Bayesian Statistical Methods for Audio and Music Processing

A. Taylan Cemgil, Simon J. Godsill, Paul Peeling, Nick Whiteley
Signal Processing and Comms. Lab, University of Cambridge
Department of Engineering, Trumpington Street, Cambridge, CB2 1PZ, UK
{atc27,sjg}@eng.cam.ac.uk

August 15, 2008

Abstract

Bayesian statistical methods provide a formalism for arriving at solutions to various problems faced in audio processing. In real environments, acoustical conditions and sound sources are highly variable, yet audio signals often possess significant statistical structure. There is a great deal of prior knowledge available about why this statistical structure is present. This includes knowledge of the physical mechanisms by which sounds are generated, the cognitive processes by which sounds are perceived and, in the context of music, the abstract mechanisms by which high-level sound structure is compiled. Bayesian hierarchical techniques provide a natural means for unification of these bodies of prior knowledge, allowing the formulation of highly-structured models for observed audio data and latent processes at various levels of abstraction. They also permit the inclusion of desirable modelling components such as change-point structures and model-order specifications.

The resulting models exhibit complex statistical structure and in practice, highly adaptive and powerful computational techniques are needed to perform inference. In this chapter, we review some of the statistical models and associated inference methods developed recently for audio and music processing. Our treatment will be biased towards musical signals, yet the modelling strategies and inference techniques are generic and can be applied in a broader context to nonstationary time series analysis. In the chapter we will review application areas for audio processing, describe models appropriate for these scenarios and discuss the computational problems posed by inference in these models. We will describe models in both the time domain and transform domains, the latter typically offering greater computational tractability and modelling flexibility at the expense of some accuracy in the models. Inference in the models is performed using Monte Carlo methods as well as variational approaches originating in statistical physics. We hope to show that this field, which is still in its infancy compared to topics such as computer vision and speech recognition, has great potential for advancement in coming years, with the advent of powerful Bayesian inference methodologies and accompanying computational power increases.

1 Introduction

In applications that need to deal with acoustical and computational modelling of sound, a fundamental obstacle is *superposition*, i.e. concurrent sound events (polyphonic music, speech or environmental sound) are mixed and altered due to reverberation present in the acoustic environment. In speech processing, this problem is referred to as the *cocktail party* problem.

In hearing aids, undesired structured environmental sources, such as wind or machine noises, contaminate the target sound and need to be filtered out; here the objective is *denoising* or perceptual enhancement. A similar situation happens in polyphonic music, where several instruments play simultaneously and one goal is to separate or identify the individual voices. In all of these domains, due to superposition, information about individual sources cannot be directly extracted, and significant focus is given in the literature to source separation, deconvolution and perceptual organisation of sound (Wang and Brown 2006).

Acoustic processing is a rather broad field and the research is driven by both scientific and technological motivations – two related but distinct goals. For technological needs, the primary motivation is to develop practical engineering solutions to enhance recognition, denoising, source separation or information retrieval. The ultimate goal here is to construct computer systems that display aspects of human level performance in automated sound understanding. In the second, the goal is scientific understanding of cognitive processes behind the human auditory system and the physical sound generation process of musical instruments or voices.

Our starting point in this article is that in both contexts, scientific or technological, Bayesian statistical methods provide a formalism to make progress. This is achieved via models which quantify prior knowledge about physical properties and semantics of sound and powerful computational techniques. The key equation, then, is Bayes' theorem and in the context of audio processing it can be stated as

$$p(\text{Structure}|\text{Audio Data}) \propto p(\text{Audio Data}|\text{Structure})p(\text{Structure})$$

Thus inference is drawn from the posterior distribution over hidden structure given observed audio data. The strength of this simple and abstract view of audio processing is that it admits a variety of tasks such as tracking, restoration, transcription, separation, identification or resynthesis can be formulated as Bayesian inference problems. The approach also inherits the benefit common to all applications of Bayesian statistical methods that the problem formulation and computational solution strategy are well separated. This differs significantly from heuristic and ad-hoc approaches to audio processing which have been popular historically and which involve the design of custom-built algorithms for solving specific tasks where problem formulation and computational solution are mixed, taking account of practical and pragmatic considerations.

1.1 Introduction to Musical Audio

The following discussion gives a basic introduction to some of the properties of musical audio signals. The discussion follows closely that of (Godsill 2004). Musical audio is highly *structured*, both in the time domain and in the frequency domain. In the time domain, *tempo* and *beat* specify the range of likely note transition times. In the frequency domain, two levels of structure can be considered. First, each note is composed of a fundamental frequency (related to the 'pitch' of the note), and partials whose relative amplitudes determine the timbre of the note. This frequency domain description can be regarded as an empirical approximation to the true process, which is in reality a complex non-linear time-domain system (McIntyre, Schumacher, and Woodhouse 1983; Fletcher and Rossing 1998). The frequencies of the partials are approximately integer multiples of the fundamental frequency, although this clearly doesn't apply for instruments such as bells and tuned percussion. Second, several notes played at the same time form chords, or polyphony. The fundamental frequencies of each note comprising a chord are typically related by simple multiplicative rules. For example, a C major chord may be composed of the frequencies 523 Hz, 659Hz $\approx 5/4\times$523 Hz and 785 Hz $\approx 3/2\times$523 Hz. Figure 2 shows a time-frequency spectrogram analysis for a simple monophonic (single note) flute recording (this may be auditioned at `www-sigproc.eng.cam.ac.uk/~sjg/haba`, where other extracts used in

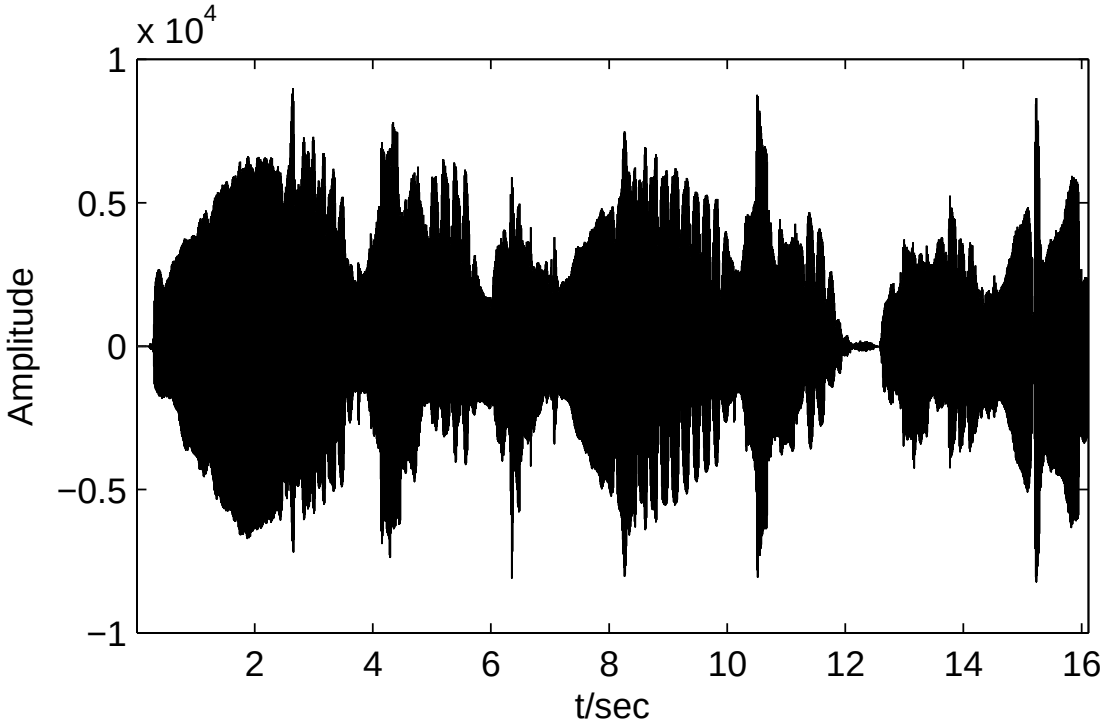


Figure 1: Time-domain waveform for a solo flute extract

this paper may also be listened to[1]), corresponding to the waveform displayed as Figure 1. In this both the temporal segmentation and the frequency domain structure are clearly visible on the plot. Focusing on a single localised time frame, at around 2s in the same extract, we can clearly see the fundamental frequency component, labelled ω_0, and the partial stucture, at frequencies $2\omega_0$, $3\omega_0$, ...of a single musical note in figure 1.1. It is clear from spectra such as figure 1.1 that it will be possible to estimate the pitch (we will refer to pitch interchangeably with ω_0, although it should be noted that *perceived* pitch is a more complex function of the fundamental and all partials) and partial information (amplitudes of partials, number of partials, etc.) from single-note data that is well segmented in time (so that there is not significant overlap between more than one separate musical note within any single segment). There are many ways to achieve this, based on sample autocorrelation functions, spectral peak locations, etc. Of course, real musical extracts don't usually arrive in conveniently segmented single note form, and much more complex structures need to be considered.

1.2 Applications

There are many tasks of interest for musical analysis in which computers can be of assistance, including (but not limited to):

1. Music-to-score transcription. This involves the analysis of raw audio signals to produce a musical 'score' representation. This is one of the most challenging and comprehensive tasks facing us in computational music analysis, and one that is certainly ill-defined, since there are many possible written scores corresponding to one performance. An expert human listener could transcribe a relatively complex piece of musical audio, but the score produced would be dissimilar in many respects to that of the composer. However, it would be reasonable to hope that the transcriber could generate a score having similar pitches

[1]Web page not yet generated - sorry!

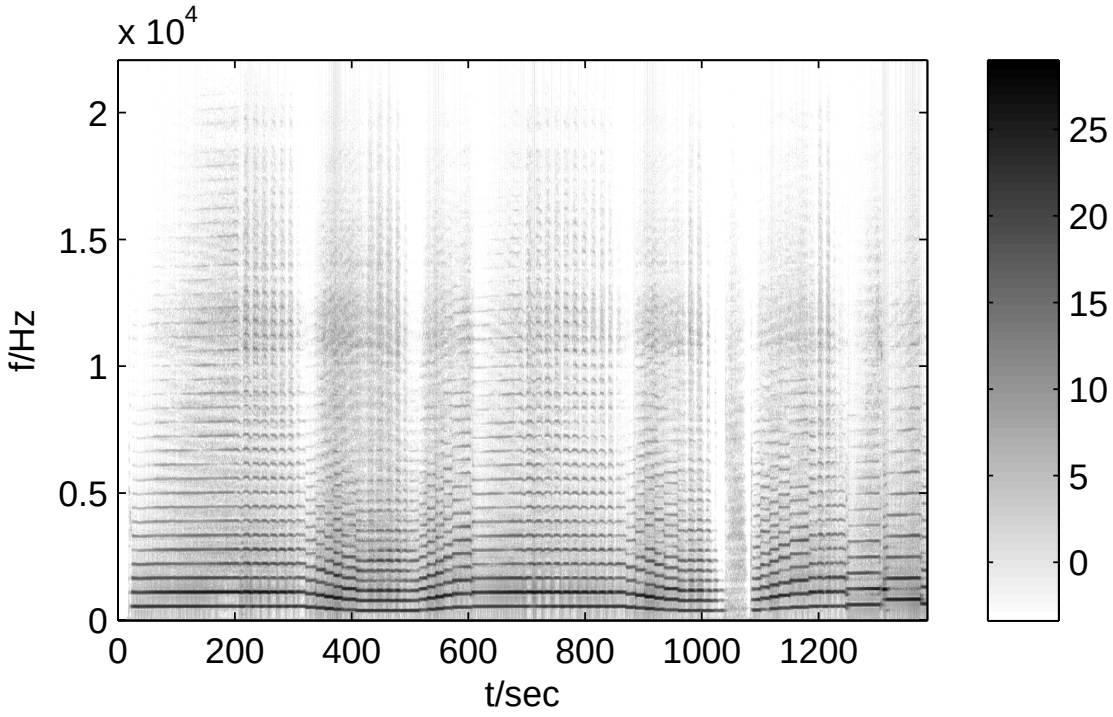


Figure 2: Time-frequency spectrogram representation for the flute recording

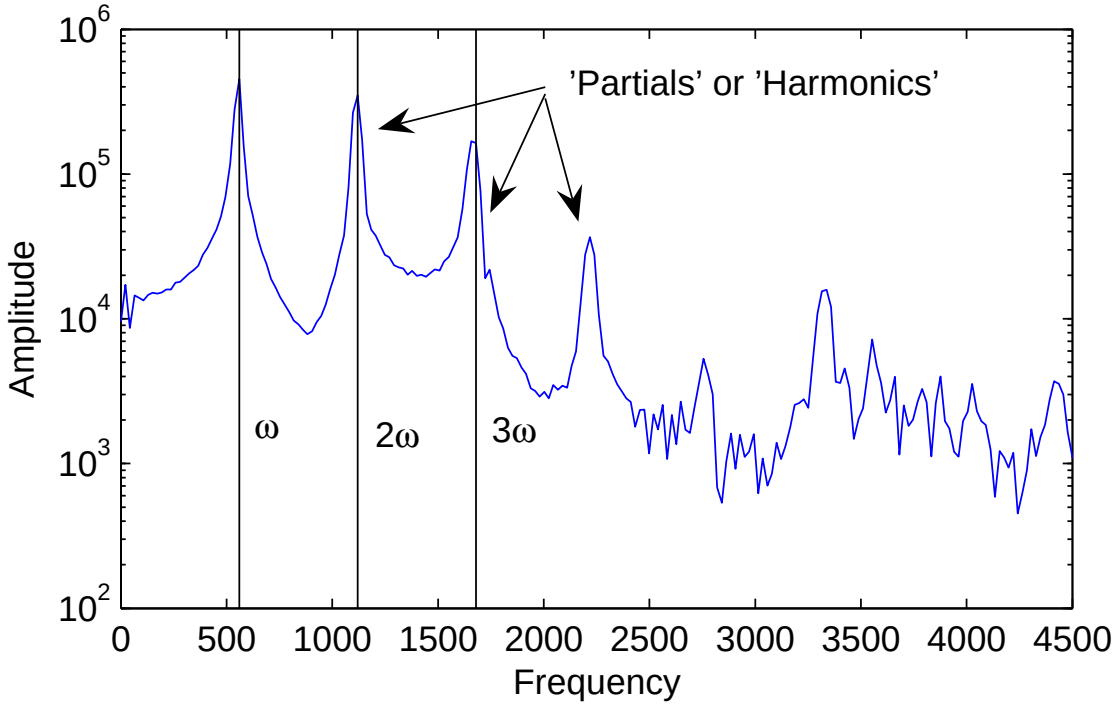


Figure 3: Short-time Fourier analysis of a single frame of data from the flute extract

and durations to those of the composer. The sub-task of generating a pitch-and-duration map of the music is the main aim of many so-called 'transcription' systems. Others have considered the task of score generation from this point on and software is available commercially for this highly subjective part of the process - we will not consider it further here. Applications that require the transcription task include analysis of ethnomusicological recordings, transcription of jazz and other improvised forms for analysis or publication of performance versions, and transcriptions of rare or historical pieces which are no longer available in the form of a printed score. Apart from applications which directly require the full transcription, there are many applications, for example those below, which are fully or partially solved as a result of a solution to the transcription problem.

2. Instrument classification is an important component of musical analysis systems, i.e. the task of recognising which instruments are playing at any given time in a piece
3. A related concept is timbre determination - extraction of the tonal character of a pitched musical note (in coarse terms, is it harsh, sweet, bright, etc.
4. Signal separation - here we attempt to separate out individual instruments or notes from a polyphonic (many-note) mixture. This finds application in many areas from sound remastering in the recording studio through to Karaoke (extraction of a principal vocal line from a source, leaving just the accompaniment). Source separation finds much wider application of course in non-musical audio, especially in source separation for hearing aids, see below.
5. Audio restoration and enhancement. In this application the quality of an audio source is enhanced, for example by reduction of background noise. This task comes as a by-product of many model-based analysis tasks, such as source separation above, since a noise-reduced version of the input signal will often be available as one of the possible inferences from the Bayesian posterior distribution.

The fundamental tasks above will find use in many varied acoustical applications. For example, with vast amount of audio data available digitally in on-line repositories, it is not reasonable to predict that almost all audio material will be available digitally in the near future. This has rendered automated processing of audio for sorting and choice of musical content an important and central information processing task, affecting literally millions of end users. For flexible interaction, it is essential that systems are able to extract structure and organize information from the audio signal directly. Our view is that the associated fundamental computational problems require both a fresh look at existing signal processing techniques and development of novel statistical methodology.

In addition, Computer based music composition and sound synthesis date back to the first days of digital computation. However, despite recent technological advances in synthesis, compression, processing and distribution of digital audio, it has yet been not possible to construct machines that can simulate the effectiveness of human listening.

Statistical methodolgies are now migrating into human computer interaction, computer games and electronic entertainment computing. Here, one ambitious research goal focuses on computational techniques to equip computers with musical listening and interaction capabilities. This is essential in construction of intelligent music systems and virtual musical instruments that can listen, imitate and autonomously interact with humans. For flexible interaction, it is essential that music systems are aware of the actual content of the music, are able to extract structure and organise information directly from acoustic input. For generating convincing performances, they need to be able to analyse and mimic master musicians.

Another vitally important application area for millions of people is hearing aids, which directly benefits from efficient and robust methods for recognition and source separation (Hamacher, Chalupper, Eggers, Fischer, Kornagel, Puder, and Rass 2005). It is estimated that there are

almost nine million hearing impaired people in the UK alone; a number which is believed to be increasing with rapidly aging population. Progress in this field is likely to improve the quality of life for a sizeable segment of society. Recently, modern hearing aids have evolved into powerful computational devices and with advances in wireless communications, it is becoming now feasible to delegate computation to external portable computing devices. This provides unprecedented possibilities along with interesting computational challenges for online adaptation, since in the next generation of hearing aids, it will be feasible to run sophisticated statistical signal processing and machine learning algorithms. Finally, computational audio processing finds application in the areas of monitoring, rescue and surveillance, computer aided music education, musicology, music perception and cognition research.

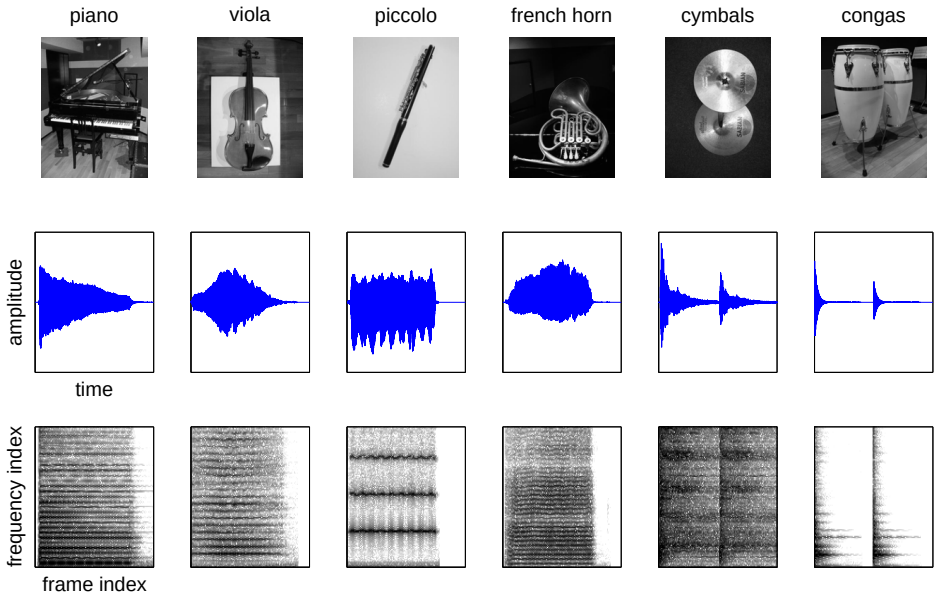


Figure 4: Some acoustical instruments, examples of typical time series and corresponding spectrograms (time varying magnitude spectra – modulus of short time Fourier transform) computed with FFT. (Audio data and images from RWCP Instrument samples database).

