# GENERATIVE MODEL BASED POLYPHONIC MUSIC TRANSCRIPTION

*Ali Taylan Cemgil, Bert Kappen*

SNN, University of Nijmegen,
The Netherlands {cemgil,bert}@snn.kun.nl

*David Barber*

Edinburgh University
dbarber@anc.ed.ac.uk

**ABSTRACT**

In this paper we present a model for simultaneous tempo and polyphonic pitch tracking. Our model, a form of Dynamical Bayesian Network [1], embodies a transparent and computationally tractable approach to this acoustic analysis problem. An advantage of our approach is that it places emphasis on modeling 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.

## 1. INTRODUCTION

When humans listen to sound, they are able to associate acoustical signals generated by different mechanisms with individual symbolic events [2]. The study and computational modeling of this human ability forms the focus of computational auditory scene analysis (CASA) and machine listening [3]. Traditionally, the focus was in speech applications. Recently, analysis of musical scenes[4] is drawing increasingly more attention, primarily because of the need for content based retrieval in digital audio databases and increasing interest in interactive music performance systems.

One of the hard problems in musical scene analysis is automatic music transcription: to infer automatically a musical notation (such as the traditional western music notation) that lists the pitch levels of notes and corresponding timestamps in a given performance. However, in its most unconstrained form, i.e., when operating on an arbitrary polyphonic acoustical input, possibly containing an unknown number of different instruments, music transcription remains yet as a difficult engineering problem. Our aim in this paper is to consider a computational framework to move us closer to a practical solution to this problem.

Music transcription has attracted quite an amount of research effort in the past. See [4] for a detailed review of early work. In speech processing, tracking the pitch of a single speaker is a fundamental problem and methods proposed in the literature fill volumes [5]. A vast majority of pitch detection algorithms are based on heuristics (e.g., picking high energy peaks of a spectrogram, correlogram, auditory filter bank, etc.) and their formulation usually lacks an explicit objective function or a signal model. Hence, it is often difficult to theoretically justify merits and shortcomings of a proposed algorithm, compare it objectively to alternatives or extend it to more complex scenarios.

Pitch tracking is inherently related to detection and estimation of sinusoidals, a topic that has also been deeply investigated in statistics, e.g. see [6]. However, ideas from statistics seem to be applied less in the context of musical sound analysis and pitch tracking. Some exceptions include the work in [7] that presents a realtime monophonic pitch tracking application based on Laplace approximation to the posterior parameter distribution of an AR(2) model. A more sophisticated Kalman filter based pitch tracker is proposed by [8] that tracks parameters of a harmonic plus noise model (HNM) for monophonic speech.

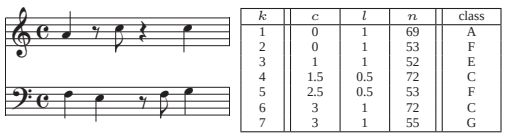
Kashino [9] is, to our knowledge, the first author to apply graphical models explicitly to the problem of polyphonic music transcription. Sterian [10] described a system that viewed transcription as a model driven segmentation of a time-frequency image. Walmsley [11] treats transcription and source separation in a full Bayesian framework. He employs a frame based generalized linear model (a sinusoidal model) and proposes inference by reversible-jump Markov Chain Monte Carlo (MCMC) algorithm. The main advantage of the model is that it makes no strong assumptions about the signal generation mechanism, and views the number of sources as well as the number of harmonics as unknown model parameters. Davy and Godsill [12] address some of the shortcomings of his model and allow changing amplitudes and deviations in frequencies of partials from integer ratios. The reported results are good, however the method is computationally expensive.

Most of the authors view automated music transcription as a "audio to piano-roll" conversion and usually view "piano-roll to score" as a separate problem. This view is partially justified, since source separation and transcription from a polyphonic source is already a challenging task. On the other hand, automated generation of a human readable score includes nontrivial tasks such as tempo tracking, rhythm quantization, meter and key induction [13]. We believe that a model that integrates these higher level symbolic prior knowledge can guide and potentially improve the inferences, as partially demonstrated by [14], both in terms of quality of the solution and computation time.

