·25·
Bayesian statistical methods for audio and music processing

A. Taylan Cemgil, Simon J. Godsill, Paul Peeling and Nick Whiteley

25.1 Introduction

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 not yet been possible to construct machines that can simulate the effectiveness of human listening – for example, an expert human listener can accurately write down a fairly complex musical score based solely on listening to the audio. 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 for the 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 semantic content of the music, are able to extract structure and can organize information directly from acoustic input. For generating convincing performances, they need to be able to analyse and mimic master musicians. These outstanding technological challenges motivate this research, in which fundamental modelling principles are applied to gain as much information as possible from ambiguous audio data.

Musical audio 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 classification, denoising, source separation or score transcription. The ultimate goal here is to construct computer systems that display aspects of human, or super-human, performance levels in an automated fashion. In the second, the goal is to aid the scientific understanding of cognitive processes behind the human auditory system (Moore 1997) and the physical sound generation process of musical instruments or voices (Fletcher and Rossing 1998).

The starting point in this chapter is that in both contexts, scientific and technological, Bayesian statistical methods provide a sound formalism for making progress. This is achieved via models which quantify prior knowledge about the physical properties and semantics of sound, combined with powerful computational methodology. 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 made from the posterior distribution for the hidden structure given observed audio data. One of the strengths of this simple and intuitive view of audio processing is that it unifies a variety of tasks such as source tracking, enhancement, transcription, separation, identification or resynthesis into a single Bayesian inference framework. 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 is in contrast with many of the more heuristic and ad hoc approaches to audio processing. Popular aproaches here involve the design of custom-built algorithms for solving specific tasks, and in which the problem formulation and computational solution are blended together, taking account of practical and pragmatic considerations only. These techniques potentially miss out on the generality and accuracy afforded by a well-defined Bayesian model and associated estimation algorithms.

We firstly consider mainstream applications of audio signal processing, give a very brief introduction to the properties of musical audio, and then proceed to pose the principal challenges as Bayesian inference tasks.

25.1.1 Applications

A fundamental task that will be a focus of this paper is music-to-score transcription (Cemgil 2004; Klapuri and Davy 2006). 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 and durations to those of the composer. The subtask 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.

Signal separation is a second fundamental challenge (Hyvärinen, Karhunen, and Oja 2001; Virtanen 2006b) – 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 hearing aids, see below. Instrument classification is a further important component of musical analysis systems, i.e. the task of recognizing which instruments are playing at any given time in a piece. 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.; Herrera-Boyer, Klapuri, and Davy 2006).

Finally, at the signal level, audio restoration and enhancement (Godsill and Rayner 1998) form another key area. 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 amounts of audio data available digitally in on-line repositories, it is not unreasonable 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 methodologies.

25.1.2 Introduction to musical audio

The following discussion gives a basic introduction to some of the properties of musical audio signals, following closely that of Godsill (2004). Musical audio is highly *structured*, both in the time domain and in the frequency domain.