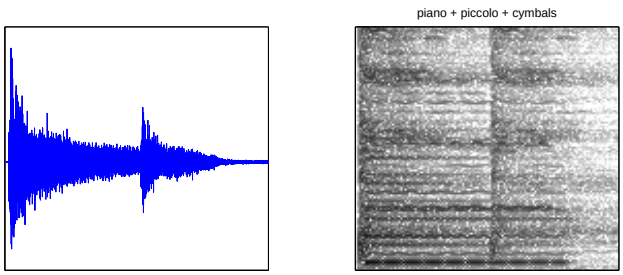


Figure 5: Superposition. The time series and the magnitude spectrogram of the resulting signal when some of the instruments play concurrently.

2 Fundamental Audio Processing Tasks

From the above discussion of the challenges facing audio processing, some fundamental tasks can be identified for treatment by Bayesian techniques. Firstly, we can hope to address the superposition task in a model-based fashion, by posing models that capture the behaviour of superimposed signals. These are similar in flavour to the latent factors analysed in some statistical modelling problems. A generic model for observed data Y, under a linear superposition assumption, will then be:

$$Y = \sum_{i=1}^{I} s_i \tag{1}$$

where the s_i represent each of the I individual audio sources present. We pose this very basic model here as a single-channel observation model, although it is straightforward to extend the model to the multi-channel case, in which case it will be usual to include also channel-specific mixing coefficients. The sources and data will typically be audio time series, but can also represent expansion coefficients of the audio in some other domain such as the Fourier or wavelet domain, as will be made clear in context later. We may make the model a little more sophisticated by making the data a stochastic function of the sources, and in this case we will specify some non-degenerate likelihood function $p(Y|\sum_{i=1}^{I} s_i)$.

We typically assume that the individual sources s_i, one or more of which may be background noise terms, are independent *a priori*. They are parameterised by θ_i, which represent information about the sound generation process for that particular source, including perhaps its pitch and other characteristics (number of partials, etc.), encoded through a conditional distribution and prior distribution for each source:

$$p(s_i, \theta_i) = p(s_i|\theta_i)p(\theta_i)$$

Dependence between the θ_i, for example to model the harmonic relationships of notes within a chord, can of course be included as desired when considering the joint distribution of sources and parameters. To this model we can add unknown hyperparameters Λ with prior $p(\Lambda)$ in the usual way, and incorporate model uncertainty through an additional prior distribution on the number of components I. The specification of suitable source models $p(s_i|\theta_i)$ and $p(\theta_i)$, as well as the form of likelihood function $p(Y|\sum_{i=1}^{I} s_i)$, will form a substantial part of the remainder of the paper.

Several fundamental inference tasks can then be identified from this generic model, including the source separation and polyphonic music transcription tasks identified above.

2.1 Source Separation

In source separation, the task is to infer the *source signals* s_i themselves, given the *observed signal* Y. Collecting the sources together as $S = \{s_i\}_{i=1}^{I}$ and the parameters as $\Theta = \{\theta_i\}_{i=1}^{I}$, the Bayesian formulation of the problem can be stated, under a fixed number of sources I, as (see for example (Mohammad-Djafari 1997; Knuth 1998; Rowe 2003; Févotte and Godsill 2006; Cemgil, Godsill, and Fevotte 2007))

$$p(S|Y) = \frac{1}{P(Y)} \int p(Y|S, \Lambda)p(S|\Theta, \Lambda)p(\Lambda)p(\Theta)d\Lambda d\Theta \tag{2}$$

where, under our deterministic model above in Eq. 1 the likelihood function $p(Y|S, \Lambda)$ will be degenerate. The marginal likelihood $P(Y)$ plays a key role when model order uncertainty is to be incorporated into the problem, for example when the number of sources N is unknown and needs to be estimated (Miskin and Mackay 2001).

Additional considerations which may additionally be included in the above framework include convolutive (filtered) and non-stationary mixing of the sources - both scenarios are of practical interest and still pose significant computational challenges. Once the posterior distribution is computed by evaluating the integral, point estimates of the sources can be obtained using suitable estimation criteria, such as marginal MAP or posterior mean estimation, although in the latter case one has to be especially careful with the interpretation of expectations in models where likelihoods and priors are invariant to source permutations.

2.2 Polyphonic Music Transcription

Music transcription refers to extraction of a human readable and interpretable description from a recording of a music performance, see Fig. 6. In cases where more than a single musical note plays at a given time instant, we term this task *polyphonic music transcription*. Interest in this problem is largely motivated by a desire to implement a program to infer automatically a musical notation, such as the traditional western music notation, listing the pitch values of notes, corresponding timestamps and other expressive information in a given performance. These quantities will be encoded in the above model through the parameters θ_i of each note present at a given time. Simple models will encode only the pitch of the note in θ_i, while more complex models can include expressive information, instrument-specific characteristics and timbre, etc.

Apart from being an interesting modelling and computational problem in its own right, automated extraction of a score-like description is potentially very useful in a broad spectrum of applications such as interactive music performance systems, music information retrieval and musicological analysis of musical performances, not to mention as an aid to the source separation task identified above. However, in its most unconstrained form, i.e., when operating on an arbitrary acoustical input, music transcription remains a very challenging problem, owing to the wide variation in acoustical conditions and characteristics of musical instruments. In spite of these difficulties, a practical engineering solution is possible by careful incorporation of prior knowledge from cognitive science, musicology, musical acoustics, and by use of computational techniques from statistics and digital signal processing.

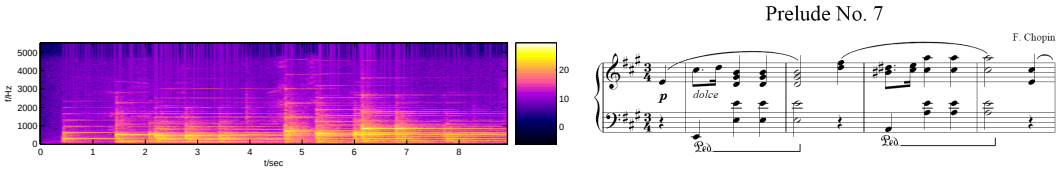


Figure 6: Polyphonic Music Transcription. The task is to generate a human readable score as shown below, given the acoustic input. The computational problem here is to infer pitch, number of notes, rhythm, tempo, meter, time signature. The inference can be achieved online (filtering) or offline (smoothing), depending upon requirements.

2.3 Hierarchical Models for Musical Audio

In a statistical sense, music transcription is an inference problem where, given a signal, we want to find a score that is consistent with the encoded music. In this context, a score can be contemplated as a collection of "musical objects" (e.g., note events) that are rendered by a performer to generate the observed signal. The term "musical object" comes directly from an analogy to visual scene analysis where a scene is "explained" by a list of objects along with a

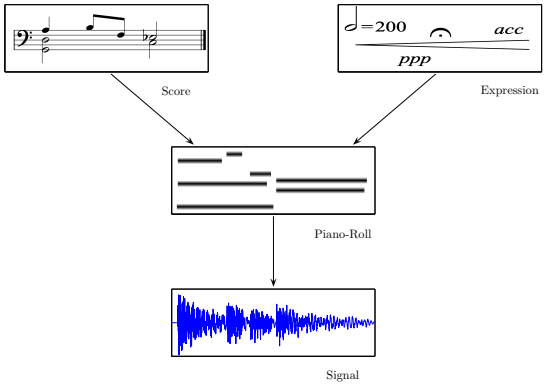


Figure 7: A hierarchical generative model for music transcription. In this model, an unknown score is rendered by a performer into a piano-roll. The performer introduces expressive timing deviations and tempo fluctuations. The piano-roll is rendered into audio by a synthesis model. The piano roll can be viewed as a symbolic representation, analogous to a sequence of MIDI events. Given the observations, transcription can be viewed as Bayesian inference of the score. Somewhat simplified, the techniques described in this article can be viewed as inference techniques as applied to subgraphs of this graphical model.

description of their intrinsic properties such as shape, color or relative position. We view music transcription from the same perspective, where we want to "explain" individual samples of a music signal in terms of a collection of musical objects where each object has a set of intrinsic properties such as pitch, tempo, loudness, duration or score position. It is in this respect that a score is a high level description of music.

Musical signals have a very rich temporal structure, and it is natural to think of them as being organized in a hierarchical way. At the highest level of this organization, which we may call as the cognitive (symbolic) level, we have a score of the piece, as, for instance, intended by a composer[2]. The performers add their interpretation to music and render the score into a collection of "control signals". Further down on the physical level, the control signals trigger various musical instruments that synthesize the actual sound signal. We illustrate these generative processes using a hierarchical graphical model (See Figure 7), where the arcs represent generative links.

This architecture is of course anything but new, and in fact underlies any music generating computer program such as a sequencer. The main difference of our model from a conventional sequencer is that the links are probabilistic, instead of deterministic. We use the sequencer analogy in describing a realistic generative process for a large class of music signals.

In describing music, we are usually interested in a symbolic representation and not so much in the "details" of the actual waveform. To abstract away from the signal details, we define an intermediate layer, that represents the control signals. This layer, that we call a "piano-roll", forms the interface between a symbolic process and the actual signal process. Roughly, the symbolic process describes how a piece is composed and performed. Conditioned on the piano-roll, the signal process describes how the actual waveform is synthesized. Conceptually, the transcription task is then to "invert" this generative model and recover back the original score. As an intermediate and less sophisticated task, we may try and invert back only as far as the piano-roll.

[2]In reality the music may be improvised and there may be actually not a written score. In this case we replace the generative model with the intentions of the performer, which can still be expressed in our framework as a 'virtual' musical score

3 Signal Models for Audio

We begin the discussion by describing some basic note and chord models for musical audio, based in the time or frequency domain. As already discussed, a basic property of most non-percussive musical sounds is a set of oscillations at frequencies related to the fundamental frequency ω_0. Consider for the moment a short-time frame of musical audio data, denoted $y(\tau)$, in which note transitions do not occur. This would correspond, for example, to the analysis of a single musical chord. Throughout, we assume that the continuous time audio waveform $y(\tau)$ has been discretised with a sampling frequency ω_s rad.s^{-1}, so that discrete time observations are obtained as $y_t = y(2\pi t/\omega_s),\ t = 0, 1, 2, \ldots, N-1$. We assume that $y(\tau)$ is bandlimited to $\omega_s/2$ rad.s^{-1}, or equivalently that it has been prefiltered with an ideal low-pass filter having cut-off frequency $\omega_s/2$ rad.s^{-1}. We will not consider for the moment the time evolution of one chord to the next, or of note changes in a melody. This critical issue is treated in later sections.

The following model for, say, the ith note out of a chord comprising I notes in total can be written as

$$s_{i,t} = \sum_{m=1}^{M_i} \alpha_{m,i} \cos(m\omega_{0,i}t) + \beta_{m,i} \sin(m\omega_{0,i}t) \tag{3}$$

for $t \in \{0, \ldots, N-1\}$. Here, $M_i > 0$ is the number of partials present in note i, $\sqrt{\alpha_{m,i}^2 + \beta_{m,i}^2}$ gives the amplitude of these partials and $\tan^{-1}(\beta_{m,i}/\alpha_{m,i})$ gives the phase of that partial. Note that $\omega_{0,i} \in (0, \pi)$ is here scaled for convenience - its actual frequency is $\frac{\omega_{0,i}}{2\pi}\omega_s$. The unknown parameters for each note are thus $\omega_{0,i}$, the fundamental frequency, M_i, the number of partials and $\alpha_{m,i}$, $\beta_{m,i}$, which determine the amplitude and phase of each partial.

The extension to the multiple note case is then straightforwardly obtained by linear superposition of a number of notes:

$$y_t = \sum_{i=1}^{I} s_{i,t} \ + \ v_t$$

where v_t is a random background noise component (compare this with the deterministic mixture in Eq. 1). In this model v_t will also have to model any residual transient noise from the musical instruments themselves. We now have in addition an unknown parameter I, the number of notes present, plus any unknown statistics of the background noise process.

Such a model is a reasonable approximation for many steady musical sounds, and has quite a lot of analytical tractability, especially if a Gaussian form is assumed for v_t and for the priors on amplitudes α and β. Nevertheless, the posterior distribution is highly non-Gaussian and multimodal, and sophisticated computational tools are required to infer accurately from this model. This was precisely the topic of the work in (Walmsley, Godsill, and Rayner 1998) and (Walmsley, Godsill, and Rayner 1999), where a reversible jump sampler was developed for such a model, under the above-mentioned Gaussian prior assumptions.

The basic form above is however over-idealised in a number of ways: principally from the assumption of constant amplitudes α and β over time, and in the fixed integer relationships between partials, i.e. partial m in note i lies exactly at frequency $m\omega_{0,i}$. The modification of the basic model to remove these assumptions was the topic of our later work (Davy and Godsill 2002; Godsill and Davy 2002; Davy, Godsill, and Idier 2006; Godsill and Davy 2005), still within a reversible jump Monte Carlo framework.[3] In particular, it is fairly straightforward to modify

[3]Editors: would you like me to write a summary of reversible jump in an appendix?

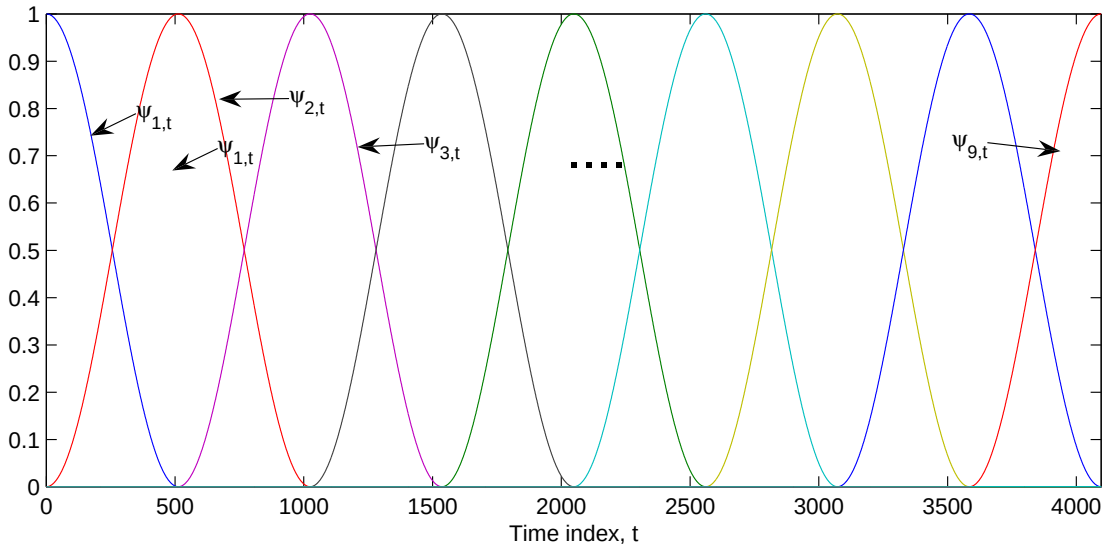


Figure 8: Basis functions $\psi_{i,t}$, $I = 9$, 50% overlapped hanning windows.

the model so that the partial amplitudes α and β vary with time,

$$s_{i,t} = \sum_{m=1}^{M_i} \alpha_{m,i,t} \cos(m\omega_{0,i}t) + \beta_{m,i,t} \sin(m\omega_{0,i}t) \tag{4}$$

and we typically expand $\alpha_{m,i,t}$ and $\beta_{m,i,t}$ on a finite set of smooth basis functions $\psi_{i,t}$ with expansion coefficients a_i and b_i:

$$\alpha_{m,i,t} = \sum_{j=1}^{J} a_i \psi_{i,t}, \quad \beta_{m,i,t} = \sum_{j=1}^{J} b_i \psi_{i,t}$$

In our work we have adopted 50%-overlapped Hamming windows for the basis functions, see Fig. 8, with support either chosen a priori by the user or treated as a Bayesian random variable (Godsill and Davy 2005).

Alternative more general representations allow a fully stochastic variation of $\alpha_{m,i,t}$ in the state-space formulation, see section **??**.

Further idealisations in these models include the assumption of constant fundamental frequencies with time and the Gaussian prior and noise assumptions, but in principle all can be addressed in a principled Bayesian fashion.

3.1 A prior distribution for musical notes

Under the above basic time-domain model we need to assign prior distributions over the unknown parameters for a single note in the mix, currently $\{\omega_{0,i}, M_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ are the vectors of parameters $\alpha_{m,i}, \boldsymbol{\beta}_{m,i}$, $m = 1, 2, ..., M_i$. Under an assumed note system such as an equally-tempered Western note system, we can augment this with a note number index n_i. A suitable scheme is the MIDI note numbering system[4] which labels middle C (or 'C4') as note number 60, and all other notes as integers relative to this - the A below this would

[4]See for example `www.harmony-central.com/MIDI/doc/table2`

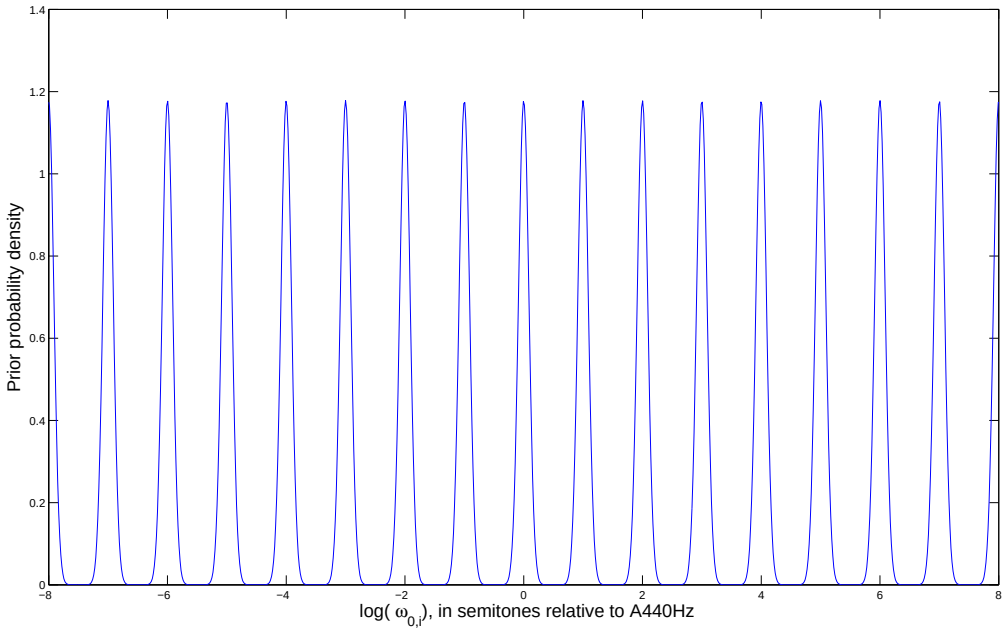


Figure 9: Prior for fundamental frequency $p(\omega_{0,i})$

be 57, for example, and the A above middle C (usually at 440Hz in modern Western tuning systems) would be note number 69. Other non-Western systems could also be encoded within variants of such a scheme. The fundamental frequency would then be expected to lie 'close' to the expected frequency for a particular note number, allowing for performance and tuning deviations from the ideal. Thus a prior for the observed fundamental frequency $\omega_{0,i}$ can be constructed fairly straightforwardly. We adopt here a truncated log-normal distribution for the note's fundamental frequency:

$$p(\log(\omega_{0,i})|n_i) \propto \begin{cases} \mathcal{N}(\mu(n_i), \sigma_\omega^2), & \log(\omega_{0,i}) \in [(\mu(n_i - 1) + \mu(n_i))/2, (\mu(n_i) + \mu(n_i + 1))/2) \\ 0, & \text{otherwise} \end{cases}$$

where $\mu(n)$ computes the expected log-frequency of note number n, i.e., when we are dealing with A440 music in the equally tempered western system,

$$\mu(n) = (n - 69)/12 \log(2) + \log(440/\omega_s) \tag{5}$$

where once again $\omega_s$rad.s^{-1} is the sampling frequency of the data. Assuming $p(n)$ is uniform for now, the resulting prior $p(\omega_{0,i})$ is plotted in Fig. 9, capturing the expected clustering of note frequencies at semitone spacings relative to A440.

