Polyphonic Pitch Identification and Bayesian Inference

Ali Taylan Cemgil
IAS, University of Amsterdam, The Netherlands
email: cemgil@science.uva.nl

Abstract

In this article, we review the essential ideas of Bayesian polyphonic pitch identification with simple examples. We describe a signal model and a simple greedy algorithm to infer the pitch classes in a sound mixture with unknown number of notes. An important advantage of our approach is that it provides a clear framework in which both high level prior information on music structure can be coupled with low level acoustic-physical knowledge in a principled manner to perform the analysis.

1 Introduction

One of the hard problems in musical scene analysis is polyphonic pitch tracking. That is, the identification of individual pitch classes from a (usually additive) mixture. This problem in various forms pops up in a vast majority of computer music applications, ranging from interactive music performance systems, music information retrieval (Music-IR) to musicological analysis of performances. In its most unconstrained form, i.e., when operating on an arbitrary polyphonic acoustical input possibly containing an unknown number of different instruments, pitch tracking remains still a challenge.

Pitch identification in polyphony has attracted significant research effort in the past – see (Scheirer 2000) for a detailed overview of early work. In speech processing, the related task of tracking the pitch of a single speaker is a fundamental problem and methods proposed in the literature are well studied (Hess 1983). However, most current pitch detection algorithms are based largely on heuristics (e.g., picking high energy peaks of a spectrogram, correlation, auditory filter bank, etc.) and their formulation usually lacks an explicit objective function or signal model. It is often difficult to theoretically justify the merits and shortcomings of such algorithms, and compare them objectively to alternatives or extend them to polyphonic scenarios.

Pitch tracking is inherently related to the detection and estimation of sinusoids. The estimation and tracking of single or multiple sinusoids is a fundamental problem in many branches of applied sciences, so it is less surprising that the topic has also been deeply investigated in statistics, (e.g. see (Quinn and Hannan 2001)). However, ideas from statistics seem to be not widely applied in the context of musical sound analysis, with only a few exceptions (Irizarry 2002; Saul, Lee, Isbell, and LeCun 2002; Parra and Jain 2001). Sterian (Sterian) described a system that viewed transcription as a model driven segmentation of a time-frequency image. Walmsley (Walmsley) 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 (Davy and Godsill) address some of the shortcomings of his model and allow changing amplitudes and frequency deviations. The reported results are encouraging, although the method is computationally very expensive.

In this article, we describe the essential ideas of Bayesian polyphonic pitch identification with simple examples. For a more detailed and extended description of our approach, the interested reader is referred to (Cemgil, Kappen, and Barber 2004).

2 A Toy Example

Consider a “one sample source identification” problem. Suppose there are two numbers (=sources) \(s_1\) and \(s_2\) which we wish to identify, but we only observe their sum (=superposition, mixture) \(y\). Given only this much of information, of course this problem does not have a unique solution since there are infinitely many number pairs with \(s_1 + s_2 = y\). Suppose now we are given additional information that the sources can be independently present or absent in the mixture. Say if the \(j\)'th source \((j = 1, 2)\) is present in the mixture, it has a Gaussian distribution with mean \(\mu_j\) and variance \(\sigma^2\), which we denote by \(\mathcal{N}(\mu_j, \sigma^2)\). If source \(j\) is absent, say it will be zero mean which we denote by \(\mathcal{N}(0, \sigma^2)\). To get a compact notation, let us denote the condition that source is present or absent with a discrete binary variable \(r_j \in \{\text{“sound”, “mute”}\}\)

\[
  r_j \sim p(r_j) \quad j = 1, 2
\]

\[
  s_j | r_j \sim [r_j \text{ = sound}]\mathcal{N}(\mu_j, \sigma^2) + [r_j \text{ = mute}]\mathcal{N}(0, \sigma^2)
\]

\[
  y = \sum_j s_j
\]

Here, and elsewhere in the article, the notation \([x = \text{state}]\) has value equal to 1 when variable \(x\) is in state text, and is zero otherwise. Now, given the model we can infer the...
The sources $s_j$ correspond to individual note events. The indicator variables $r_j$ denote whether a particular sound is present in the observed mixture $y$. Of course, in reality $y$ and each of $r_j$ and $s_j$ will be time series. Provided we know how the individual sources are generated, source separation and source identification (= transcription) correspond to inference of posterior quantities $p(s_j|y)$ and $p(r_j|y)$ respectively. In the following section, we will define a signal model $p(y|s)p(s|r)$.

