrix defined as $C = [1,\ 0]$. The phase and amplitude characteristics of y_t are determined by the initial conditions s_0. The damping factor $0 \leq \rho \leq 1$ specifies the rate s_t contracts to 0. The transition noise term ϵ_s summarizes contributions of unknown factors, e.g., error terms due to nonlinearities that we are not modelling.

In reality, musical instruments (with a definite pitch) have several modes of oscillation that are roughly located at integer multiples of the fundamental frequency ω. Hence, we can model such signals by a bank of simple oscillators giving a block diagonal transition matrix

$$A_t(\omega_t, \rho_t) = \text{diag}\left(\rho_{1,t} B(\omega_t), \rho_{2,t} B(2\omega_t), \dots \rho_{H,t} B(H\omega_t)\right)$$

where H denotes the number of *harmonics*, assumed to be known. The state s_t of this system is concatenation of individual oscillator states. To reduce the number of free parameters, we further assume that $\rho_{h,t} = \rho_t^h$, motivated by the fact that damping factors of harmonics in a vibrating string scale approximately geometrically with respect to that of the fundamental frequency, i.e. higher harmonics decaying faster.

We model the transient component z_t as white noise with exponentially decaying variance

$$\begin{aligned} q_t &= \alpha q_{t-1} \\ z_t &= q_t^{1/2} \epsilon_{z,t} [r_t = \text{sound}] + \epsilon_0 \end{aligned}$$

where $\epsilon_{z,t} \sim \mathcal{N}(0,1)$, $\epsilon_0 \sim \mathcal{N}(0, R)$ and $0 \leq \alpha < 1$. We assume here that all the transient component parameters (initial variance q_0, variance decay parameter α and the variance R of the "steady state" noise ϵ_0 is known. The parameter update equations for each sound generator $j = 1 \dots M$

$$\begin{aligned} \omega_{\text{new}} &\sim f(\omega|n_{j,t}) \quad s_{\text{new}} \sim f(s) \\ \text{onset}_j &= (r_{j,t-1} = \text{mute} \wedge r_{j,t} = \text{sound}) \\ \log \omega_{j,t} &= (\log \omega_{j,t-1} + \epsilon_\omega)[\neg\ \text{onset}_j] + \log \omega_{\text{new}}[\text{onset}_j] \\ \rho_{j,t} &= \rho_{\text{sound}_j}[r_{j,t} = \text{sound}] + \rho_{\text{mute}}[r_{j,t} = \text{mute}] \\ q_{j,t} &= \alpha q_{j,t-1}[\neg \text{onset}_j] + q_0[\text{onset}_j] \end{aligned}$$

where ρ_{sound} and ρ_{mute} are decay coefficients such that $1 \geq \rho_{\text{sound}} > \rho_{\text{mute}} > 0$. We use a deterministic mapping $f(\omega|n_{j,t})$ to generate the rotation angle given the pitch label. To allow for mistuned notes one can also use a narrow Gaussian. We assume a Gaussian initial state distribution $f(s) = \mathcal{N}(0, S)$. The total energy injected into the string at an onset (mute $\rightarrow$ sound transition in r_j) is proportional to $\det S$ and the covariance structure of S describes

how this total energy is distributed among the harmonics. Thus, $f(s)$ captures the timbre characteristics of the sound. Given the parameters, each sound generator $j = 1 \dots M$ produces the next sample

$$\begin{aligned} s_{j,t} &= A_t(\omega_{j,t}, \rho_{j,t}) s_{t-1}[\neg \text{ onset}] + s_{\text{new}}[\text{onset}] + \epsilon_{s,j,t} \\ z_{j,t} &= q_{j,t}^{1/2} \epsilon_{z,j,t}[r_{j,t} = \text{sound}] + \epsilon_0 \\ y_{j,t} &= C s_{j,t} + z_{j,t} \end{aligned}$$

In the above, C is a $1 \times 2H$ projection matrix $C = [1, 0, 1, 0, \dots, 1, 0]$ with zero entries on the even components. This effectively sums contributions of each harmonic. Finally, the observed audio signal is the superposition of the outputs of all sound generators.