The prior model for a note is completed with two components. Firstly a prior for the number of partials, $p(M_i|\omega_{0,i})$, is specified as uniform over the range $\{M_{\min}, ..., M_{\max}\}$, with limits truncated to prevent partials at frequencies greater than $\omega_s/2$, the Nyquist rate. Secondly a prior for the amplitude parameters $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ must be specified. This turns out to be quite crucial to the modelling performance and here we initially proposed a Gaussian form. It is expected however that partials at high frequencies will have lower energy than those at high frequencies, generally following a low-pass filter shape in the frequency domain. Coefficents $\alpha_{m,i}$ and $\beta_{m,i}$ are then assigned independent Gaussian prior distributions such that their amplitudes are assumed

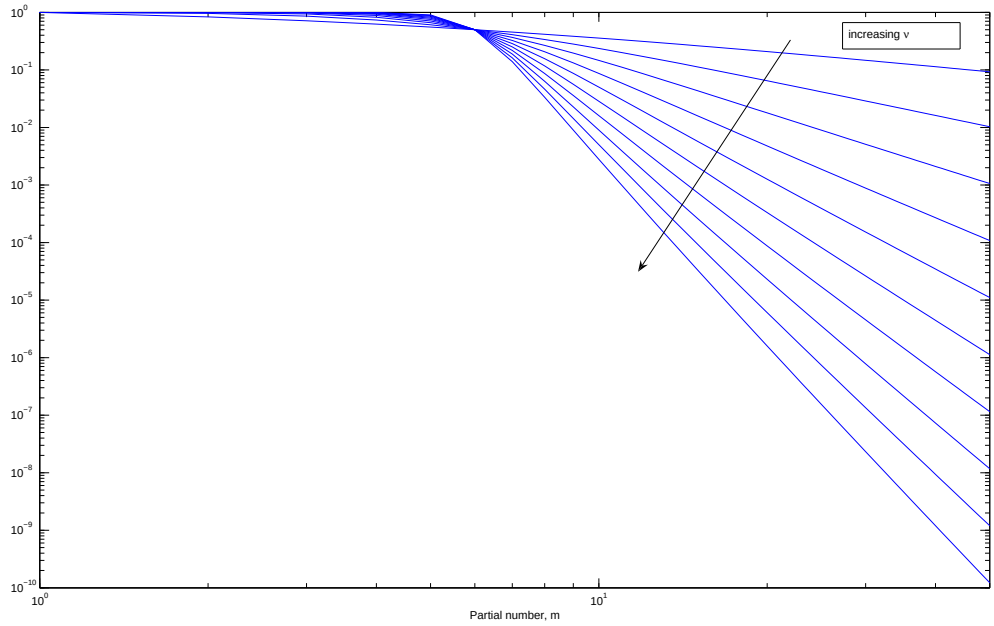


Figure 10: Family of k_m curves (log-log plot), $T = 5$, $\nu = 1, ..., 10$.

to decay with increasing frequency of the partial number m. The general form of this is

$$p(\alpha_{m,i}, \beta_{m,i}) = \mathcal{N}(\beta_{m,i}|0, g_i^2 k_m)\mathcal{N}(\alpha_{m,i}|0, g_i^2 k_m)$$

Here g_i is a scaling factor common to all partials in a note and k_m is a frequency-dependent scaling factor to allow for the expected decay with increasing frequency for partial amplitudes. Following (Godsill and Davy 2005) the amplitudes are assumed to decay as follows:

$$k_m = 1/(1 + (Tm)^\nu)$$

where ν is a decay constant and T determines the cut-off frequency. Such a model is based on empirical observations of the partial amplitudes in many real instrument recordings, and essentially just encodes a low pass filter with unknown cut-off frequency and decay rate. See for example the family of curves with $T = 5$, $\nu = 1, 2, ..., 10$, Fig. 10. It is worth pointing out that this model does not impose very stringent constraints on the precise amplitude of the partials: the Gaussian distribution will allow for significant departures from the $k_m = 1/(1 + (Tm)^\nu)$ rule, as dictated by the data, but it does impose a generally low-pass shape to the harmonics across frequency. It is possible to keep these parameters as unknowns in the MCMC scheme (see (Godsill and Davy 2005)), although in the examples presented here we fix these to appropriately chosen values for the sake of computational simplicity. g_i, which can be regarded as the overall 'volume' parameter for a note, is treated as an additional random variable, assigned an inverted Gamma distribution for its prior. The Gaussian prior structure outlined here for the α and β parameters is readily extended to the time-varying amplitude case of Eq. (4), in which case similar Gaussian priors are applied directly to the expansion coefficients a and b, see (Davy, Godsill, and Idier 2006).

In the simplest case, a polyphonic model is then built by taking an independent prior over

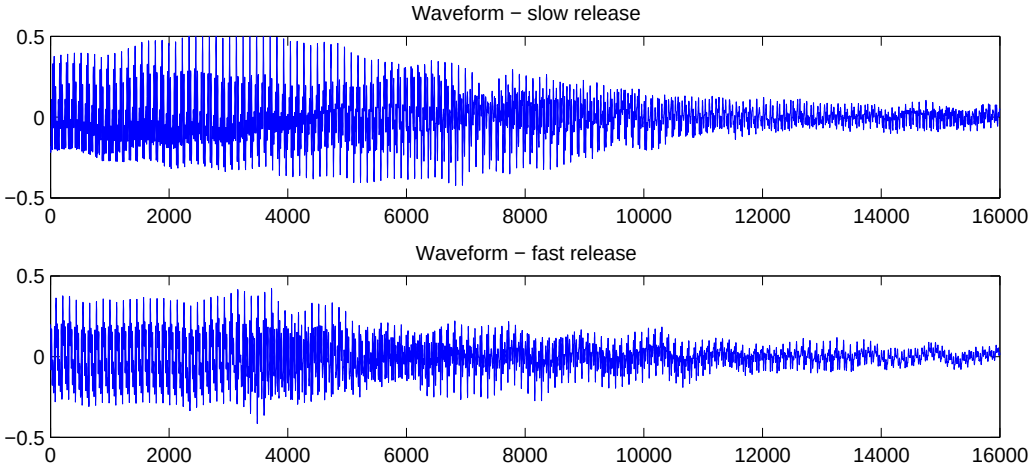


Figure 11: Waveforms for release transient on pipe organ. Top: slow release; bottom: fast release.

the individual notes and the number of notes present:

$$p(\Theta) = p(I) \prod_{i=1}^{I} p(\theta_i)$$

where

$$\theta_i = \{n_i, \omega_{0,i}, M_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, g_i\}$$

This model can be explored using MCMC methods, in particular the reversible jump MCMC method (Green 1995), and results from this and related models can be found in (Godsill and Davy 2005; Davy, Godsill, and Idier 2006). In later sections, however, we discuss simple modifications to the generative model in the frequency domain which render the computations much more feasible for large polyphonic mixtures of sounds.

The models of this section provide a quite accurate time-domain description of many musical sounds. The inclusion of additional effects such as inharmonicity and time-varying partial amplitudes (Godsill and Davy 2005; Davy, Godsill, and Idier 2006) makes for additional realism.

3.2 Example: musical transient analysis with the harmonic model

A useful case in point is the analysis of musical transients, i.e. the start or end of a musical note, when we can expect rapid variation in partial amplitudes with time. Here we take as an example a pipe organ transient, analysed under different playing conditions: one involving a rapid release at the end of the note, and the other involving a slow release, see Fig. 11. There is some visible (and audible) difference between the two waveforms, and we seek to analyse what is being changed in the structure of the note by the release mode. Such questions are of interest to acousticians and instrument builders, for example.

We analyse these datasets using the prior distribution of the previous section and the model of Eq. (4). A fixed length hanning window of duration 0.093s was used for the basis functions. The resulting MCMC output can be used in many ways. For example, examination of the expansion coefficients a_i and b_i allows an analysis of how the partials vary with time under each playing condition. In both cases the reversible jump MCMC identifies 9 significant partials in the data. In Figs. 12 and 13 we plot the first five ($m = 1, ..., 5$) partial energies $a_{m,i}^2 + b_{m,i}^2$ as a function of time.

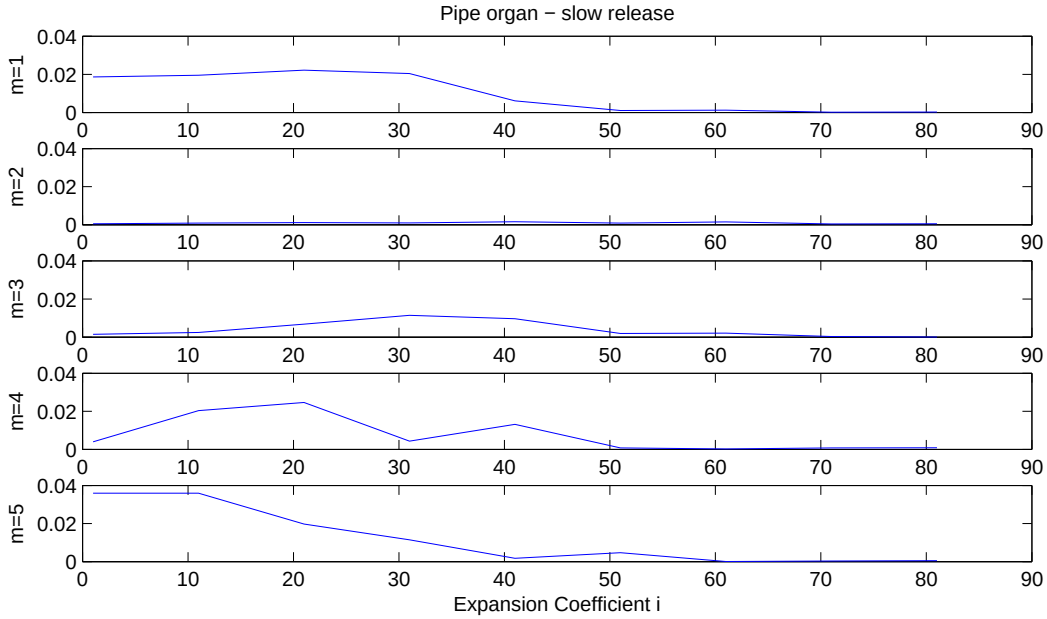


Figure 12: Magnitudes of partials with time: slow release.

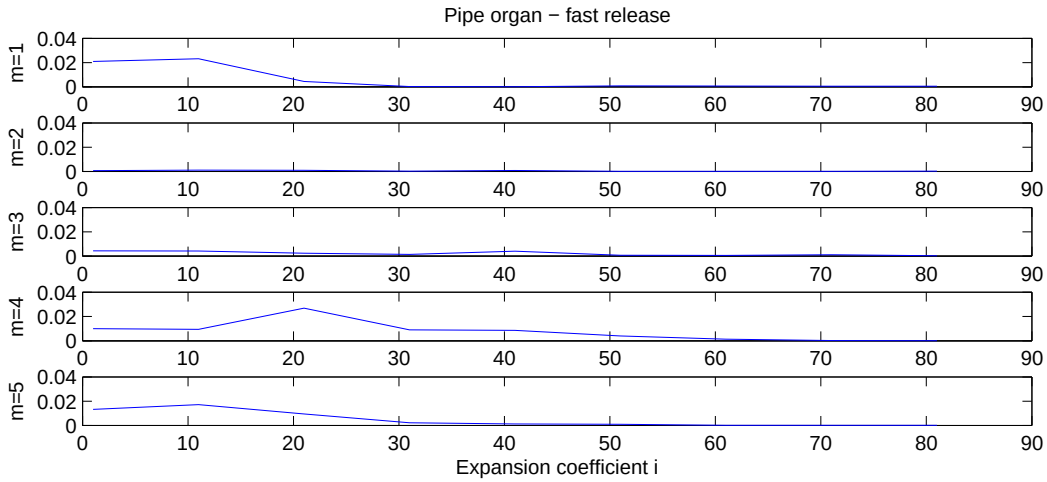


Figure 13: Magnitudes of partials with time: fast release.

Examining the behaviour from the MCMC output we can see that the third partial is substantially elevated during the slow release mode, between coefficients $i = 30$ to 40. Also, in the slow release mode, the fundamental frequency ($m = 1$) decays at a much later stage relative to, say, the fifth partial, which itself decays more slowly in that mode. One can also use the model output to perform signal modification; for example time stretching or pitch shifting of the transient are readily achieved by reconstructing the signal using the MCMC-estimated parameters but modifying the hanning window basis function length (for time-stretching) or reconstructing with modified fundamental frequency ω_0, see `www-sigproc.eng.cam.ac.uk/~sjg/haba`. The details of our reversible jump MCMC scheme are quite complex, involving a combination of specially designed independence Metropolis-Hastings proposals and random walk-style proposals for the note frequency variables. In the frequency-domain models described in section 5 we use essentially the same MCMC scheme, with simpler likelihood functions - some more details of the proposals used are given there.

4 Dynamical State Space Models

The models introduced in the previous section are generalised linear models, where the expansion coefficients a_i and b_i are assumed to be a-priori independent across frames. While these models are quite useful, they are not very realistic models of the underlying physics, as audio is essentially the result of unfolding dynamical processes. It is possible to introduce random walk dynamics on the expansion coefficients. In contrast, we will describe the evolution of the expansion coefficients in state space form, via state space models. Such representations can be derived from well known sinusoidal signal representations such described in the appendix A and are one step towards physical models.

The audio signal can be written as a sum of exponentially decaying or windowed sinusoids (See Eq. (3) or examples in appendix sections A.2 and A.3). However, richer processes with more complex behaviour can be described by state space models. This involves defining stochastic signal representations and higher level, unobserved stochastic elements, such as change point processes and model-order indicators, which are combined to form hierarchical Bayesian models. The reader will notice that these models exhibit a variety of structures that arise from the interaction between the low-level signal models and the high-level latent processes. We also describe associated inference tasks and efficient methods for their solution which take advantage of the model structures.

4.1 Conditionally Linear Dynamical Systems with regime switching

We start this section with a general framework to highlight and unify the basic modelling ideas. Our goal is to construct a model that can mimic the qualitative behaviour of acoustical systems (such as musical instruments shown in figure 4) yet still has some analytical structure which allows efficient inference. Our starting point is the conditionally-linear state space model. This is motivated by the fact that the harmonic structure of audio signals, arising the physical processes by which they are generated, can conveniently be formulated in state-space form.

Here, we construct a generic model as a cascade of two systems with an excitation (e.g., vibrating string) feeding into a resonator (e.g., the body of an acoustic instrument). Respectively,

Figure 14: A damped oscillator in state space form.

s^e and s^f are the states of the excitation system and the resonating system:

$$\begin{aligned} s_k^e &\sim \mathcal{N}\left(s_k^e; A_k s_{k-1}^e, Q_k\right) \\ s_k^f &= A_{\text{ft}} s_{k-1}^f + B_{\text{ft}} s_k^e \\ \bar{y}_k &\sim \mathcal{N}\left(\bar{y}_k; C s_k^f, R\right) \end{aligned} \tag{6}$$

where C is an observation matrix and R is the observation noise. This model is, of course, a particular parametrisation of the general linear dynamical system. The main idea, in the general sense, is to define a nonstationary process with $p(\bar{A}, \bar{Q})$ over the sequence of transition matrices $\bar{A} \equiv \{A_k\}_{k\geq 0}$ and the transition noise covariances $\bar{Q} \equiv \{Q_k\}_{k\geq 0}$. The posterior estimates of these latent parameters, when integrated over latent states s_k will describe the signal in a compact way. There are clearly many possibilities in defining prior distributions over $\bar{A}$ and $\bar{Q}$. In the sequel, we will define several realistic models in this framework.

We define a sequence of discrete switch variables $r = r_{0:K-1}$ and draw the state matrices conditionally. Here this indicators r are abstract, but in practice will correspond to onsets, offsets or note labels, depending upon the context of the task at hand.

$$\begin{aligned} p(r) &= p(r_{0:K-1}) = p(r_0) \prod_{k=1}^{K-1} p(r_k|r_{k-1}) \\ A_k &\sim p(A_k|r_k) \qquad\qquad Q_k \sim p(Q_k|r_k) \end{aligned}$$

This includes the special cases such as when $A_k = A(r)$, i.e. a deterministic function of r (with the choice $p(A_k|r) = \delta(A_k - A(r))$), similar with Q_k.

In principle, one could work with in any state space coordinate system by appropriately choosing the state matrices. However, we prefer to work in a representation to maintain the interpretability of the parameters. This representation is closely related to the sinusoidal models and described in detail in the appendixA.

4.1.1 Dynamic Harmonic model

To highlight our specific construction, we start this section with an example. We consider a second order oscillator systems, driving a second order resonator, following Eq.(6). We specify the models by a specific choice of transition matrices and transition noise covariance

$$A_k = \tilde{Z}(\gamma_k, \omega_k) \qquad A_{\text{ft}} = \tilde{Z}(\gamma_{\text{ft}}, \omega_{\text{ft}}) \qquad Q_k = q_k I$$

where

$$\tilde{Z}(\gamma, \omega) \equiv e^{-L\gamma_\nu} \begin{pmatrix} \cos(L\omega_\nu) & -\sin(L\omega_\nu) \\ \sin(L\omega_\nu) & \cos(L\omega_\nu) \end{pmatrix}^\top$$

We have a $L \times L$ observation noise covariance matrix R, and an $L \times 2K$ observation matrix C, where each column is a damped sinusoidal (see Eq.(32) in the appendix). For simplicity, we consider the case when frame length is $L = 1$, where we generate the signal sample by

sample. See Fig.14. Note that in this representation the state vector s is simply the expansion coefficients (real amplitudes) a and b, introduced in the previous section.

The transition matrix A_k generates a damped sinusoidal which is fed into the system with transition matrix A_{ft} via the 2×2 input matrix B_{ft}, here taken as $B_{\text{ft}} = I$. The driving noise of the excitation has time dependent variance $Q_k = q_k I$. The observation matrix in this case is $C = \begin{pmatrix} 1 & 0 \end{pmatrix}$. The discrete variables r_k in this model encode onsets and offsets. We define a Markov chain r_k for $k = 0, 1, \ldots$, where $r_k \in \{\text{on} = 1, \text{off} = 0\}$, with the state transition distribution parametrised as $p(r_k = \text{on} | r_{k-1} = \text{on}) = \pi_{\text{on}}$ and $p(r_k = \text{off} | r_{k-1} = \text{off}) = \pi_{\text{off}}$. Conditioned on r_k and r_{k-1}, we let

$$q_k = \begin{cases} q_{\text{onset}} & \text{if } r_k = \text{on and } r_{k-1} = \text{off} \\ q_{\text{on}} & \text{if } r_k = \text{on and } r_{k-1} = \text{on} \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma_k = \begin{cases} \gamma_{\text{on}} & \text{if } r_k = \text{on} \\ \gamma_{\text{off}} & \text{if } r_k = \text{off} \end{cases}$$

The hyper parameters of this model are the prior state probabilities π_{on} and π_{off} transition variances q_{onset}, q_{on}, the excitation model transition damping constants γ_{on}, γ_{off} and frequency ω_{ex}, the resonator parameters γ_{ft} and ω_{ft} and the observation noise variance R. With each new onset event ($r_{k-1,k} = (\text{off}, \text{on})$)), the excitation state is reinitialised from a Gaussian with variance q_{onset} and driven with noise until the next onset time.