3 Model

It is well known that musical instruments tend to create oscillations with modes that are roughly related by integer ratios, albeit with strong damping effects and transient attack characteristics (Fletcher and Rossing 1998; Serra and Smith 1991; Rodet 1998). It is common to model such signals as the sum of a periodic component and a transient non-periodic component. The sinusoidal model (McAulay and Quatieri 1986) is often a good approximation that provides a compact representation for the periodic component. 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 derive online algorithms.

First we consider how to generate a damped sinusoid $y_t$ through time, with angular frequency $\omega$. Consider a Gaussian process where typical realizations $y_1:T$ are damped “noisy” sinusoids with angular frequency $\omega$:

$$s_t \sim N(\rho_1 B(\omega) s_{t-1}, Q)$$

(1)  

$$y_t \sim N(C s_t, R)$$

(2)  

$$s_0 \sim N(0, S)$$

(3)

We use $N(\mu, \Sigma)$ to denote a multivariate Gaussian distribution with mean $\mu$ and covariance $\Sigma$. Here $B(\omega) = \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}$ is a Givens rotation matrix that rotates two dimensional vector $s_t$ 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 condition $s_0$ drawn from a prior with covariance $S$. The damping factor $0 \leq \rho_1 \leq 1$ specifies the rate at which $s_t$ contracts to 0. See Figure 1 for an example. The transition noise variance $Q$ is used to model deviations from an entirely deterministic linear model. The observation noise variance $R$ models background noise.

In reality, musical instruments (with a definite pitch) have several modes of oscillation that are roughly located at integer multiples of the fundamental frequency $\omega$. We can model such signals by a bank of oscillators giving a block diagonal transition matrix $A_t = A(\omega, \rho_t)$ defined as

$$
A_t = 
\begin{pmatrix}
\rho_t^{(1)} B(\omega) & 0 & \cdots & 0 \\
0 & \rho_t^{(2)} B(2\omega) & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \rho_t^{(H)} B(H\omega)
\end{pmatrix}
$$

(4)

where $H$ denotes the number of harmonics, assumed to be known. $A(\omega, \rho_t)$ is the transition matrix at time $t$ and en-
among the harmonics. The covariance matrix \( S \) structure of \( S \) is proportional to the determinant of \( S \). The stochastic model assumes that there are two independent additive noise components that corrupt the state vector \( s \) and the sample \( y \), so the resulting waveform \( y_{1:T} \) is a damped sinusoid with both phase and amplitude noise.

codes the physical properties of the sound generator as a first order Markov Process. The state of the sound generator is represented by \( s_t \), a 2\( H \) dimensional vector that is obtained by concatenation of all the oscillator states in (1).

3.1 From Piano-Roll to Audio Signal

A piano-roll is a collection of indicator variables \( r_{j,t} \), where \( j = 1, \ldots, M \) runs over sound generators (i.e. notes or “keys” of a piano) and \( t = 1, \ldots, T \) runs over time. Each sound generator has a unique fundamental frequency \( \omega_j \) associated with it. For example, we can choose \( \omega_j \) such that we cover all notes of the tempered chromatic scale in a certain frequency range. This choice is arbitrary and for a finer pitch analysis a denser grid with smaller intervals between adjacent notes can be used.

Each indicator is binary, with values “sound” or “mute”. The essential idea is that, if previously muted, \( r_{j,t-1} = \text{“mute”} \), \( r_{j,t} \) is an onset for the sound generator \( j \) occurs if \( r_{j,t} = \text{“sound”} \). The generator continues to sound (with a characteristic damping decay) until it is again set to “mute”, when the generated signal decays to zero amplitude (much) faster. The piano-roll, being a collection of indicators \( r_{1:M,1:T} \), can be viewed as a binary sequence, e.g. see Figure 2. Each row of the piano-roll \( r_{j,1:T} \) controls onsets and offsets of the underlying sound generator by changing the damping coefficient.