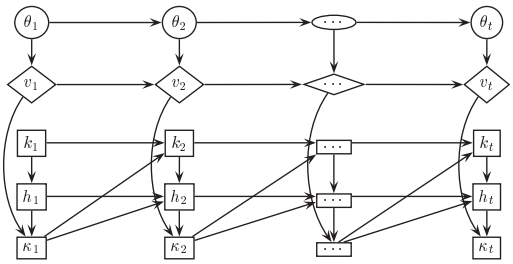
In a statistical sense, music transcription, (as many other perceptual tasks such as visual object recognition or robot localization) can be viewed as a latent state estimation problem: given the audio signal, we wish to infer the underlying score (i.e. collection of onset times, note durations, pitch classes, etc. ). We assume that we have one microphone which we sample with a constant sampling frequency $F_s$. We will denote the audio samples $\{y_1, y_2, \ldots\ldots, y_T\}$ by $y_{1:T}$. Our approach considers the desired quantities as 'hidden' (unobserved), 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. 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. The prior term $p(\mathcal{H}_{1:T})$ reflects our knowledge about the hidden



| $k$ | $c$ | $l$ | $n$ | 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 | 5 | ... | 465 | 466 | 467 | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | -1 | -0.99 | -0.99 | -0.97 | -0.96 | ... | -1.23 | -1.23 | -1.22 | ... |
| $v$ | 0 | 0.002 | 0.004 | 0.006 | 0.008 | ... | 0.998 | 1.001 | 1.003 | ... |
| $k$ | 0 | 0 | 1 | 2 | 2 | ... | 2 | 2 | 3 | ... |
| $h$ | 0 | 0 | 0 | 1 | 1 | ... | 1 | 1 | 1.5 | ... |
| $\kappa$ | 0 | 1 | 1 | 0 | 0 | ... | 0 | 1 | 0 | ... |

Figure 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) The graphical model for the timer process and a possible realization. $\kappa$ values are shown as $[\kappa = \text{onset}]$. Other variables are described in the text.

variables. Our hidden variables will contain, in addition to the score, other variables (e.g. tempo) required to complete the sound generation procedure.

## 2. 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 in a symbolic representation (such as a score) and not so much in 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.

### 2.1. 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. In Figure 1, we show a simple polyphonic score and the corresponding note sequence. 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.

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 and an indicator variable, $\kappa$, is set to the 'onset' state. We represent the tempo in log-period by $\theta_t$. 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 $\theta$ 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 $h_t$, the *onset threshold*, that specifies the score position of the *next* note $c_{\text{new}}$

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

Above $f(c|h_{t-1}, k_t)$ 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 $v_t \geq h_t$ ,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

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

The sigmoid parameter $a$ adjusts the timing accuracy: a smaller $a$ allows for more deviation from the value specified by the threshold $h_t$. The graphical submodel of the timer process and a numerical example are shown in in Figure 1. 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 (e.g. 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 $f(m|k_t)$ would be a deterministic distribution. 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 $g_{j,t}$ that denote the score position of

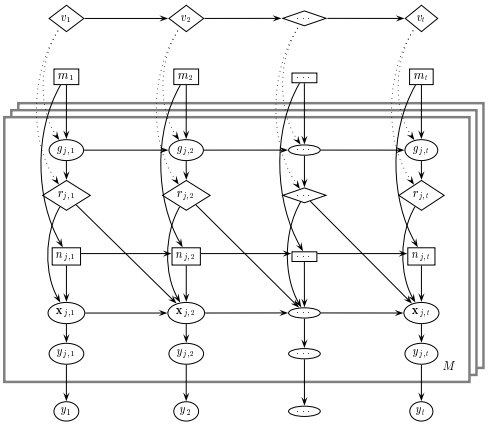



Figure 2: Graphical Model of the signal process. Timer model variables and their links are omitted for clarity. Parameters $\omega_t$, $\rho_t$, transient noise process $z_t$ and periodic process $s_t$ are also not explicitly shown, but are summarized as $\mathbf{x}$. The rectangle box denotes "plates", $M$ replications of the nodes inside.

the next note offset

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

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 $g_{j,t}$ and $v_t$, the state of the indicator $r_{j,t}$ is deterministic:

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

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

### 2.2. 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 [15]. It is convenient to model such signals as the sum of a periodic component and a transient component [16, 17]. 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 [8, 12]. 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 $\omega$:

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

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 $\omega$ 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 $\epsilon_s$ 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 $\omega$. 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 a 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 $\alpha$ and the variance $R$ of the "steady state" noise $\epsilon_0$ is known. The parameter update equations for each sound generator $j = 1 \dots M$

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

where $\rho_{\text{sound}}$ and $\rho_{\text{mute}}$ 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

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

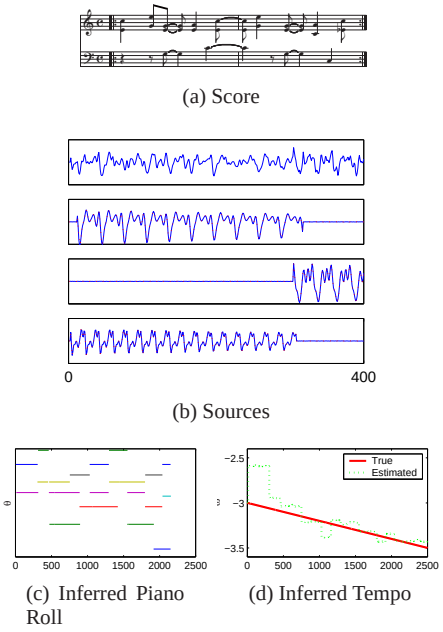
In the above, $C$ is a $1 \times 2H$ projection matrix $C = [1, 0, 1, 0, \dots, 1, 0]$ with zero entries on the even components. This effectively sums contributions of each harmonic. Finally, the observed audio signal is the superposition of the outputs of all sound generators where $y_t = \sum_j y_{j,t}$.

## 3. RESULTS AND DISCUSSION

The dynamical model introduced here is a dynamic Bayesian network [1] 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 [13]. Such a hybrid approach enables us to exploit analytical structure and deterministic relations. For example, the signal model, given $\omega$ 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. 3 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. 3(a) with an accelerating tempo. A small segment of this sequence is shown in the upper part of Fig. 3(b). In this example, to demonstrate tempo tracking and pitch tracking where 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. 3(c) and the tempo. 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. By integrating tempo tracking with signal analysis one can potentially design fast approximation techniques for detection of onsets, i.e. change points. For example, if a performance has almost constant tempo, a correct estimate of the tempo gives a lot of information about locations of future onsets.

The work presented here is a model driven approach where transcription is viewed as a Bayesian inference problem, similar to previous work of [11, 12, 14]. On the other hand, in our knowledge, our work is the first demonstration of a compact and realistic generative model for musical signals that combines a dynamical segment model and a signal model. Our model, with minor modifications, can be potentially useful in applications other than transcription. For example, we can 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 $p(\omega_{1:M,1:t}|y_{1:t})$.




(a) Score

(b) Sources

(c) Inferred Piano Roll

(d) Inferred Tempo

Figure 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.

## A. REFERENCES


[1] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. dissertation, University of California, Berkeley, 2002.

[2] A. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[3] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 2, pp. 297–336, 1994.

[4] E. D. Scheirer, "Music-listening systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[5] W. J. Hess, *Pitch Determination of Speech Signal*. New York: Springer, 1983.

[6] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*. Cambridge University Press, 2001.

[7] K. L. Saul, D. D. Lee, C. L. Isbell, and Y. LeCun, "Real time voice processing with audiovisual feedback: toward autonomous agents with perfect pitch," in *NIPS*2002*, Vancouver, 2002.

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

[9] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of bayesian probability network to music scene analysis," in *Proc. IJCAI Workshop on CASA*, Montreal, 1995, pp. 52–59.

[10] A. Sterian, "Model-based segmentation of time-frequency images for musical transcription," Ph.D. dissertation, University of Michigan, Ann Arbor, 1999.

[11] P. J. Walmsley, "Signal separation of musical instruments," Ph.D. dissertation, University of Cambridge, 2000.

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

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

[14] C. Raphael, "Automatic transcription of piano music," in *Proc. ISMIR*, IRCAM/Paris, 2002.

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

[16] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1991.

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