Even this rather simple model displays quite rich behaviour, as shown by typical realisations in Fig. 4.1.1. Given the observed signals, posterior inference for latent variables gives structural information about the signal. For example, $\mathsf{E}(\omega_k | \bar{y}_{1:k})$ will give an online estimate of the instantaneous frequency of the excitation and the posterior transition variance estimate $\mathsf{E}(r_k | \bar{y}_{1:K})$ will give an indication of onsets and offsets.

4.2 Dynamic Harmonic model and Changepoint models

As previously discussed, acoustic systems, and in particular pitched musical instruments tend to create oscillations with modes at frequencies that are roughly related by ratios of integers (Fletcher and Rossing 1998). The fundamental frequency, corresponding the largest common divisor of mode frequencies, is strongly correlated with the perceived pitch in music. For transcription, we need to estimate the fundamental frequency as well as the onsets and offsets to mark the beginning and the end of each note. This problem can be formalised using the harmonic models introduced in section 3, coupled to a change-point structure related to the mixture model of the previous section. We now combine a number of oscillators that are harmonically related by a fundamental frequency. We define the block diagonal state evolution matrix

$$A_k = \mathbf{diag}(Z_{0,k}, \ldots, Z_{\nu,k}, \ldots Z_{W-1,k})$$

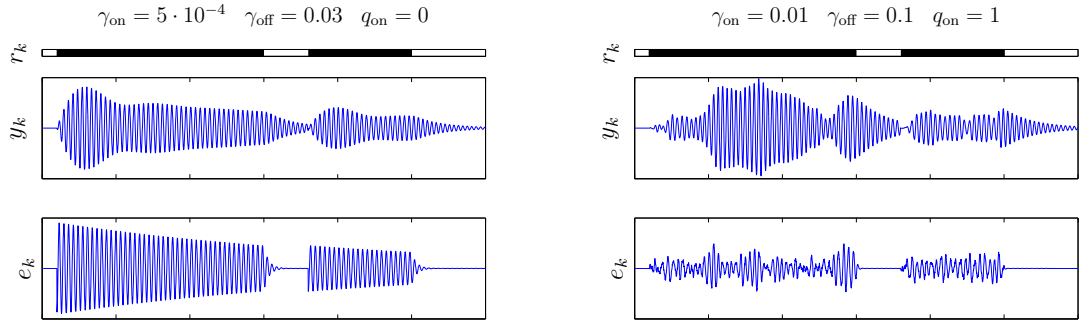
with possible choices

$$Z_{\nu,k} = \tilde{Z}(\gamma_k, \omega_k)^\nu = \tilde{Z}(\nu\gamma_k, \nu\omega_k)^\nu \tag{7}$$

Here, the power ν adjust both the damping and the frequency. This ensures that all oscillators are tuned to a multiple of a base fundamental frequency.

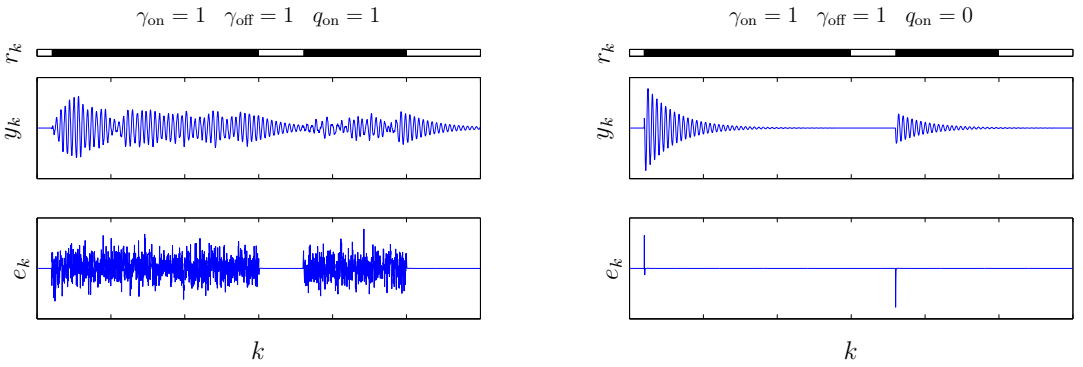
Both γ and ω can assume positive real values and the exact posterior has a complicated form due to the nonlinear relationship with observed data. One simplification, in contrast with models introduced earlier in section 3, is choosing the fundamental frequencies from a finite set taking values on a prespecified grid such as the tempered scale or finer gradatation according to the desired frequency resolution,

$$\omega_k = \omega(m_k) \qquad m_k \in 1 \ldots M$$



(a) Only periodic excitation

(b) Periodic + noise excitation



(c) Only noise excitation

(d) Only impulsive excitation

Figure 15: Time series y_k generated by a cascade of two phasors in Eq.6 (middle), conditioned on the indicator sequence $r_{0:K-1}$, (top – black = on, white = off). The excitation signals are shown at the bottom $e_k = (1 \;\; 0)\, s_k$. The other hyperparameters are fixed at $\gamma_{\text{ft}} = 0.005$, $\omega_{\text{ft}} = \pi/15$, $\omega_{\text{ex}} = \pi/16$, $q_{\text{onset}} = 1$

Here, m is a discrete index variable and $\omega(m)$ is a function to the associated fundamental frequency. For example, when m corresponds to the pitch label; $\omega(m)$ corresponds to the fundamental frequency, as in (5) for example. We define the following pair of discrete latent variables

$$d_k \quad = \quad (r_k, m_k)$$

and obtain a discrete chain that can visit one of the $|d| = 2M$ states at each time slice k. The prior can be taken Markovian as $p(d_k|d_{k-1})$. The most likely onsets, offsets and the fundamental frequency can be inferred via calculating the marginal maximum a-posteriori (MMAP) trajectory

$$d^*_{0:K-1} \quad = \quad \arg\max_{d_{0:K-1}} \int p(y_{0:K-1}|s_{0:K-1})p(s_{0:K-1}|d_{0:K-1})p(d_{0:K-1})ds_{0:K-1}$$

Alternatively, when online estimates are required, as is the case in real-time interaction, the filtering density can be computed recursively

$$\begin{aligned} p(s_k, d_k|y_{0:k}) \quad &\propto \quad \sum_{d_{k-1}} \int p(y_k|s_k)p(s_k, d_k|s_{k-1}, d_{k-1})p(s_{k-1}, d_{k-1}|y_{0:k-1})ds_{k-1} \\ p(d_k|y_{0:k}) \quad &= \quad \int p(s_k, d_k|y_{0:k})ds_k \end{aligned}$$

For general switching state space models, exact inference of the above quantities is not tractable. Whilst in principle the filtering distribution can be represented exactly as a Gaussian mixture and propagated in closed form, we have to still resort to approximations since the number of mixture components needed for exact representation of $p(s_k, d_k|y_{0:k})$ increases exponentially with increasing k. However, there is an interesting special case, when condtioned on a particular configuration of d, there is a "forgetting" property, i.e., if

$$p(s_k|d_{k-1:k} = \bar{d}, s_{k-1}) \quad = \quad p(s_k|d_{k-1:k} = \bar{d})$$

In this case, the exact MMAP trajectory or the filtering density (Fearnhead 2003; Cemgil, Kappen, and Barber 2004) can be computed in polynomial time. This can be shown by considering all trajectories $d^{(j)}_{0:k}$ for $j = 1 \dots |d|^{k+1}$ of the discrete states d upto time k. One can show that trajectories (j') which are dominated by (j) in terms of conditional marginal likelihood $z(j') \leq z(j) \equiv p(y_{0:k}, r^{(j)}_{0:k})$ can be discarded without destroying optimality. This greedy pruning strategy is optimal and leaves only a number of trajectories that is increasing polynomially with time (Cemgil, Kappen, and Barber 2004).

4.3 Polyphony and factorial models

The models described so far are useful for modelling complex sound sources. Yet, extensions are required for source separation or polyphonic transcription. This is typically done via *factorial models*, i.e., by constructing models over the product spaces of the individual sources, as was done for the static case in section 3, eq. (3).

One possible construction assumes that the audio is a superposition of several sources, indexed via $i = 1 \dots I$, with each source i modelled using the same model class, yet with different hyperparameter instantiations. Here, we first consider a factorial switching state space model for music transcription. Here, each latent process $\nu = 1 \dots W$ corresponds to a "piano key". Indicators $d_{0:W-1,0:K-1}$ encode a latent piano roll. Given this model, polyphonic transcription can be obtained as

$$d^*_{0:W-1,0:K-1} \quad = \quad \operatorname*{argmax}_d \int p(y|s)p(s|d)p(d)ds$$

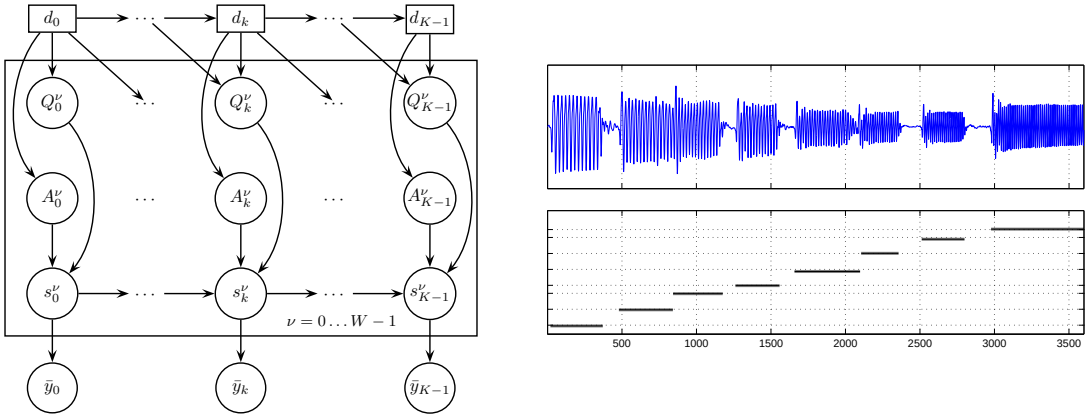


Figure 16: (Left) Switching state space model with a block diagonal state transition matrix where each block is denoted by ν. The dynamic harmonic model corresponds to the case when the blocks of the transition matrix are chosen as $A_k = \tilde{Z}(\gamma_k, \omega_k)$. (Right) Pitch detection and onset selection of a signal recorded from a bass guitar. The MMAP state trajectory is show as $r_k = \text{on} = \text{black}$ and the vertical axis denotes the pitch label index m_k.

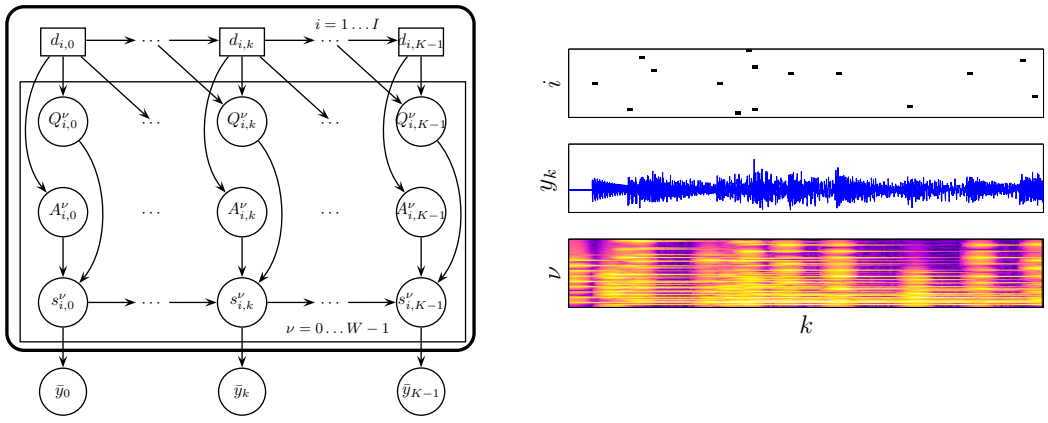


Figure 17: A Factorial Switching state space model and typical realisations. The task of polyphonic transcription is to find the maximum marginal a-poteriori estimate of the latent switches d^* given observations

A factorial switching state space model for polyphonic transcription is introduced in (Cemgil, Kappen, and Barber 2004) and some further strategies have been investigated in (Cemgil 2007). The model and a typical realisation is shown in Fig. 17. However, efficient inference in factorial switching state space models is still an open problem. In the presence of a large number of concurrent sources I, even a single time slice is intractable as the discrete variables in the product space have a cardinality that scales exponentially as $O(|d|^I)$. Even conditioned on d, the latent continuous state dimension $|s|$ can be very large for realistic models. and analytical integration over s via Kalman filtering techniques is computationally heavy. The main reason for this bottleneck in inference is the coupling between states s. When these are integrated out, all time slices of the discrete indicators become coupled and the exact inference problem is reduced to an intractable combinatorial optimisation problem. In the sequel, we will discuss models where the couplings are ignored.

5 Frequency domain models

The previous two sections described various time domain models for musical audio, including sinusoidal models and state-space models. The models are quite accurate for many examples of audio, although they show some non-robust properties in the case of signals which are far from steady-state oscillation and for instruments which do not closely obey the laws described above. Perhaps more critically, for large polyphonic mixes of many notes, each having potentially many partials, the computations can become very expensive, in particular the calculation of marginal likelihood terms in the presence of many Gaussian components $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. Computing the marginal likelihood is costly as this requires computation of Kalman filtering equations for a large state space (that scales with the number of tracked harmonics) and for very long time series (as typical audio signals are sampled at 44.1 KHz). Hence, either efficient approximations need to be developed or simplified models need to be constructed.

In this section we at least partially bypass the computational issues by working with approximate models in the frequency domain. These allow for direct likelihood calculations without resorting to expensive matrix inversions and determinant calculations. Later in the chapter these models will be elaborated further to give sophisitcated Bayesian non-negative matrix factorisation algorithms which are capable of learning the structure of the audio events in a semi-blind fashion. Here initially, though, we work with simple model-based structures in the frequency domain that are analogous to the time domain priors of the section 3. There are several routes to a frequency domain representation, including multi-resolution transforms, wavelets, etc., though here we use a simple windowed discrete Fourier transform as examplar. We now propose two versions of a frequency domain likelihood model, both of which bypass the main computational burden of the high-dimensional time-domain Gaussian models.

5.0.1 Gaussian frequency-domain model

The first model proposed is once again a Gaussian model. In the frequency domain we will have typically complex-valued expansion coefficients of the data on a one-dimensional lattice of frequency values $\nu \in N$, i.e. a set of spectrum values y_ν. The assumption is that the contribution of each musical source term to the expansion coefficients is as independent zero-mean (complex) Gaussians, with variance determined by the parameters of the musical note:

$$s_{i,\nu} \sim \mathcal{N}_C(0, \lambda_\nu(\theta_i))$$

where $\theta_i = \{n_i, \omega_{0,i}, M_i, g_i\}$ has the same interpretation as for the earlier time-domain model, but now we can neglect the α and β coefficients since the random behaviour is now directly

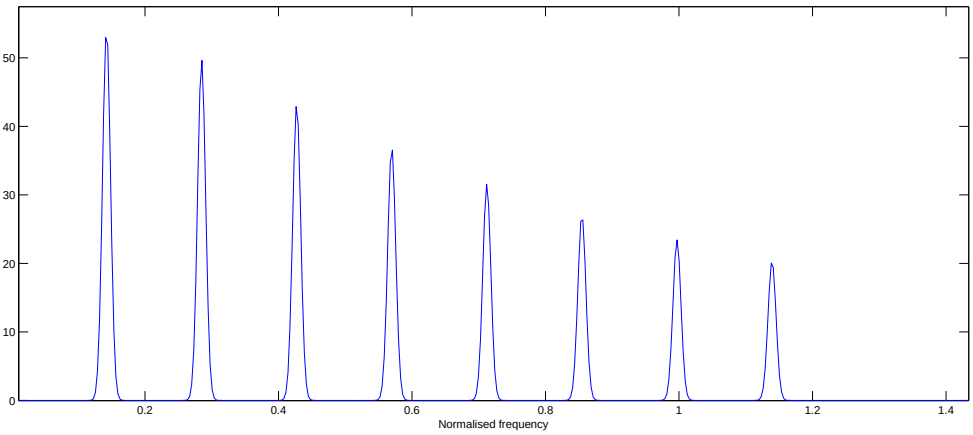


Figure 18: Template function $\lambda_\nu(\theta_i)$ with $M_i = 8$, $\omega_{0,i} = 0.71$, Gaussian pulse shape.

modelled by $s_{i,\nu}$. This is a very natural formulation for generation of polyphonic models since we can add a number of sources together to make a single complex Gaussian data model:

$$y_\nu \sim \mathcal{N}_C(0, S_{v,\nu} + \sum_{i=1}^{I} \lambda_\nu(\theta_i))$$

Here, $S_{v,\nu} > 0$ models a Gaussian background noise component in a manner analogous to the time-domain formulation's v_t and it then remains to design the positive-valued 'template' functions λ. Once again, Fig. 1.1 gives some guidance as to the general characteristics required. We then model the template using a sum of positive valued pulse waveforms ϕ_ν, shifted to be centred at the expected partial position, and whose amplitude decays with increasing partial number:

$$\lambda_\nu(\theta_i) = \sum_{m=1}^{M_i} g_i^2 k_m \phi_{\nu - m\omega_{0,i}} \tag{8}$$

where k_m, g_i and M_i have exactly the same interpretation as in the time-domain model. An example template construction is shown in Fig. 18, in which a Gaussian pulse shape has been utilised.

5.0.2 Point process frequency-domain model

The Gaussian frequency domain model requires a knowledge of the conditional distribution for the whole range of spectrum values. However, the salient features in terms of pitch estimation appear to be the *peaks* of the spectrum see Fig. 1.1. Hence a more parsimonious likelihood model might work only with the peaks detected from the Fourier magnitude spectrum. Thus we propose as an alternative to the Gaussian spectral model, a point process model for the peaks in the spectrum. Specifically, if the peaks in the spectrum of an individual note are assumed to be drawn from a one-dimensional inhomogeneous Poisson point process having intensity function $\lambda_\nu(\theta_i)$ (considered as a function of continuous frequency ν), then the combined set of peaks from many notes may be combined, under an independence assumption, to give a Poisson point process whose intensity function is the sum of the individual intensities (Grimmett and Stirzaker 2001). Suppose we detect a set of peaks in the magnitude spectrum $\{p_j\}_{j=1}^{J}$, $\nu_{\min} < p_j < \nu_{\max}$.