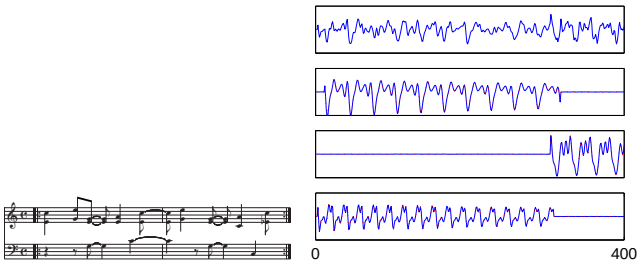
$$y_t = \sum_j y_{j,t}$$

IV. Results and Discussion

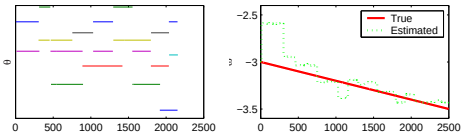
The dynamical model introduced here is a dynamic Bayesian network [4] in which exact computation of posterior features is intractable. We are currently investigating efficient approximation methods mainly focusing on Rao Blackwellized sequential importance sampling and iterative improvement [1]. Such a hybrid approach enables us to exploit analytical structure and deterministic relations. For example, the signal model, given ω and the indicators r, is a factorial Kalman filter model, where integrations can be computed analytically. Space here does not allow us to detail a full inference procedure for our model, which will be described elsewhere (in preparation).

In Fig. IV we show some preliminary results for tempo and pitch tracking, using sequential Monte Carlo. We have rendered a signal y_t from the score Fig. IV(a) with an accelerating tempo. A small segment of this sequence is shown in the upper part of Fig. IV(b). In this example, to demonstrate tempo tracking and pitch tracking we assume that we know $\kappa_{1:T}$. The lower part show that we can reconstruct the original signals essentially perfectly. Knowing the onsets and observation sequence alone, we can infer accurately the hidden pitch labels Fig. IV(c) – that is, which 'notes' are being played in the given section of sound. Similarly, the tempo can be inferred reasonably well. These preliminary results are encouraging, but do not yet constitute a full and efficient procedure for inferring all hidden quantities. However, these initial results demonstrate that accurate pitch and tempo tracking is possible using our framework, although computational obstacles still need to be overcome to achieve real-time performance.

The work presented here is a model driven approach where transcription is viewed as a Bayesian inference problem. In this respect, our approach is similar to previous work of [12], [2], [7]. On the other hand, in our knowledge our work the first demonstration of a compact and realistic generative model for musical signals that combines a dynamical segment model and a signal model. By integrating tempo tracking with signal analysis one can design fast approximation techniques for detection of onsets, i.e. change



(a) Score (b) Sources



(c) Inferred Piano Roll (d) Inferred Tempo

Fig. 3. (a) Original Score (b) The upper plot shows a section of the original acoustic signal y_t and reconstructed signals of the first three notes for the same time window. These reconstructions are indistinguishable from the original sources. Added together, the sources almost perfectly reconstruct the original signal y_t. (c) Given the onsets and note durations, we can estimate the pitch, which is an exact representation of the original score. (d) Assuming the correct onset sequence, we can estimate the tempo.

points. For example, in a performance is almost constant, the tempo gives a lot of information about locations of future onsets. The model can also be used to construct a score follower, essentially by just clamping the score variables and inferring the score position pointer. Similarly, a multipitch tracker can be formulated as a procedure to infer $p(\omega_{1:M,1:t}|y_{1:t})$.

References

[1] A. T. Cemgil and H. J. Kappen. Monte Carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, 18:45–81, 2003.

[2] M. Davy and S. J. Godsill. Bayesian harmonic models for musical signal analysis. In *Bayesian Statistics 7*, 2003.

[3] N. H. Fletcher and T. Rossing. *The Physics of Musical Instruments*. Springer, 1998.

[4] K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[5] L. Parra and U. Jain. Approximate Kalman filtering for the harmonic plus noise model. In *Proc. of IEEE WASPAA*, New Paltz, 2001.

[6] L. R. Rabiner. A tutorial in hidden Markov models and selected applications in speech recognation. *Proc. of the IEEE*, 77(2):257–286, 1989.

[7] Christopher Raphael. Automatic transcription of piano music. In *Proceedings of the International Symposium on Music Information Retrieval*, IRCAM/Paris, 2002.

[8] X. Rodet. Musical sound signals analysis/synthesis: Sinusoidal + residual and elementary waveform models. *Applied Signal Processing*, 1998.

[9] R Rowe. *Machine Musichanship*. MIT Press, 2001.

[10] X. Serra and J. O. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1991.

[11] Barry L. Vercoe, William G. Gardner, and Eric D. Scheirer. Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proc. IEEE*, 86:5:922–940, May 1998.

[12] P. J. Walmsley. *Signal Separation of Musical Instruments*. PhD thesis, University of Cambridge, 2000.