We characterise the damping factor for each note \( j = 1, \ldots, M \) by two decay coefficients \( \rho_{\text{sound}} \) and \( \rho_{\text{mute}} \) such that \( 1 \geq \rho_{\text{sound}} > \rho_{\text{mute}} > 0 \). The piano-roll \( r_{j,1:T} \) controls the signal process by changing the damping coefficient of the transition matrix. We denote the transition matrix as \( A_{j}^{\text{mute}} = A(\omega_j, \rho_{\text{mute}}) \); similarly for \( A_{j}^{\text{sound}} \).

3.1.1 Piano-Roll : Onsets

At each new onset, i.e. when \( (r_{j,t-1} = \text{“mute”}) \rightarrow (r_{j,t} = \text{“sound”}) \), the old state \( s_{t-1} \) is “forgotten” and a new state vector is drawn from a Gaussian prior distribution \( \mathcal{N}(0, S) \). This models the energy injected into a sound generator at an onset (this happens, for example, when a guitar string is plucked). The amount of energy injected is proportional to the determinant of \( S \) and the covariance structure of \( S \) describes how this total energy is distributed among the harmonics. The covariance matrix \( S \) thus captures some of the timbre characteristics of the sound. The transition and observation equations are given by

\[
\begin{align*}
\text{isonset}_{j,t} & = (r_{j,t-1} = \text{“mute”} \land r_{j,t} = \text{“sound”}) \\
A_{j,t} & = [r_{j,t} = \text{“mute”}]A_{j}^{\text{mute}} + [r_{j,t} = \text{“sound”}]A_{j}^{\text{sound}} \\
s_{j,t} & \sim [-\text{isonset}_{j,t}]\mathcal{N}(A_{j,t}s_{t-1}, Q) + [\text{isonset}_{j,t}]\mathcal{N}(0, S) \\
y_{j,t} & \sim \mathcal{N}(Cs_{j,t}, R)
\end{align*}
\]

In the above, \( C \) is a \( 1 \times 2H \) projection matrix \( C = [1, 0, 1, 0, \ldots, 1, 0] \) with zero entries on the even components. Hence \( y_{j,t} \) has a mean being the sum of the damped harmonic oscillators. \( R \) models the variance of the noise in the output of each sound generator. Finally, the observed audio signal is the superposition of the outputs of all sound generators,

\[
y_t = \sum_j y_{j,t}
\]

This generative model can be described qualitatively by the graphical model in Figure 3. The important message is that this model has conceptually the same structure as the toy model described in section 2.

4 Simulation Results

To infer the most likely piano-roll we need to compute \( \arg \max_{r_{1:M,1:T}} p(r_{1:M,1:T} | y_{1:T}) \) as in the toy example. Although the exact calculation of this quantity is intractable, we are able to derive a simple approximate inference algorithm which we describe in a chord identification problem.
4.1 Vertical Problem: Chord identification

Chord identification is the simplest polyphonic transcription task. Here we assume that a given audio signal \( y_{1:T} \) is generated by a piano-roll constant piano roll where there are no onsets or offsets for \( t > 1 \). The task is to find the MAP configuration

\[
\hat{r}^*_{1:M} = \arg\max_{r_{1:M}} p(y_{1:T}|r_{1:M})
\]

Each configuration corresponds to a chord. The two extreme cases are “silence” and “cacophony” that correspond to configurations with all \( r_j \) mute or sound respectively. The size of the search space in this case \( 2^M \), which is prohibitive for direct computation.

A simple approximation is based on greedy search: we start iterative improvement from an initial configuration (silence, or randomly drawn from the prior). At each iteration \( i \), we evaluate the probability of all neighbouring configurations of that can be reached by adding or removing single notes. If all neighbours have smaller probability, the algorithm terminates, having found a local maximum. Otherwise, we pick the neighbour with the highest probability and iterate until convergence. We illustrate the algorithm on a signal sampled from the generative model, see Figure 4. We observe that for many examples this procedure is able to identify the correct chord. Using multiple restarts from different initial configurations will improve the quality of the solution at the expense of computational cost.