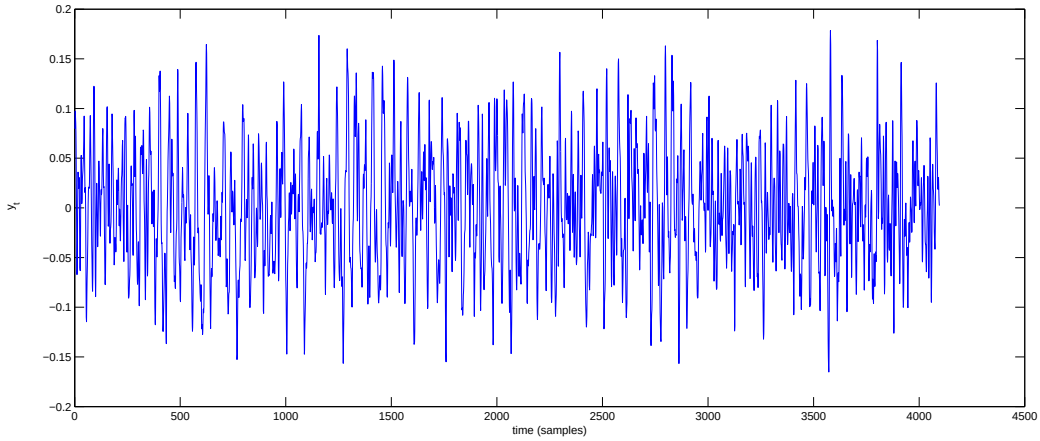


Figure 19: Audio waveform - single chord data.

Then the likelihood may be readily computed using:

$$p(\{p_j\}_{j=1}^{J}, J|\Theta) = \mathrm{Po}(J|Z(\Theta)) \prod_{j=1}^{J} \frac{\left(S_{v,p_j} + \sum_{i=1}^{I} \lambda_{p_j}(\theta_i)\right)}{Z(\Theta)}$$

where $Z(\Theta) = \int_{\nu_{\min}}^{\nu_{\max}} \left(S_{v,\nu} + \sum_{i=1}^{I} \lambda_{\nu}(\theta_i)\right) d\nu$ is the normalising constant for the overall intensity function. Here once again we include a background intensity function $S_{v,\nu}$ which models 'false detections', i.e. detected peaks that belong to no existing musical note. The form of the template functions λ can be very similar to that in the Gaussian frequency model, Eq. 8. A modified form of this likelihood function was successfully applied for chord detection problems in (Peeling, Li, and Godsill 2007).

5.1 Example: inference in the frequency domain models

The frequency domain models provide a substantially faster likelihood calculation than the earlier time-domain models, allowing for rapid inference in the presence of significantly larger chords and tone complexes. Here we present example results for a tone complex containing many different notes, played on a pipe organ. Analysis is performed on a very short segment of 4096 data points, sampled at a rate of $\omega_s = 2\pi \times 44,100$ rad.s^{-1} - hence just under 0.1s of data, see Fig. 19. From the score of the music we know that there are four notes simultaneously playing: C5, F♯5, B5, and D6, or MIDI note numbers 72, 78, 83 and 86. However, the mix is complicated by the addition of pipes one octave below and one or more octaves above the principal pitch, and hence we have at least 12 notes present in the complex, MIDI notes 60, 66, 71, 72, 74, 78, 83, 84, 86, 90, 95, and 98. Since the upper octaves share all of their partials with notes from one or more octaves below, it is not clear whether the models will be able to distinguish all of the sounds as separate notes. We run the frequency-domain models using the prior framework of Section 3.1 and a reversible jump MCMC scheme of the same form as that used in the previous transient analysis example. Firstly, using the Gaussian frequency domain model of section 5.0.1, the MCMC burn-in for the note number vector $n = [n_1, n_2, ..., n_I]$ is shown in Fig. 20. This is a variable dimension vector under the reversible jump MCMC and we can see notes entering or leaving the vector as iterations proceed. We can also see large moves of an octave (± 12 notes) or a fifth (+7 or -5 notes), corresponding to specialised Metropolis-

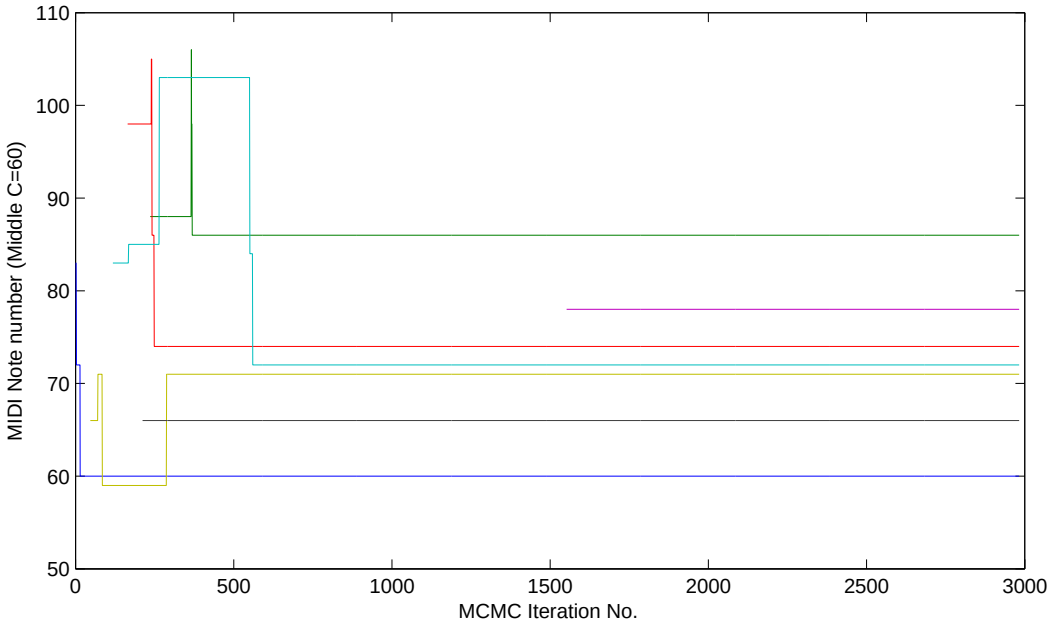


Figure 20: Evolution of the note number vector with iteration number - single chord data. Gaussian frequency domain model.

Hastings moves which center their proposals on the octave or fifth as well as the locality of the current note. As is typical of these models, the MCMC becomes slow moving once converged to a good mode of the distribution and further large moves only occur occasionally. There is a good case here for using adaptive or population MCMC schemes to improve the properties of the MCMC. Nevertheless, convergence is much faster than for the earlier proposed time domain models, particularly in terms of the model order sampling, which was here initialised at $I = 1$, i.e. one single note present at the start of the chain. Specialised independence proposals have also been devised, based on simple pitch estimation methods applied to the raw data. These are largely responsible for the initiation of new notes in the MCMC chain. In this instance the MCMC has identified correctly 7 out of the (at least) 12 possible pitches present in the music: 60, 66, 71, 72, 74, 78, 86. The remaining 5 unidentified pitches share all of their partials with lower pitches estimated by the algorithm, and hence it is reasonable that they remain unestimated. Examination of the discrete Fourier magnitude spectrum (Fig. 21) shows that the higher pitches (with the possible exception of $n_7 = 83$, whose harmonics are modelled by $n_3 = 71$) are generally buried at very low amplitude in the spectrum and can easily be absorbed into the model for pitches one or more octaves lower in pitch.

We can compare these results with those obtained using the Poisson model of section 5.0.2. The MCMC was run under identical conditions to the Gaussian model and we plot the equivalent note index output in Fig. 22. Here we see that fewer notes are estimated, since the basic point process model takes no account of the amplitudes of the peaks in the spectrum, and hence is happy to assign all harmonics to the lowest possible fundamental pitch. The four predominant pitches estimated are the four lowest fundamentals: 60, 66, 71 and 74. The sampler is, however, generally more mobile and we see a better and more rapid exploration of the posterior.

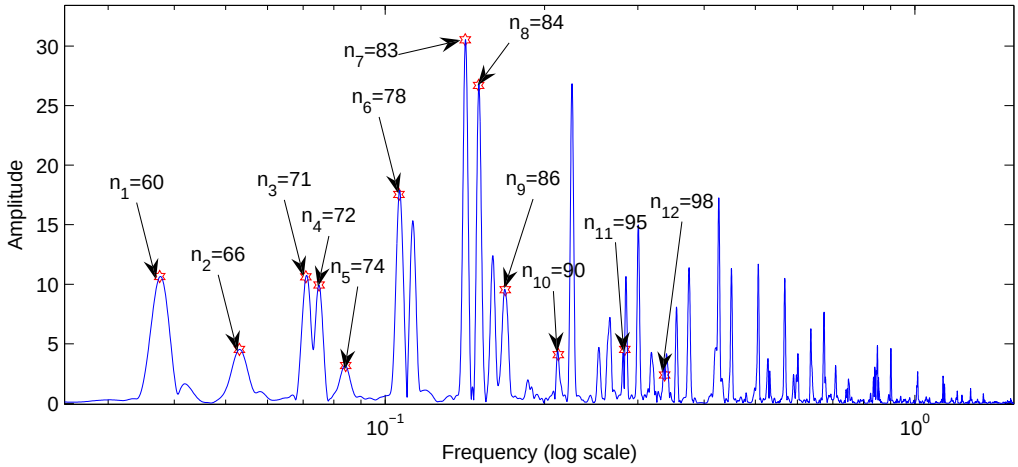


Figure 21: Discrete Fourier magnitude spectrum for 12-note chord. True note positions marked with red pentagram.

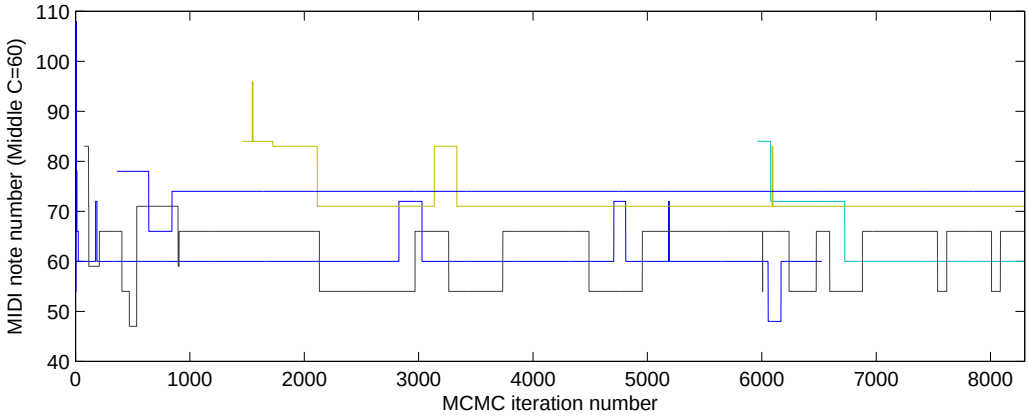


Figure 22: Evolution of the note number vector with iteration number - single chord data. Poisson frequency domain model.

5.2 Further prior structures for Transform domain representations

In audio processing, the energy content of a signal is typically time-varying hence it is natural to model audio with a process with a time varying power spectral density on a time frequency plane (Reyes-Gomez, Jojic, and Ellis 2005; Wolfe, Godsill, and Ng 2004; Févotte, Daudet, Godsill, and Torrésani 2006), and several prior structures are proposed in the literature for modelling the expansion coefficients. The central idea is choosing a latent variance model varying over time and frequency bins

$$\begin{aligned} s_{\nu,k}|Q_{\nu,k} &\sim \mathcal{N}(s_{\nu,k}; 0, Q_{\nu,k}) \\ Q_{\nu,k} &= q_{\nu,k} I \end{aligned}$$

In (Wolfe, Godsill, and Ng 2004), the following structure is proposed under the name *Gabor Regression*

$$q_{\nu,k}|r_{\nu,k} \sim [r_{\nu,k} = \text{on}]\, \mathcal{IG}(q_{\nu,k}; a, b/a) + [r_{\nu,k} = \text{off}]\, \delta(q_{\nu,k})$$

Moreover, the joint distribution over the latent indicators $r = r_{0:W-1,0:K-1}$ is taken as a pairwise Markov Random field where u denotes a double index $u = (\nu, k)$

$$p(r) \propto \prod_{(u,u')\in\mathcal{E}} \phi(r_u, r_{u'})$$

5.3 Gamma chains and fields

An alternative model is introduced in (Cemgil and Dikmen 2007; Cemgil, Peeling, Dikmen, and Godsill 2007), where a Markov Random field is directly placed on the variance terms as

$$p(q) = \int d\lambda p(q, \lambda)$$

using a so-called gamma field.

To understand the construction of a Gamma field, it is instructive to look first at a chain, where we have an alternating sequence of Gamma and inverse Gamma random variables

$$q_u|\lambda_u \sim \mathcal{IG}(q_u; a_q, a_q\lambda) \qquad \lambda_{u+1}|q_u \sim \mathcal{G}(\lambda_{u+1}; a_\lambda, q_u/a_\lambda)$$

Note that this construction leads to conditionally conjugate Markov blankets that are given as

$$\begin{aligned} p(q_u|\lambda_u, \lambda_{u+1}) &\propto \mathcal{IG}(q_u; a_q + a_\lambda, a_q\lambda_u + a_\lambda\lambda_{u+1}) \\ p(\lambda_u|q_{u-1}, q_u) &\propto \mathcal{G}(\lambda_u; a_\lambda + a_q, a_\lambda q_{u-1}^{-1} + a_q q_u^{-1}) \end{aligned}$$

Moreover it can be shown that any pair of variables q_i and q_j are positively correlated, and q_i and λ_k are negatively correlated. Note that this is a particular *stochastic volatility* model useful for characterisation of non-stationary behaviour observed in time series (Shepard 2005).

We can represent a chain by a graphical model where the edge set is $\mathcal{E} = \{(u, u)\} \cup \{(u, u+1)\}$. Considering the Markov structure of the chain, we define a gamma field $p(q, \lambda)$ as a bipartite undirected graphical model consisting of the vertex set $\mathcal{V} = \mathcal{V}_\lambda \cup \mathcal{V}_q$, where partitions $\mathcal{V}_\lambda$ and $\mathcal{V}_q$ denotes the collection of variables λ and q that are conditionally distributed $\mathcal{G}$ and $\mathcal{IG}$ respectively. We define an edge set $\mathcal{E}$ where an edge $(u, u') \in \mathcal{E}$ such that $\lambda_u \in \mathcal{V}_\lambda$ and $q_{u'} \in \mathcal{V}_q$, if the joint distribution admits the following factorisation

$$p(\lambda, q) \propto \left(\prod_{u\in\mathcal{V}_\lambda} \lambda_u^{(\sum_{u'} a_{u,u'} - 1)} \right) \left(\prod_{u'\in\mathcal{V}_q} q_u^{-(\sum_u a_{u,u'} + 1)} \right) \left(\prod_{(u,u')\in\mathcal{E}} \exp(-a_{u,u'} \frac{\lambda_u}{q_{u'}}) \right)$$

Here, the shape parameters play the role of coupling strengths; when $a_{u,u'}$ is large, adjacent nodes are correlated. Given, this construction, various signal models can be developed figure 23.

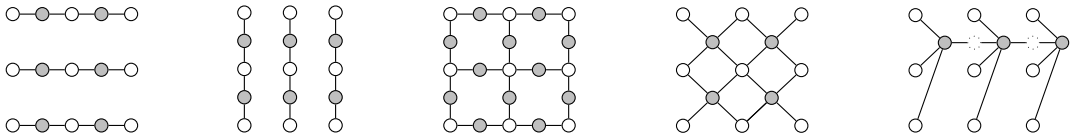


Figure 23: Possible model topologies for Gamma fields. White and gray nodes corresponds to $\mathcal{V}_q$ and $\mathcal{V}_\lambda$ nodes respectively. The horizontal and vertical axis corresponds to frequency ν and frame index k. Each model describes how the prior variances are coupled as a function of time-frequency index. For example, the first model from the left corresponds to a source model with "spectral continuity", energy content of a given frequency band changes only slowly. The second model is useful for modelling impulsive sources where energy is concentrated in time but spread across frequencies.

5.4 Models based on Latent Variance/Intensity factorisation

The various Markov random field priors of the previous introduced couplings between the latent variances $q_{\nu,k}$. Yet, another alternative is to decompose the latent variances as a product. Here, we define the following hierarchical model (see Fig. 25)

$$\begin{aligned} s_{\nu,k} &\sim \mathcal{N}(s_{\nu,k}; 0, q_{\nu,k}) & q_{\nu,k} &= t_\nu v_k \\ t_\nu &\sim \mathcal{IG}(t_\nu; a^t_\nu, a^t_\nu b^t_\nu)) & v_k &\sim \mathcal{IG}(v_k; a^v_k, a^v_k b^v_k)) \end{aligned} \tag{9}$$

Such models are also particularly useful for modelling acoustic instruments. Here, the t_ν variables can be interpreted as average expected energy template as a function of frequency bin. At each time, this template is modulated by v_ν, to adjust the overall volume. An example is given in Figure 24 to represent a piano sound. Here, the template gives the harmonic structure of the pitch and the excitation characterises the time varying energy.

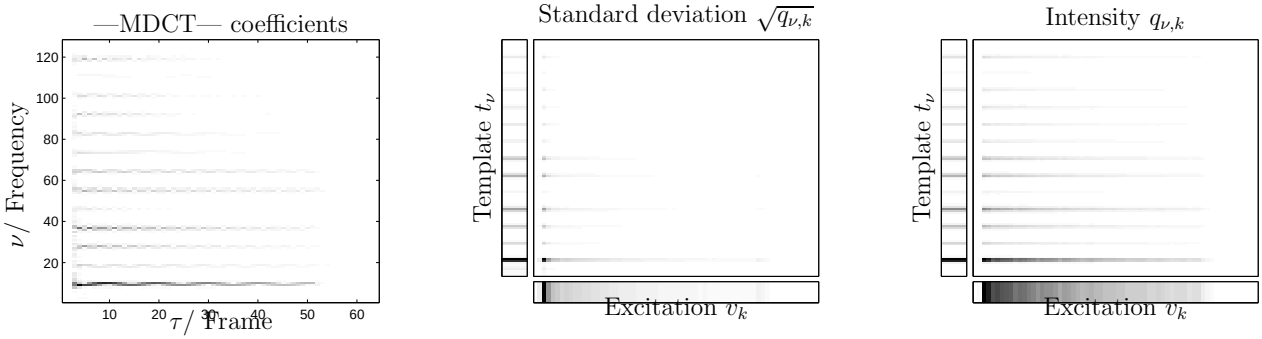


Figure 24: (Left) The spectrogram of a piano $|s_{\nu,k}|^2$. (Middle) Estimated templates and excitations using the conditionally Gaussian model defined in 9, where $q_{\nu,k}$ is the latent variance (Right) Estimated templates and excitations using the conditionally Poisson model defined in the next section 15

A simple factorial model, that uses the gamma chain prior models introduced in section 5.3 is constructed as follows:

$$x_{\nu,k} = \sum_i s_{\nu,i,k} \qquad s_{\nu,i,k} \sim \mathcal{N}(s_{\nu,i,k}; 0, q_{\nu,i,k}) \qquad Q = \{q_{\nu,i,k}\} \sim p(Q|\Theta^t) \tag{10}$$

The computational advantage of this class of models is the conditional independence of the latent sources given the latent variance variables. Given the latent variances and data, the

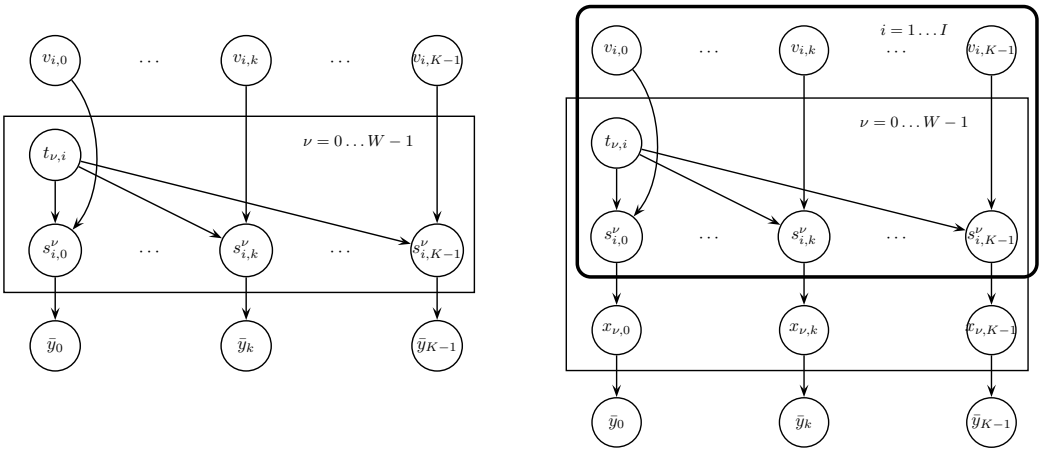


Figure 25: (Left) Latent variance/intensity models in product form (Eq.9). Hyperparameters are not shown. (Right) Factorial version of the same model, used for polyphonic estimation as used in section 5.5.3.

posterior of the sources is a product of Gaussian distributions. In particular, the individual marginals are given in closed form as

$$\begin{aligned} p(s_{\nu,i,k}|X,Q) &= \mathcal{N}(s_{\nu,i,k}; \kappa_{\nu,i,k} x_{\nu,k}, q_{\nu,i,k}(1-\kappa_{\nu,i,k})) \\ \kappa_{\nu,i,k} &= q_{\nu,i,k} / \sum_{i'} q_{\nu,i',k} \end{aligned}$$

This means that if the latent variances can be estimated, source separation can be easily accomplished. The choice of prior structures on the latent variances $p(Q|\cdot)$ is key here.

Below we illustrate this approach in single channel source separation for transient/harmonic decomposition. Here, we assume that there are two sources $i = 1, 2$. The prior variances of the first source $i = 1$ are tied across time frames using a gamma chain and aims to model a source with harmonic continuity. The prior has the form $\prod_\nu p(q_{\nu,i=1,1:K})$. This model simply assumes that for a given source the amount of energy in a frequency band stays roughly constant. The second source $i = 2$ is tied across frequency bands and has the form $\prod_k p(q_{1:W,i=2,k})$; this model tries to capture impulsive/percusive structure (for example compare the piano and conga examples in Fig.4). The model aims to separate the sources based on harmonic continuity and impulsive structure.

We illustrate this approach to separate a piano sound into its constituent components and drum separation. We assume that $J = 2$ components are generated independently by two Gamma chain models with vertical and horizontal topology. In figure 26-(b), we observe that the model is able to separate transients and harmonic components. The sound files of these results can be downloaded and listened at the following url: `http://www-sigproc.eng.cam.ac.uk/~sjg/haba`, which is perhaps the best way assess the sound quality.

The variance/intensity factorisation models described in Eq. 9 have also straightforward factorial extensions

$$\begin{aligned} x_{\nu,k} &= \sum_i s_{\nu,i,k} \\ s_{\nu,i,k} &\sim \mathcal{N}(s_{\nu,i,k}; 0, q_{\nu,i,k}) \qquad q_{\nu,i,k} = t_{\nu,i} v_{i,k} \qquad (11) \\ T &= \{t_{\nu,i}\} \sim p(T|\Theta^t) \qquad V = \{v_{i,k}\} \sim p(V|\Theta^v) \qquad (12) \end{aligned}$$

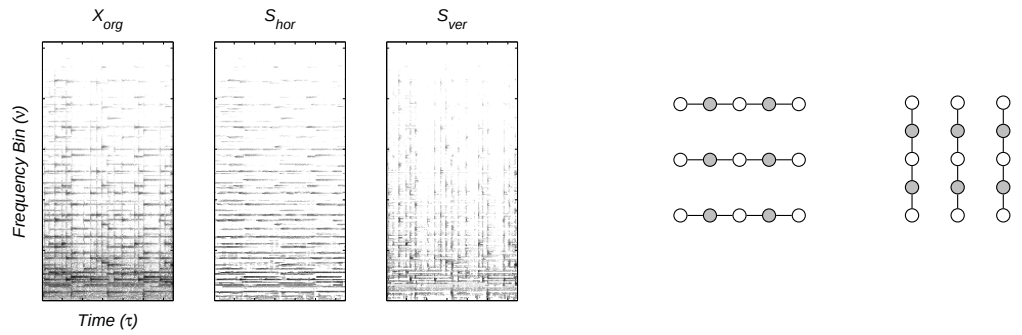


Figure 26: Single channel Source Separation example, left to right, log-MDCT coefficients of the original signal and reconstruction with horizontal and vertical IGMRF models.

If we integrate out the latent sources, the marginal is given as

$$x_{\nu,k} \sim \mathcal{N}(x_{\nu,k}; 0, \sum_i t_{\nu,i} v_{i,k})$$

Note that, as $\sum_i t_{\nu,i} v_{i,k} = [TV]_{\nu,k}$, the variance "field" Q is given compactly as the matrix product $Q = TV$. This resembles closely a matrix factorisation and is used extensively in audio modelling. In the next section, we discuss models of this type.

5.5 Non-negative Matrix factorisation models

Until now, we have described conditionally Gaussian models. Recently, a popular branch of source separation and analysis of musical audio literature has focused on non-negativity of the magnitude spectrogram $X = \{x_{\nu,\tau}\}$ with $x_{\nu,\tau} \equiv \|s_{\nu,k}\|_2^{1/2}$, where $s_{\nu,k}$ are expansion coefficients obtained from a time frequency expansion. The basic idea is representing a spectrogram by enforcing a factorisation as $X \approx TV$ where both T and V are matrices with positive entries (Smaragdis and Brown 2003; Abdallah and Plumbley 2006; Virtanen 2006; Kameoka 2007; Bertin, Badeau, and Richard 2007; Vincent, Bertin, and Badeau 2008). In music signal analysis, T can be interpreted as a codebook of templates, corresponding to spectral shapes of individual notes and V is the matrix of activations, somewhat analogous to a musical score. Often, the following objective is minimised:

$$(T, V)^* = \min_{T,V} D(X||TV) \tag{13}$$

where D is the information (Kullback-Leibler) divergence, given by

$$D(X||\Lambda) = \sum_{\nu,\tau} \left(x_{\nu,\tau} \log \frac{x_{\nu,\tau}}{\lambda_{\nu,\tau}} - x_{\nu,\tau} + \lambda_{\nu,\tau} \right) \tag{14}$$

Using Jensen's inequality (Cover and Thomas 1991) and concavity of $\log x$, it can be shown , that $D(\cdot)$ is nonnegative and $D(X||\Lambda) = 0$ if and only if $X = \Lambda$. The objective in (13) could be minimised by any suitable optimisation algorithm. (Lee and Seung 2000) have proposed an efficient variational bound minimisation algorithm that has attractive convergence properties, that has been since successfully applied to various applications in signal analysis and source separation. Although not widely acknowledged, it can be shown that the minimisation algorithm is in fact an EM algorithm with data augmentation (Cemgil 2008). More precisely, it can be shown that minimising D w.r.t., T and V is equivalent finding the ML solution of the following

hierarchical model

$$
\begin{aligned}
x_{\nu,k} &= \sum_i s_{\nu,i,k} \\
s_{\nu,i,k} &\sim \mathcal{PO}(s_{\nu,i,k}; 0, \lambda_{\nu,i,k}) \qquad \lambda_{\nu,i,k} = t_{\nu,i} v_{i,k} \qquad (15) \\
t_{\nu,i} &\sim \mathcal{G}(t_{\nu,i}; a^t_{\nu,i}, b^t_{\nu,i}/a^t_{\nu,i}) \qquad v_{i,k} \sim \mathcal{G}(v_{i,k}; a^v_{i,k}, b^v_{i,k}/a^v_{i,k}) \qquad (16)
\end{aligned}
$$

The computational advantage of this models is the conditional independence of the latent sources given the variance variables. In particular, we have

$$
\begin{aligned}
p(s_{\nu,i,k}|X,T,V) &= \mathcal{BI}(s_{\nu,i,k}; x_{\nu,k}, \kappa_{\nu,i,k}) \\
\kappa_{\nu,i,k} &= \lambda_{\nu,i,k} / \sum_{i'} \lambda_{\nu,i',k}
\end{aligned}
$$

This means that if the latent variances can be estimated somehow, source separation can be easily accomplished as $\mathsf{E}(s)_{\mathcal{BI}(s;x,\kappa)} = \kappa x$.

5.5.1 Variational Bayes

It is also possible to estimate the marginal likelihood $p(X)$ by integrating out all the templates and excitations. This can be done via Gibbs sampling or using a variational approach. The variational approach is very similar to the EM algorithm, with an additional approximation step. We sketch here the Variational Bayes (VB) (Ghahramani and Beal 2000; Bishop 2006) method to bound the marginal loglikelihood as

$$
\begin{aligned}
\mathcal{L}_X(\Theta) &\equiv \log p(X|\Theta) \geq \sum_S \int d(T,V) q \log \frac{p(X,S,T,V|\Theta)}{q} \qquad (17) \\
&= \mathsf{E}(\log p(X,S,V,T|\Theta))_q + H[q] \equiv \mathcal{B}_{VB}[q] \qquad (18)
\end{aligned}
$$

where, $q = q(S,T,V)$ is an instrumental distribution and $H[q]$ is its entropy. The bound is tight for the exact posterior $q(S,T,V) = p(S,T,V|X,\Theta)$, but as this distribution is complex we assume a factorised form for the instrumental distribution by ignoring some of the couplings present in the exact posterior

$$
q(S,T,V) = q(S)q(T)q(V) = \left(\prod_{\nu,\tau} q(s_{\nu,1:I,\tau})\right)\left(\prod_{\nu,i} q(t_{\nu,i})\right)\left(\prod_{i,\tau} q(v_{i,\tau})\right) \equiv \prod_{\boldsymbol{\alpha}\in\mathcal{C}} q_{\boldsymbol{\alpha}}
$$

where $\boldsymbol{\alpha} \in \mathcal{C} = \{\{S\},\{T\},\{V\}\}$ denotes set of disjoint clusters. Hence, we are no longer guaranteed to attain the exact marginal likelihood $\mathcal{L}_X(\Theta)$. Yet, the bound property is preserved and the strategy of VB is to optimise the bound. Although the best q distribution respecting the factorisation is not available in closed form, it turns out that a local optimum can be attained by the following fixed point iteration:

$$
q^{(n+1)}_{\boldsymbol{\alpha}} \propto \exp\left(\mathsf{E}(\log p(X,S,T,V|\Theta))_{q^{(n)}_{\neg\boldsymbol{\alpha}}}\right) \qquad (19)
$$

where $q_{\neg\boldsymbol{\alpha}} = q/q_{\boldsymbol{\alpha}}$. This iteration monotonically improves the individual factors of the q distribution, i.e. $\mathcal{B}[q^{(n)}] \leq \mathcal{B}[q^{(n+1)}]$ for $n = 1,2,\ldots$ given an initialisation $q^{(0)}$. The order is not important for convergence, one could visit blocks in arbitrary order. However, in general, the attained fixed point depends upon the order of the updates as well as the starting point $q^{(0)}(\cdot)$. This approach is computationally rather attractive and is very easy to implement (Cemgil 2008).

5.5.2 Variational update equations and sufficient statistics

The expectations of $\mathsf{E}(\log p(X, S, T, V|\Theta))$ are functions of the sufficient statistics of q. The fixed point iteration for the latent sources S (where $m_{\nu,\tau} = 1$), and excitations V leads to the following

$$q(s_{\nu,1:I,\tau}) = \mathcal{M}(s_{\nu,1:I,\tau}; x_{\nu,\tau}, p_{\nu,1:I,\tau}) \qquad q(v_{i,\tau}) = \mathcal{G}\left(v_{i,\tau}; \alpha^v_{i,\tau}, \beta^v_{i,\tau}\right) \tag{20}$$

$$p_{\nu,i,\tau} = \exp(\mathsf{E}(\log t_{\nu,i}) + \mathsf{E}(\log v_{i,\tau}))/\sum_i \exp(\mathsf{E}(\log t_{\nu,i}) + \mathsf{E}(\log v_{i,\tau})) \tag{21}$$

$$\alpha^v_{i,\tau} = a^v_{i,\tau} + \sum_\nu m_{\nu,\tau}\mathsf{E}(s_{\nu,i,\tau}) \qquad \beta^v_{i,\tau} = \left(\frac{a^v_{i,\tau}}{b^v_{i,\tau}} + \sum_\nu m_{\nu,\tau}\mathsf{E}(t_{\nu,i})\right)^{-1} \tag{22}$$

The variational parameters of $q(t_{\nu,i}) = \mathcal{G}\left(t_{\nu,i}; \alpha^t_{\nu,i}, \beta^t_{\nu,i}\right)$ are found similarly. The hyperparameters can be optimised by maximising the variational Bound. While this does not guarantee to increase the true marginal likelihood, it leads in this application to quite practical and fast algorithms.

5.5.3 Example: Polyphonic pitch estimation

In this section, we illustrate Bayesian NMF for polyphonic pitch detection. The approach consists of two stages:

1. Estimation of hyperparameters given a corpus of piano notes
2. Estimation of templates and excitations given new polyphonic data and fixed hyperparameters

In the first stage, we estimate the hyperparameters $a^t_{\nu,i} = a^t_i$ and $b^t_{\nu,i}$ (see Eq. 16), via maximisation of the variational bound given in Eq. 18. Here, the observations are matrices X_i matrix is a spectrogram computed given each note $i = 1 \ldots I$. In figure 27, we show the estimated scale parameters $b^t_{\nu,i}$ as a function of frequency band ν and note index i. The harmonic structure of each note is clearly visible.

To test the approach, we synthesize a music piece (here, a short segment from the beginning of "Für Elise" by Beethoven), given a midi piano roll and recordings of isolated notes from a piano by simply appropriately shifting each time series and adding. The piano roll and the the spectrogram of the synthesized audio are shown in Figure 28. The pitch detection task is infering the excitations given the hyperparameters and the spectrogram.

The results are shown in Figure 29. The top figure shows the excitations estimated give the prior Eq. 16. The notes are visible here but there are some artifacts. The middle figure shows results from a model where excitations are tied across time using a Gamma chain introduced in section 5.3. This prior is highly effective here and we are able to get a more clearer picture. The bottom figure displays results obtained from a real recording of "Für Elise", performed on electric guitar. Interestingly, whilst we are still using the hyperparameters estimated from a piano, the infered excitations show significant overlap with the original score.

6 Probabilistic Models for Tempo, Rhythm, Meter

An important feature of music information retrieval and interactive performance systems is the capacity to infer attributes related to the temporal structure of audio signals. Detecting the *pulse* or foot-tapping rate of musical audio signals has been a focus of research in musical signal processing for several years (Cemgil 2004; Klapuri and Davy 2006). However, little attention has been paid to the task of extracting information about the more abstract musical concepts

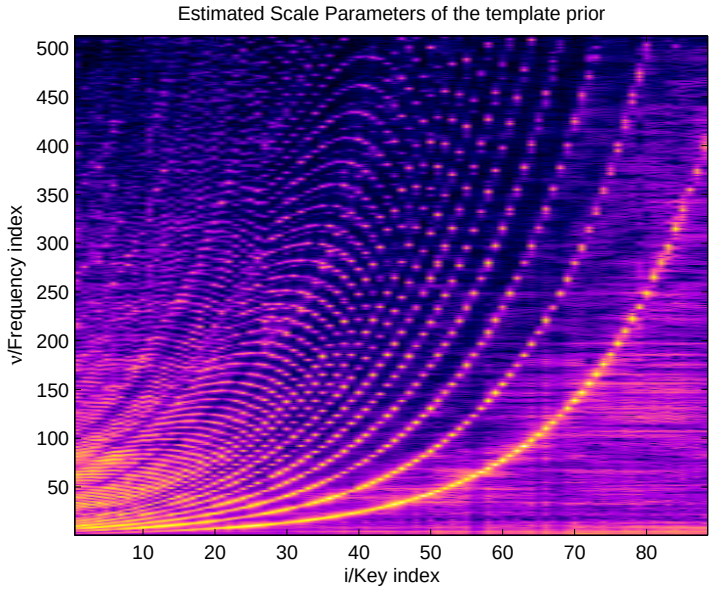


Figure 27: Estimated template hyperparameters $b^t_{\nu,i}$.