4.2 Piano-roll estimation

The piano-roll estimation problem can be viewed as an extension of chord identification in that we also detect onsets and offsets for each note within the analysis frame. A practical approach is to analyze the signal in sufficiently short time windows and assume that for each note, at most one change-point can occur within the window. The results of this algorithm on real data are shown in Figure 5.

5 Discussion and Conclusion

We have presented a model driven approach where transcription is viewed as a Bayesian inference problem. In this respect, at least, our approach parallels the previous work of (Walmsley; Davy and Godsill; Raphael). We believe, however, that our formulation, based on a switching state space model, has several advantages. We can remove the assumption of a frame based model and this enables us to analyse music online and to sample precision. Practical approximations to an eventually intractable exact posterior can be carried out frame-by-frame, such as by using a fixed time-lag smoother. This, however, is merely a computational issue (albeit an important one). We may also discard samples to reduce computational burden, and account for this correctly in our model.
Figure 5: Polyphonic transcription of a short segment from a recording of a bass guitar. (Top) The signal, original sampling rate of 22050 Hz is downsampled with a factor of $D = 5$. (Middle) Spectrogram (Short time Fourier transform modulus) of the downsampled signal. Horizontal and vertical axes correspond to time and frequency, respectively. Grey level denotes the energy in a logarithmic scale. The low frequency notes are not well resolved due to short window length. Taking a longer analysis window would increase the frequency resolution but smear out onsets and offsets. (Bottom) Estimated piano-roll. The model used $M = 30$ sound generators where fundamental frequencies were placed on a chromatic scale that spanned the 2.5 octave interval between the low A (second open string on a bass) and a high D (highest note on the forth string). Model parameters are estimated by an expectation-maximization algorithm recorded from the same instrument. The analysis is carried out using a window length of $W = 450$ samples, without overlap between analysis frames (i.e. $L = W$). The greedy procedure was able to identify the correct pitch classes and their onsets to sample precision. For this example, the results were qualitatively similar for different window lengths $W$ around $300 \sim 500$ and downsampling factors $D$ up to 8.

An additional advantage of our formulation is that we can still deliver a pitch estimate even when the fundamental and lower harmonics of the frequency band are missing. This is related to so called virtual pitch perception (Terhardt 1974): we tend to associate notes with a pitch class depending on the relationship between harmonics rather than the frequency of the fundamental component itself.

One of the advantages of our generative model based approach is that we can infer a chord given any subset of data. For example, we can simply downsample $y_{1:T}$ (without any preprocessing) by an integer factor of $D$ and view the discarded samples simply as missing values. Of course, when $D$ is large, i.e. when we throw away many samples, due to aliasing, higher harmonics will overlap with harmonics in the lower frequency band which will cause a more diffuse posterior on the piano-roll, eventually degrading performance.

Although our approach has many desirable features (automatically deducing number of correct notes, high temporal resolution etc.), one of the main disadvantage of our method is computational cost associated with updating large covariance matrices in Kalman filtering. It would be very desirable to investigate approximation schemas that employ fast transformations such as the FFT to accelerate computations.

When transcribing music, human experts rely heavily on prior knowledge about the musical structure – harmony, tempo or expression. Such structure can be captured by training probabilistic generative models on a corpus of compositions and performances by collecting statistics over selected features (e.g. (Raphael and Stoddard 2003)). One of the important advantages of our approach is that such prior knowledge about the musical structure can be formulated as an informative prior on a piano-roll; thus can be integrated in signal analysis in a consistent manner. We believe that investigation of this direction is important in designing robust and practical music transcription systems.

We have not yet tested our model for more general scenarios, such as music fragments containing percussive instruments or bell sounds with inharmonic spectra. Our simple periodic signal model would be clearly inadequate for such a scenario. On the other hand, we stress the fact that the framework presented here is not only limited to the analysis of signals with harmonic spectra, and in principle applicable to any family of signals that can be represented by a switching state space model. This is already a large class since many real-world acoustic processes can be approximated well with piecewise linear regimes. We can also formulate a joint estimation schema for unknown parameters and integrate them out (e.g. see (Davy and Godsill)). However, this is currently a hard and computationally expensive task. If efficient and accurate approximate integration methods can be developed, our model will be applicable to mixtures of many different types of acoustic signals and may be useful in more general auditory scene analysis problems.
References