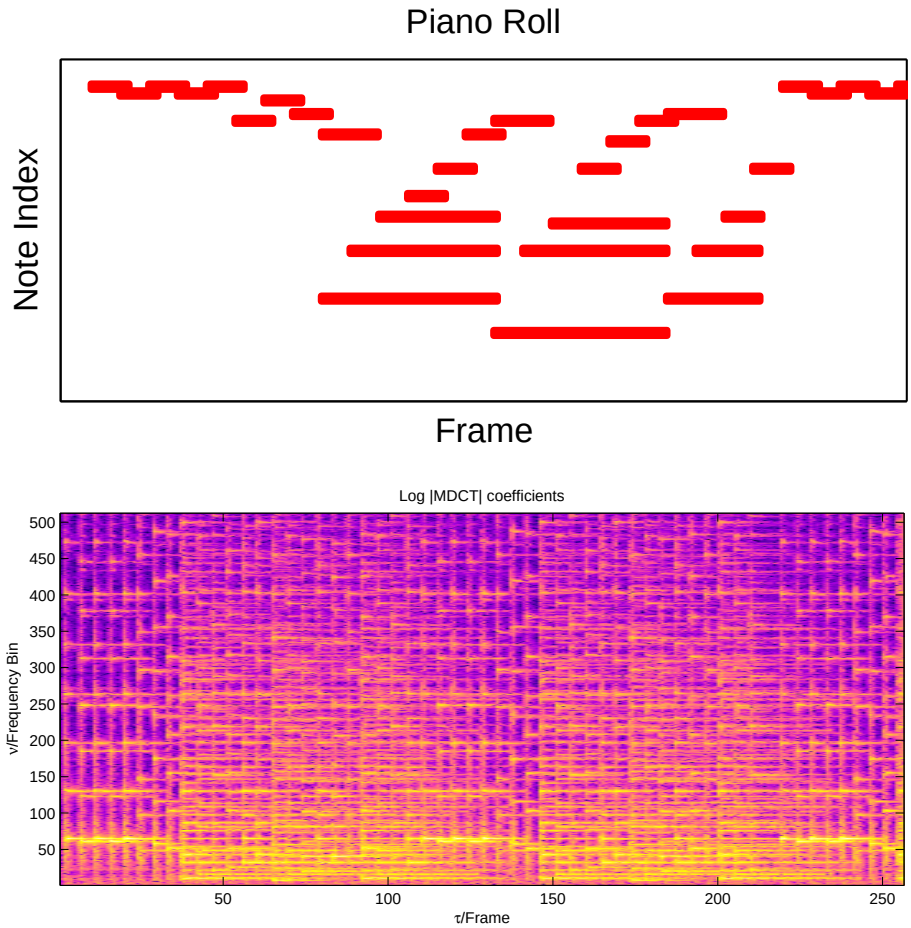


Figure 28: The ground truth piano roll and the spectrum of the polyphonic data

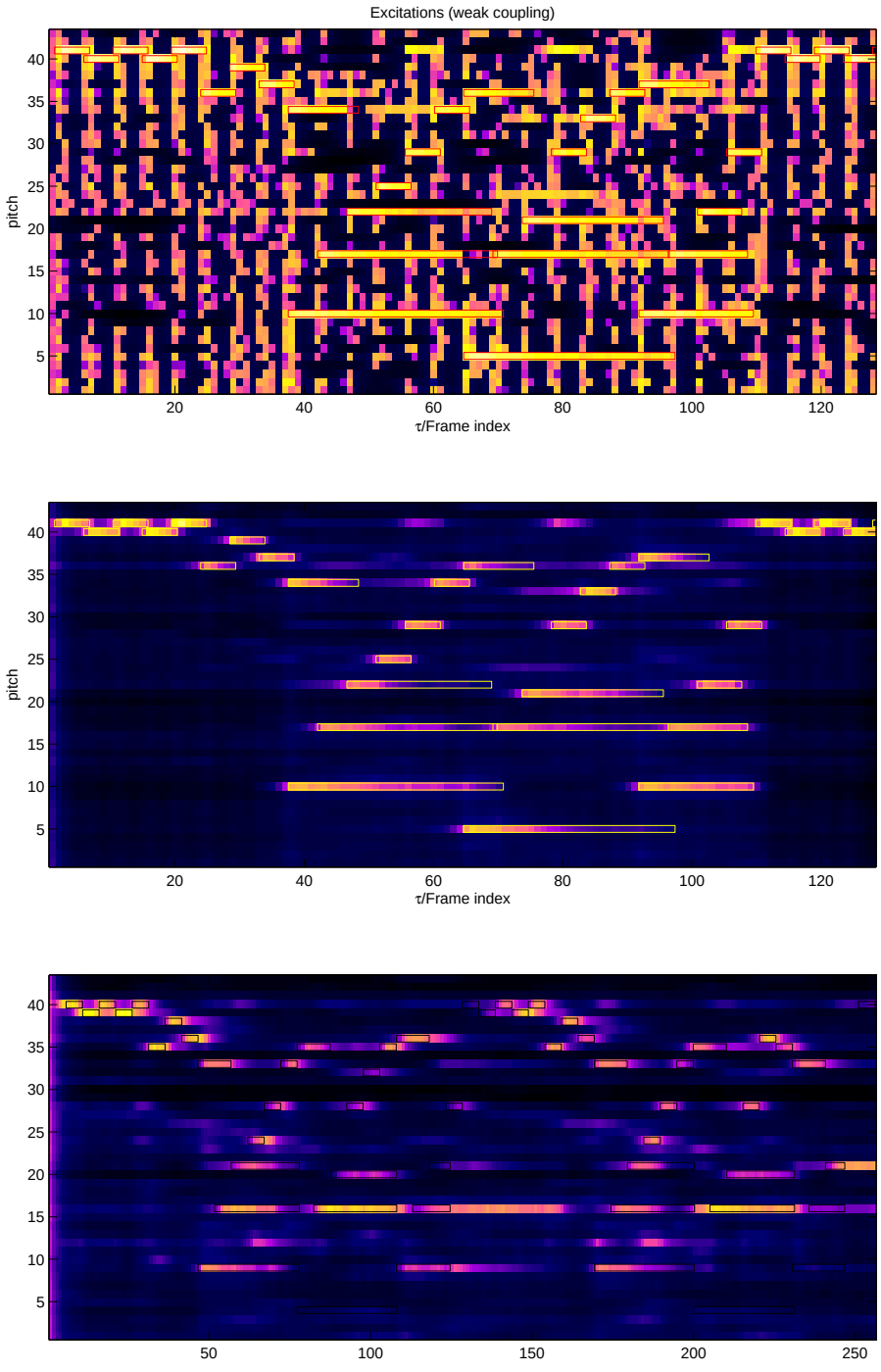


Figure 29: Polyphonic Pitch detection. Estimated expected excitations (Top) Uncoupled excitations (Middle) Tied excitations using a Gamma chain, ground truth shown in white (Bottom) Excitations estimated from a guitar using the hyperparameters estimated from a piano - ground truth shown in black.

of *tempo*, *rhythm* and *meter*. An hierarchical Bayesian approach is ideally suited to this task. In order to quantify the musical concepts of interest, we interpret tempo as being equivalent to the performed rate of quarter notes and rhythmic pattern as indicating the regions within a musical bar in which note onsets are likely to occur. In this section we summarise a modelling framework which permits inference of these quantities from observed MIDI onset events or raw audio samples (Whiteley, Cemgil, and Godsill 2006; Whiteley, Cemgil, and Godsill 2007).

6.1 A Bar-pointer Model

Central to the method is a dynamical model of a *bar-pointer*, an hypothetical, hidden, dynamical object which maps the positive real line to one period of a latent rhythmical pattern, i.e. one musical bar. Different rhythms are represented by rhythmic pattern functions which control the hyper-parameters of a conjugate gamma-Poisson observation model. The trajectory of the bar-pointer modulates these functions onto the time line. Conditional upon this trajectory and hyper-parameter values, observed MIDI onset events are modelled as being a non-homogeneous Poisson Process or raw audio samples are modelled as being a Gaussian process. The inference task is then to estimate the state of the bar-pointer and values of rhythmic pattern indicator variables. Both filtering and smoothing are of interest in different applications.

For a discrete time index k and Δ a positive constant, at time $t_k = k\Delta$ denote by ϕ_k the position of the bar-pointer, which takes values in $[0, 1)$. Denote by $\dot{\phi}_k$ its velocity which takes values in $[\dot{\phi}_{min}, \dot{\phi}_{max}]$, where $\dot{\phi}_{min} > 0$ and $\dot{\phi}_{max}$ represents the mximum tempo to be considered. The motion of the bar-pointer is modelled as being a piece-wise constant velocity process:

$$\phi_{k+1} = (\phi_k + \Delta\dot{\phi}_k) \text{mod } 1,$$
$$p(\dot{\phi}_{k+1}|\dot{\phi}_k) \propto \mathcal{N}(\dot{\phi}_k, \sigma^2_\phi) \times \mathbb{I}_{[\dot{\phi}_{min} \leq \dot{\phi}_{k+1} \leq \dot{\phi}_{max}]},$$

where $\mathbb{I}_{[x]}$ takes the value 1 when x is true and zero otherwise. The velocity of the bar pointer is defined to be proportional to tempo.

At each time index k, a rhythmic pattern indicator, R_k, takes one value in a finite set, for example $\mathcal{S} = \{0, 1\}$. The elements of the set $\mathcal{S}$ correspond to different rhythmic patterns, which are described in further detail below, with examples given in figure 30. For now we deal with the simple case in which there are only two such patterns, and switching between values of R_k is modelled as occurring if a bar line is crossed, i.e.:

if $\phi_k < \phi_{k-1}$,

$$Pr(R_k = s|r_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} p_r, & s \neq r_{k-1}, \\ 1 - p_r, & s = r_{k-1}, \end{cases}$$

otherwise, $R_k = r_{k-1}$, where p_r is the probability of a change in rhythmic pattern. In summary, $\mathbf{x}_k \equiv [\phi_k \ \dot{\phi}_k \ r_k]^T$ specifies the state of the system at time index k. We remark that the dynamical model can be extended to include the musical notion of *meter*, see (Whiteley, Cemgil, and Godsill 2006) for details.

6.2 Poisson Observation Model

MIDI onset events are treated as being Poisson distributed with an intensity parameter which is conditioned on the position of the bar-pointer and the rhythm indicator variable. Defining the Poisson intensity in this fashion allows quantification of the postulate that for a given rhythm, there are regions in one bar in which onsets occur with high probability.

Each *rhythmic pattern function*, $\mu_r : [0, 1) \to \mathbb{R}^+$, maps the position of the bar pointer ϕ_k to the mean of a gamma prior distribution on an intensity parameter λ_k. The value of $\mu_r(\phi_k)$

combined with a constant variance Q_λ, determines the shape and rate parameters of the gamma distribution:

$$a_r(\phi_k) = \mu_r(\phi_k)^2/Q_\lambda,$$
$$b_r(\phi_k) = \mu_r(\phi_k)/Q_\lambda,$$

For brevity, denote $a_k \equiv a_r(\phi_k)$, and $b_k \equiv b_r(\phi_k)$. Then conditional on ϕ_k and r_k, the prior density over λ_k is:

$$p(\lambda_k|\phi_k, r_k) = \begin{cases} \lambda_k^{a_k-1}\frac{b_k^{a_k}\exp(-b_k\lambda)}{\Gamma(a_k)}, & \lambda_k \geq 0 \\ 0, & \lambda_k < 0 \end{cases}$$

This combination of prior distributions provides robustness against variation in the data.

Let y_k denote the number of onset events observed in the kth non-overlapping frame of length Δ, centred at time t_k. The number y_k is modelled as being Poisson distributed as follows:

$$p(y_k|\lambda_k) = \frac{(\lambda_k\Delta)^{y_k}\exp(-\lambda_k\Delta)}{y_k!}.$$

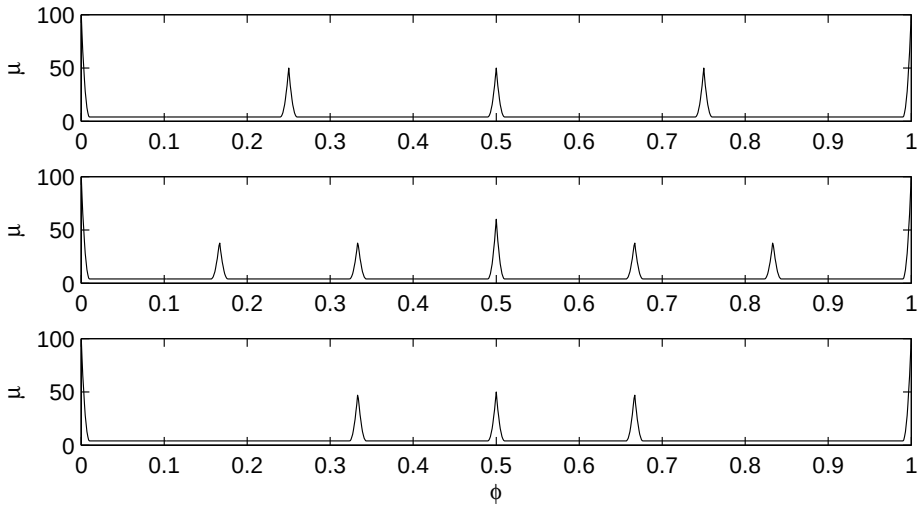


Figure 30: Examples of rhythmic pattern functions, each corresponding to a different value of r_k. Top - a bar of duplets in 4/4 meter, middle - a bar of triplets in 4/4 meter, bottom - 2 against 3 polyrhythm. The widths of the peaks model arpeggiation of chords and expressive performance. Construction in terms of splines permits flat regions between peaks, corresponding to an onset event 'noise floor'.

Inference for the intensity λ_k is not required so it is integrated out. This may be done analytically, yielding:

$$\begin{aligned} p(y_k|\phi_k, r_k) &= \int_0^\infty p(y_k|\lambda_k)p(\lambda_k|\phi_k, r_k)d\lambda_k \\ &= \frac{b_k^{a_k}\Gamma(a_k + y_k)}{y_k!\Gamma(a_k)(b_k + \Delta)^{a_k+y_k}}. \end{aligned}$$

6.3 Gaussian Observation Model

For every time index k, denote by z_k a vector of ν samples constituting the kth non-overlapping frame of a raw audio signal. The time interval Δ is then given by $\Delta = \nu/f_s$, where f_s is the

sampling rate. The samples are modelled as independent with a zero mean Gaussian distribution:

$$p(z_k|\sigma_k^2) = \frac{1}{(2\pi\sigma_k^2)^{\nu/2}} \exp\left(-\frac{z_k^T z_k}{2\sigma_k^2}\right).$$

An inverse-gamma distribution is placed on the variance σ_k^2. The shape and scale parameters of this distribution, denoted by c_k and d_k respectively, are determined by the location of the bar pointer, ϕ_k and the rhythmic pattern indicator variable r_k, again via a rhythmic pattern function, μ_k.

$$p(\sigma_k^2|\phi_k, r_k) = \frac{d_k^{c_k} \exp(-d_k/\sigma_z^2)}{\Gamma(c_k)} \sigma_k^{-2(c_k+1)},$$

$$c_k = \mu_k^2/Q_s + 2,$$

$$d_k = \mu_k \left(\frac{\mu_k^2}{Q_s} + 1\right),$$

where Q_s is the variance of the inverse-gamma distribution and is chosen to be constant.

The variance of the Gaussian distribution, σ_k^2, may be integrated out analytically to yield:

$$p(z_k|\phi_k, r_k) = \frac{d_k^{c_k}\Gamma(c_k + \nu/2)}{(2\pi)^{\nu/2}\Gamma(c_k)} \left(\frac{z_k^T z_k}{2} + d_k\right)^{-(c_k+\nu/2)}.$$

6.4 Results

Of practical interest are the posterior filtering and smoothing distributions, e.g for some date record of length K frames, $p(\phi_{1:K}, \dot{\phi}_{1:K}, r_{1:K}|y_{1:K})$ and its marginals.

The performance of this model is demonstrated by a smoothing task for an excerpt of a MIDI performance of 'Michelle' by the Beatles, using the Poisson observation model (results for the Gaussian observation model are reported in (Whiteley, Cemgil, and Godsill 2006)). This demonstrates the joint tempo-tracking and rhythm recognition capability of the method. The performance, by a professional pianist, was recorded using a Yamaha Disklavier C3 Pro Grand Piano. The top two rhythmic patterns in figure 30 were employed. For purposes of exposition, the state-space of the bar pointer was uniformally discretised to $M = 1000$ position and $N = 20$ velocity points. The discretised position and velocity of bar pointer are denoted by m_k and n_k respectively, with the discretised dynamical model being that with probability 0.99, $n_k = n_{k-1}$ and, within the allowed range, n_k is incremented or decremented with equi-probability. Uniform initial prior distributions were set on m_k, n_k and r_k. The time frame length was set to $\Delta = 0.02s$, corresponding to the range of tempi: $12 - 240$ quarter notes per minute. The probability of a change in rhythmic pattern was set to $p_r = 0.1$. The variance of the gamma distribution was set $Q_\lambda = 10$.

This section of 'Michelle' presents a significant challenge for tempo tracking because of the triplets, each of which by definition has a duration of 3/2 quarter notes. A performance of this excerpt could be wrongly interpreted as having a local change in tempo in the second bar, when in fact the rate of quarter notes remains constant; the bar of triplets is just a change in rhythm.

For the discretised model, exact inference is possible (Sequential Monte Carlo methods for the non-discretised model are investigated in (Whiteley, Cemgil, and Godsill 2007)). In figure 31, the strong diagonal stripes in the image of the posterior smoothing distributions for m_k correspond to the *maximum a-posteriori* (MAP) trajectory of the bar pointer. The change to a triplet rhythm in the second bar and the subsequent reversion to duplet rhythm are correctly identified. The MAP tempo is given by the darkest stripe in the image for the velocity log-smoothing distribution - it is roughly constant throughout.

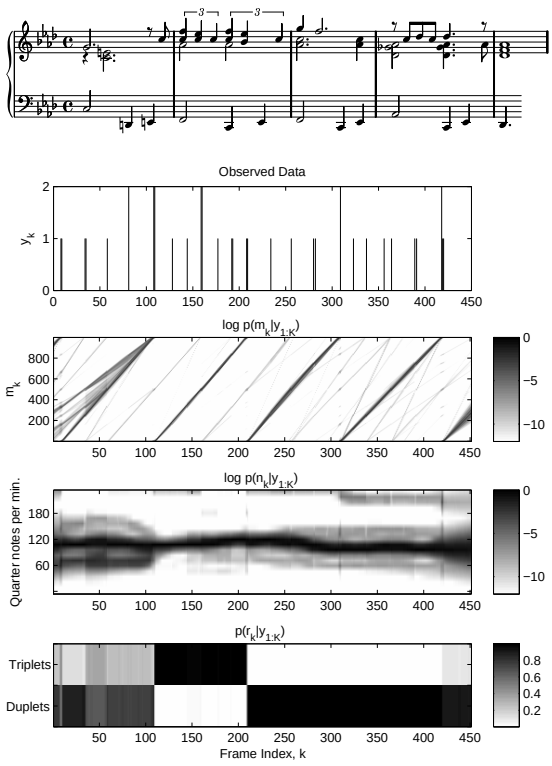


Figure 31: Results for joint tempo tracking and rhythmic pattern recognition on a MIDI performance of 'Michelle' by the Beatles. The top figure is the score which the pianist was given to play. Each image consists of smoothing marginal distributions for each frame index.

7 Discussion and Conclusions

In this review, we sketched Bayesian methods for analysis of audio signals. The Bayesian models exhibit complex statistical structure and in practice, highly adaptive and powerful computational techniques are needed to perform inference.

In this chapter, we have reviewed some of these statistical models, including generalised linear models, dynamic systems and other hierarchical models, and described how various problems in audio and music processing can be cast into Bayesian posterior inference. We have also illustrated inference methods based on Monte Carlo simulation or other deterministic techniques (such as mean field, variational Bayes) originating in statistical physics to tackle computational problems posed by inference in these models. We describes models in both the time domain and transform domains, the latter typically offering greater computational tractability and modelling flexibility at the expense of some accuracy in the models.

The Bayesian approach has two key advantages over more traditional engineering solutions: it provides both a suite for model construction and a framework for algorithm development. Apart from the pedagogical advantages (such as highlighting algorithmic similarities, convergence characteristics and computational requirements), the framework facilitates development of sophisticated models and to automation of code generation procedures. We believe that the field of computer hearing, which is still in its infancy compared to topics such as computer vision and speech recognition, has great potential for advancement in coming years, with the advent of powerful Bayesian inference methodologies and accompanying computational power increases.

A Review of sinusoidal models

This appendix introduces several low-level, deterministic representations of audio signals and highlights the links between various representations such as damped sinusoids, state space and Gabor representations. In the main text, these representations are used to formulate statistical models for observed data, given the values of latent parameters. These deterministic representations are highly structured and the statistical models into which they are assimilated therefore exhibit rich temporal and spectral properties. Furthermore, the specific parameterisations of these signal models allows their statistical counter-parts to be coupled to higher-level model components, such as change-point processes and model-order indicators.

A.1 Static Models

Sound signals are emitted by vibrating objects that can be modelled as a cascade of second order systems. Hence, it is convenient to represent them as a sum of sinusoids and a transient non-periodic component (See e.g. (McAulay and Quatieri 1986; Serra and Smith 1991; Rodet 1998; Irizarry 2002; Parra and Jain 2001; Davy and Godsill 2003)). We will start our exposition using a deterministic and static model, which we later generalise in several directions. For a time series y_n of length N with $n = 0 \dots N-1$, the (deterministic) discrete time sinusoidal model can be written as

$$y_n = \sum_{\nu=0}^{W-1} \alpha_\nu e^{-\gamma_\nu n} \cos(\omega_\nu n + \phi_\nu) \tag{23}$$

Here, $\nu = 0 \dots W-1$ denotes the sinusoidal index and W is the total number of sinusoidal components. The parameters of this model are all real numbers, amplitudes α_ν, log-damping coefficients $\gamma_\nu > 0$, the frequencies ω_ν and the phases ϕ_ν. Using the fact that $\cos(\theta) = (e^{j\theta} + e^{-j\theta})/2$ we can write

$$y_n = \sum_{\nu=0}^{W-1} (c_\nu z_\nu^n + c_\nu^* z_\nu^{*n}) \tag{24}$$

where c_ν is the complex amplitude $c_\nu = (\alpha_\nu/2)e^{j\phi_\nu}$ and z_ν is the complex pole $z_\nu = e^{-\gamma_\nu + j\omega_\nu}$. It is obvious that we can write the model in matrix notation as:

$$\bar{y} \equiv \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ z_0 & z_0^* & \dots & z_{W-1} & z_{W-1}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_0^n & z_0^{*n} & \dots & z_{W-1}^n & z_{W-1}^{*n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_0^{N-1} & z_0^{*N-1} & \dots & z_{W-1}^{N-1} & z_{W-1}^{*N-1} \end{pmatrix} \begin{pmatrix} c_0 \\ c_0^* \\ \vdots \\ c_{W-1} \\ c_{W-1}^* \end{pmatrix} \equiv F_{0:N-1} c$$

$F_{0:N-1} = F_{0:N-1}(\boldsymbol{\gamma}, \boldsymbol{\omega})$ is $N \times 2W$ and each column is a damped complex exponential. The $n+1$'th row will be denoted as F_n. It is possible to estimate all parameters using subspace techniques (Laroche 1993; Badeau, Boyer, and David 2002).

Many popular models appear as special cases of the above model, and can be obtained by tying various parameters. The *harmonic model*, with fundamental frequency ω and some log-damping coefficient γ is defined as (23) by the following

$$\omega_\nu = \nu\omega \qquad \gamma_\nu = \nu\gamma \tag{25}$$

This model is particularly useful for pitched musical instruments that tend to create oscillations with modes that are roughly related by integer ratios. It is possible to let the damping coefficients free; or tie them as above to model the fact that higher frequencies get damped faster. A related model is the *inharmonic model* that assumes

$$\omega_\nu = B(\nu, \omega) \tag{26}$$

where B is a function that "stretchs" the harmonic frequencies. Stretching of harmonics is a phenomena that is observed especially in piano strings.

It is important to note that the inverse Fourier transform is obtained as a special case of the harmonic model by ignoring the damping $\gamma_\nu = \gamma = 0$ and choosing the frequencies on a uniform grid with $\omega_\nu = 2\pi\nu/N$ for $\nu = 0 \dots W-1$ with $W = N/2$. We call all these models and their variations *static* models.

A.2 State space Representations and Dynamic Models

An alternative, *dynamic* representation exploits the *rotational invariance property*:

$$\begin{pmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{n+L-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ z_0 & z_0^* & \dots & z_{W-1} & z_{W-1}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_0^{L-1} & z_0^{*L-1} & \dots & z_{W-1}^{L-1} & z_{W-1}^{*L-1} \end{pmatrix} \begin{pmatrix} z_0 & & & & \\ & z_0^* & & & \\ & & \ddots & & \\ & & & z_{W-1} & \\ & & & & z_{W-1}^* \end{pmatrix}^n \begin{pmatrix} c_0 \\ c_0^* \\ \vdots \\ c_{W-1} \\ c_{W-1}^* \end{pmatrix}$$

$$y_{n:n+L-1} = F_{0:L-1}\, \mathbf{diag}(Z_0, Z_1, \dots, Z_{W-1})^n c \tag{27}$$

with $Z_\nu = Z_\nu(\gamma_\nu, \omega_\nu) \equiv \mathbf{diag}(z_\nu, z_\nu^*)$, $c \equiv \begin{pmatrix} c_0 & c_0^* & \dots & c_{W-1} & c_{W-1}^* \end{pmatrix}^\top$. By appropriately rotating the complex amplitudes, the basis matrix stays shift invariant in time. This representation allows us to write a model for arbitrary frame lengths and "interpolate" between dynamic and static interpretations. For example, given a suitable frame length L, we can reorganise the time series in N/L nonoverlapping frames (assuming L divides N). Then we can write using (27)

$$\bar{y}_k \equiv (y_{kL}, \dots, y_{(k+1)L-1})^\top = F_{0:L-1}\, \mathbf{diag}(Z_0^L, Z_1^L, \dots, Z_{W-1}^L)^k c$$

Here, k denotes a frame index such that $k = 0 \dots N/L - 1$. One can easily envision models where subsequent frames are possibly of different length L_k such that $\sum_k L_k = N$. Of course, mathematically speaking, we haven't really gained anything and in the noiseless case, all formulations for any frame length L are equivalent. However, the dynamical system perspective opens up interesting possibilities to extend the basic model in a way that is not apparent in the static formulation. In particular, we can write the following state space parametrisation:

$$\begin{aligned} s_0^{\mathrm{d}} &= \begin{pmatrix} c_0 & c_0^* & \dots & c_{W-1} & c_{W-1}^* \end{pmatrix}^\top = c \\ s_{k+1}^{\mathrm{d}} &= A^{\mathrm{d}} s_k^{\mathrm{d}} \\ \bar{y}_k &= F_{0:L-1} s_k^{\mathrm{d}} \end{aligned}$$

where $A^{\mathrm{d}} \equiv \mathbf{diag}(Z_1^L, Z_2^L, \dots, Z_{W-1}^L)$ is a $2W \times 2W$ diagonal matrix with poles appearing in conjugate pairs and $F_{0:L-1}$ is $L \times 2W$, as defined in (27). This representation is known as the diagonal realisation. It is well known that the representation of $\bar{y}_k$ is not unique; for any invertible transform matrix $\mathcal{T}$, we could define $s_k = \mathcal{T} s_k^{\mathrm{d}}$

$$\mathcal{T} s_{k+1}^{\mathrm{d}} = \mathcal{T} A^{\mathrm{d}} \mathcal{T}^{-1} \mathcal{T} s_k^{\mathrm{d}} \qquad \Longleftrightarrow \qquad s_{k+1} = A s_k \tag{28}$$

$$\bar{y}_k = F_{0:L-1} \mathcal{T}^{-1} \mathcal{T} s_k^{\mathrm{d}} \qquad \Longleftrightarrow \qquad \bar{y}_k = C_{0:L-1} s_k \tag{29}$$

with new state evolution and observation matrices $A \equiv \mathcal{T} A^{\mathrm{d}} \mathcal{T}^{-1}$ and $C_{0:L-1} \equiv F_{0:L-1} \mathcal{T}^{-1}$. In particular, to avoid complex arithmetic, one can apply the following block diagonal transform

$$\begin{aligned} \mathcal{T} &= \mathbf{diag}(T, T, \ldots, T) \\ T &\equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ j & -j \end{pmatrix} \qquad\qquad TT^H = T^H T = I \end{aligned} \tag{30}$$

This renders the $n+1$'th row of $C_{0:L-1}$

$$\begin{aligned} C_n &= F_n \mathcal{T}^H \\ &= \sqrt{2} \begin{pmatrix} e^{-n\gamma_0} \cos(n\omega_0) & e^{-n\gamma_0} \sin(n\omega_0) & \ldots & e^{-n\gamma_{W-1}} \cos(n\omega_{W-1}) & e^{-n\gamma_{W-1}} \sin(n\omega_{W-1}) \end{pmatrix} \end{aligned} \tag{31, 32}$$

Further we obtain

$$\begin{aligned} A &= \mathbf{diag}(A_0, \ldots, A_\nu, \ldots, A_{W-1}) \\ A_\nu &= A_\nu(\gamma_\nu, \omega_\nu) = T Z_\nu^L T^H = e^{-L\gamma_\nu} \begin{pmatrix} \cos(L\omega_\nu) & -\sin(L\omega_\nu) \\ \sin(L\omega_\nu) & \cos(L\omega_\nu) \end{pmatrix}^\top \end{aligned} \tag{33}$$

$$\begin{aligned} s_0 &= \begin{pmatrix} s_{0,0}^\top & \ldots & s_{0,\nu}^\top & \ldots & s_{0,W-1}^\top \end{pmatrix}^\top \\ s_{0,\nu} &= T \begin{pmatrix} c_\nu & c_\nu^* \end{pmatrix}^\top = \sqrt{2} \begin{pmatrix} \mathrm{Re}\{c_\nu\} & -\mathrm{Im}\{c_\nu\} \end{pmatrix}^\top \end{aligned} \tag{34}$$

The harmonic and inharmonic models of the previous section can be written in state space form by constraining the poles to have related frequencies. For example, the harmonic model with damping coefficient γ and fundamental frequency ω can be written as

$$A_\nu = A_\nu(\gamma_\nu, \omega_\nu) = T Z(\gamma, \omega)^{L\nu} T^H = e^{-L\gamma\nu} \begin{pmatrix} \cos(L\omega\nu) & -\sin(L\omega\nu) \\ \sin(L\omega\nu) & \cos(L\omega\nu) \end{pmatrix}^\top$$

A.3 Nonstationarity and Time-Frequency Representations

The sinusoidal model, as we have introduced it in the previous section, is rather restrictive for modelling long sequences and realisations from nonstationary processes, as the poles of the model are taken to be fixed. One possibility is to assume that the poles are time varying; where one can take

$$A_{\nu,k} = A_\nu(\gamma_{\nu,k}, \omega_{\nu,k}) \qquad\qquad C_k = C(\gamma_{0:W-1,k}, \omega_{0:W-1,k})$$

However, the associated estimation problem is highly nonlinear. The inverse problem is also ill posed as one can generate any signal if one can arbitrarily modulate the pole frequencies and damping coefficients. Hence, an appropriate regulariser, such as a random walk in pole frequencies, needs to be assumed. In certain cases, such as when the harmonic model is used, the pole frequencies can be discretised and deterministic grid based inference techniques (Tabrikian, Dubnov, and Dikalov 2004; Cemgil, Kappen, and Barber 2004) or more powerful sequential Monte Carlo techniques can be employed (Dubois and Davy 2007).

An arguably simpler approach is to use a basis that is localised both in time and frequency. Such representations are known as *Gabor representations*. In the Gabor representation (Mallat 1999; Wolfe, Godsill, and Ng 2004), a real valued time series is represented on a $W \times K$ time-frequency lattice by

$$y_n = \sum_{\tau=0}^{K-1} \sum_{\nu=0}^{W-1} \left(c_{\nu,\tau} h_{n,\tau} z_\nu^n + c_{\nu,\tau}^* h_{n,\tau} z_\nu^{*n} \right)$$

where $\nu = 0 \dots W-1$ is the frequency band index and $\tau = 0 \dots K-1$ is the time frame index, where the poles are $z_\nu = z_\nu(\omega) = e^{j\omega\nu}$ with $\omega = 2\pi/W$. The coefficients $h_{n,\tau}$ are fixed and are determined from a prototype *window function* $h(n)$ as

$$h_{n,\tau} \;=\; h(n - \tau\frac{N}{K})$$

The real valued and non-negative window function is typically chosen as a symmetric bell-shaped curve with compact support[5] to give a suitable time-frequency localisation. The ratio N/K denotes the effective number of samples that we shift the window with each frame. The ratio KW/N is the oversampling rate and gives an indication of the redundancy of the representation. It is informative to write the model in matrix form

$$\bar{y} \;=\; Gc$$

where the expansion coefficients are given as

$$c \;=\; \begin{pmatrix} c_{0,0} & c^*_{0,0} & \dots & c_{1,\tau} & \dots & c_{\nu,1} & \dots & c^*_{W-1,K-1} \end{pmatrix}^\top$$

and G is given as a $N \times 2WK$ matrix

$$\begin{pmatrix} g_{0,0}(0) & g^*_{0,0}(0) & \dots & g_{0,\tau}(0) & \dots & g_{\nu,0}(0) & \dots & g^*_{W-1,K-1}(0) \\ \vdots & \dots & \dots & \vdots & \vdots & \vdots & \vdots & \\ g_{0,0}(n) & g^*_{0,0}(n) & \dots & g_{0,\tau}(n) & \dots & g_{\nu,0}(n) & \dots & g^*_{W-1,K-1}(n) \\ \vdots & \dots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ g_{0,0}(N-1) & g^*_{0,0}(N-1) & \dots & g_{0,\tau}(N-1) & \dots & g_{\nu,0}(N-1) & \dots & g^*_{W-1,K-1}(N-1) \end{pmatrix} \quad (35)$$

Here, each column is a basis vector with a modulated and translated version of the window function $g_{\nu,\tau}(n) = h_{n,\tau} z_\nu^n$. To avoid complex arithmetic, the block diagonal transformation in (30) can be employed as $\bar{y} = G\mathcal{T}^{-1}\mathcal{T}c = \tilde{G}s$ to obtain real matrices and expansion coefficients, as in (32) and (34). For certain choices of the time-frequency lattice parameters with $KW = N$, the matrix G can be rendered square and orthonormal $G^H G = I$. One such choice is taking the window h rectangular with a support of $L = N/K$ samples such that subsequent windows do not overlap in time and using $W = L$ frequency bins. This is equivalent to modelling the sequence as independent time frames. Many other techniques, such as lapped orthogonal transforms or modified discrete cosine transforms, to obtain orthagonality with other type of window and basis functions exist and it is beyond our purpose to review this rather broad literature.

The Gabor representation is in a way a generalisation of the sinusoidal model in Eq. (24), that represents the signal by exponentially decaying window function $e^{-\gamma n}$, using a single sinusoid per frequency band. In contrast, the Gabor representation chooses at each time frame and each frequency band a new expansion coefficient, hence does not enforce phase continuity and allows for smooth amplitude variations. One potential shortcomming of the Gabor representation is that it is not very powerful in representing abrupt changes in the signal characteristics, such as chaingepoints or frequency modulations.

References

Abdallah, S. A. and M. D. Plumbley (2006, January). Unsupervised analysis of polyphonic music using sparse coding. *IEEE Transactions on Neural Networks 17*(1), 179–196.

[5]However, the original paper by Gabor assumed Gaussian functions with infinite support

Badeau, R., R. Boyer, and B. David (2002, September). Eds parametric modelling and tracking of audio signals. In *DAFx-02*, Hamburg, Germany.

Bertin, N., R. Badeau, and G. Richard (2007). Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. of Internation conference on audio, speech and signal processing (ICASSP)*, Honolulu.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cemgil, A. T. (2004). *Bayesian Music Transcription*. Ph. D. thesis, Radboud University of Nijmegen.

Cemgil, A. T. (2007). Strategies for sequential inference in factorial switching state space models. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 07)*, Honolulu, Hawaii, pp. 513–516.

Cemgil, A. T. (2008, July). Bayesian inference in non-negative matrix factorisation models. Technical Report CUED/F-INFENG/TR.609, University of Cambridge.

Cemgil, A. T. and O. Dikmen (2007, September). Conjugate gamma Markov random fields for modelling nonstationary sources. In *ICA 2007, 7th International Conference on Independent Component Analysis and Signal Separation*.

Cemgil, A. T., S. J. Godsill, and C. Fevotte (2007). Variational and Stochastic Inference for Bayesian Source Separation. *Digital Signal Processing 17*.

Cemgil, A. T., H. J. Kappen, and D. Barber (2004). A generative model for music transcription. *Accepted to IEEE Transactions on Speech and Audio Processing*.

Cemgil, A. T., P. Peeling, O. Dikmen, and S. J. Godsill (2007, October). Prior structures for time-frequency energy distributions. In *Accepted to Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. New York: John Wiley & Sons, Inc.

Davy, M., S. Godsill, and J. Idier (2006, April). Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America 119*(4).

Davy, M. and S. J. Godsill (2002). Detection of abrupt spectral changes using support vector machines. an application to audio signal segmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

Dubois, C. and M. Davy (2007, May). Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach. *IEEE transactions on Speech, Audio and Language Processing 15*(4), 1283–1295.

Fearnhead, P. (2003). Exact and efficient bayesian inference for multiple changepoint problems. Technical report, Dept. of Math. and Stat., Lancaster University.

Févotte, C., L. Daudet, S. J. Godsill, and B. Torrésani (2006, May). Sparse regression with structured priors: Application to audio denoising. In *Proc. ICASSP*, Toulouse, France.

Févotte, C. and S. Godsill (2006). A Bayesian approach for blind separation of sparse sources. *IEEE Trans. on Speech and Audio Processing*.

Fletcher, N. H. and T. Rossing (1998). *The Physics of Musical Instruments*. Springer.

Ghahramani, Z. and M. Beal (2000). Propagation algorithms for variational Bayesian learning. In *Neural Information Processing Systems 13*.

Godsill, S. (2004). Computational modeling of musical signals. *Chance Magazine (American Statistical Association 17*(4).

Godsill, S. and M. Davy (2005, October). Bayesian computational models for inharmonicity in musical instruments. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY.

Godsill, S. J. and M. Davy (2002). Bayesian harmonic models for musical pitch estimation and analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

Green, P. J. (1995). Reversible jump Markov-chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes* (Third Edition ed.). Oxford University Press.

Hamacher, V., J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass (2005). Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP J. Appl. Signal Process. 2005*(1), 2915–2929.

Irizarry, R. A. (2002). Weighted estimation of harmonic components in a musical sound signal. *Journal of Time Series Analysis 23*.

Kameoka, H. (2007). *Statistical Approach to Multipitch Analysis*. Ph. D. thesis, University of Tokyo.

Klapuri, A. and M. Davy (Eds.) (2006). *Signal Processing Methods for Music Transcription*. New York: Springer.

Knuth, K. H. (1998, Jul.). Bayesian source separation and localization. In *SPIE'98: Bayesian Inference for Inverse Problems*, San diego, pp. 147–158.

Laroche, J. (1993). Use of the matrix pencil method for the spectrum analysis of musical signals. *Journal of Acoustical Society of America 94*, 1958–1965.

Lee, D. D. and H. S. Seung (2000). Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press.

McAulay, R. J. and T. F. Quatieri (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*(4), 744–754.

McIntyre, M. E., R. T. Schumacher, and J. Woodhouse (1983). On the oscillations of musical instruments. *J. Acoustical Society of America 74*, 1325–1345.

Miskin, J. and D. Mackay (2001). Ensemble learning for blind source separation. In S. J. Roberts and R. M. Everson (Eds.), *Independent Component Analysis*, pp. 209–233. Cambridge University Press.

Mohammad-Djafari, A. (1997, Jul.). A Bayesian estimation method for detection, localisation and estimation of superposed sources in remote sensing. In *SPIE'97*, San Diego.

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

Peeling, P. H., C. Li, and S. J. Godsill (2007, April). Poisson point process modeling for polyphonic music transcription. *Journal of the Acoustical Society of America Express Letters 121*(4), EL168–EL175. Reused with permission from Paul Peeling, The Journal of the Acoustical Society of America, 121, EL168 (2007). Copyright 2007, Acoustical Society of America.

Reyes-Gomez, M., N. Jojic, and D. Ellis (2005). Deformable spectrograms. In *AI and Statistics Conference*, Barbados.

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

Rowe, D. B. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. Chapan & Hall/CRC.

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

Shepard, N. (Ed.) (2005). *Stochastic Volatility, Selected Readings*. Oxford University Press.

Smaragdis, P. and J. Brown (2003). Non-negative matrix factorization for polyphonic music transcription. In *WASPAA, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Tabrikian, J., S. Dubnov, and Y. Dikalov (2004, Jan). Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. *IEEE Transactions on Speech and Audio Processing 12*(1), 76–87.

Vincent, E., N. Bertin, and R. Badeau (2008). Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas. IEEE.

Virtanen, T. (2006, November). *Sound Source Separation in Monaural Music Signals*. Ph. D. thesis, Tampere University of Technology.

Walmsley, P., S. J. Godsill, and P. J. W. Rayner (1998, September). Multidimensional optimisation of harmonic signals. In *Proc. European Conference on Signal Processing*.

Walmsley, P. J., S. J. Godsill, and P. J. W. Rayner (1999, October). Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State*, Mohonk, NY State.

Wang, D. and G. J. Brown (Eds.) (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, Applications*. Wiley.

Whiteley, N., A. T. Cemgil, and S. J. Godsill (2006). Bayesian modelling of temporal structure in musical audio. In *Proceedings of International Conference on Music Information Retrieval*, Victoria, Canada.

Whiteley, N., A. T. Cemgil, and S. J. Godsill (2007, April). Sequential Inference of Rhythmic Structure in Musical Audio. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 07)*, pp. 1321–1324. IEEE.

Wolfe, P. J., S. J. Godsill, and W. Ng (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *Journal of the Royal Statistical Society*.