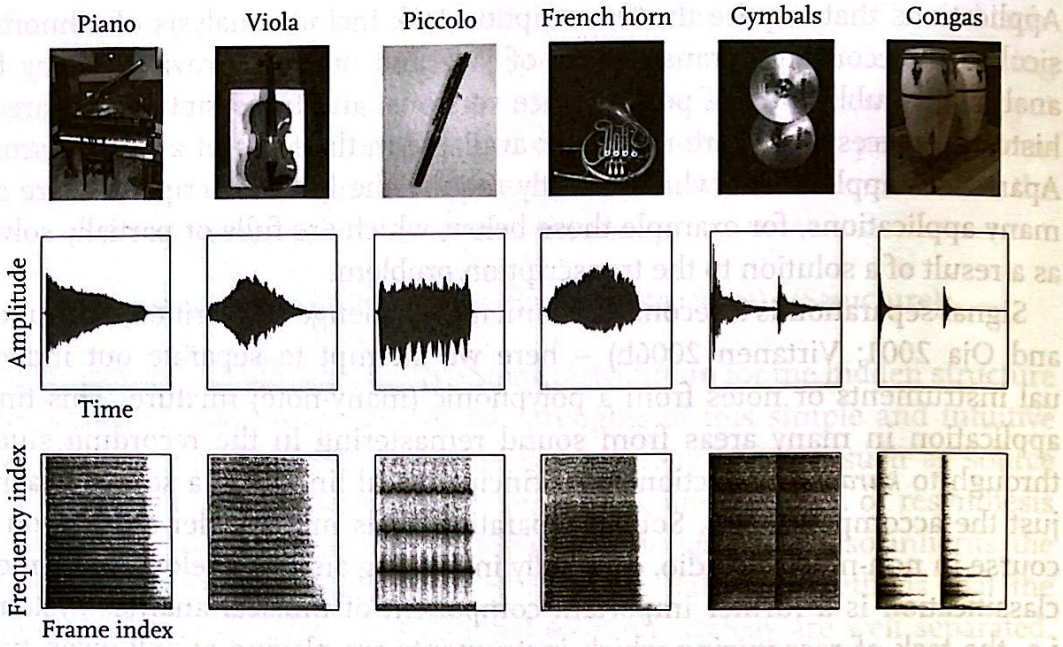

Fig. 25.1 Some acoustical instruments, examples of typical time series and corresponding spectrograms (time varying magnitude spectra – modulus of short time Fourier transform) computed with FFT. (Audio data and images from RWCP Instrument samples database.)

In the time domain, *tempo* and *beat* specify the range of likely times where note transitions occur. 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 nonlinear 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. See Figure 25.1 for examples of pitched and percussive musical instruments and typical time series. Second, several notes played at the same time form chords, or polyphony (Figure 25.2). 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, 659 Hz $\approx 5/4 \times 523$ Hz and 785 Hz $\approx 3/2 \times 523$ Hz. Figure 25.3 shows the waveform for a simple monophonic (single note) flute recording and Figure 25.4 shows the corresponding time – frequency spectrogram analysis (this may be auditioned at `www-sigproc.eng.cam.ac.uk/~sjg/haba`, where other extracts used in this paper may also be found). In this both the temporal segmentation and the frequency-domain structure are clearly visible on the plot. Focusing on a single localized time frame, at around 2 s in the same extract, we can clearly see

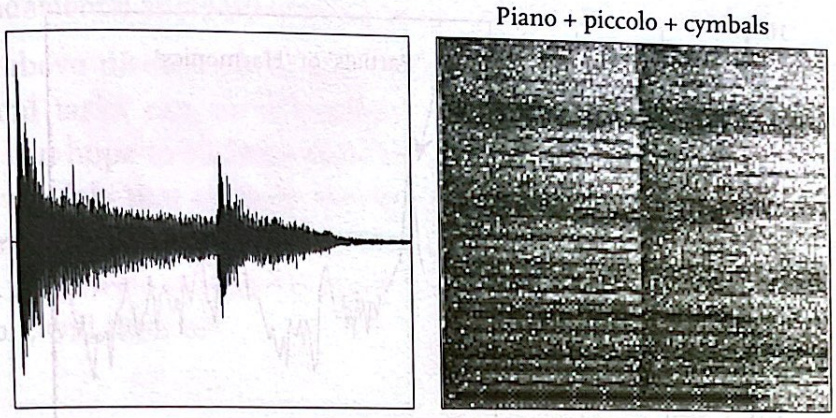

Fig. 25.2 Superposition. The time series and the magnitude spectrogram of the resulting signal when some of the instruments play concurrently.

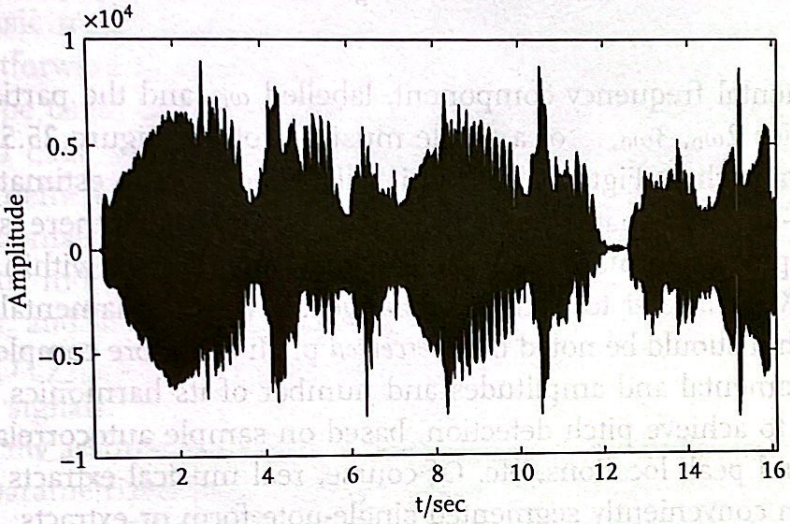

Fig. 25.3 Time-domain waveform for a solo flute extract.

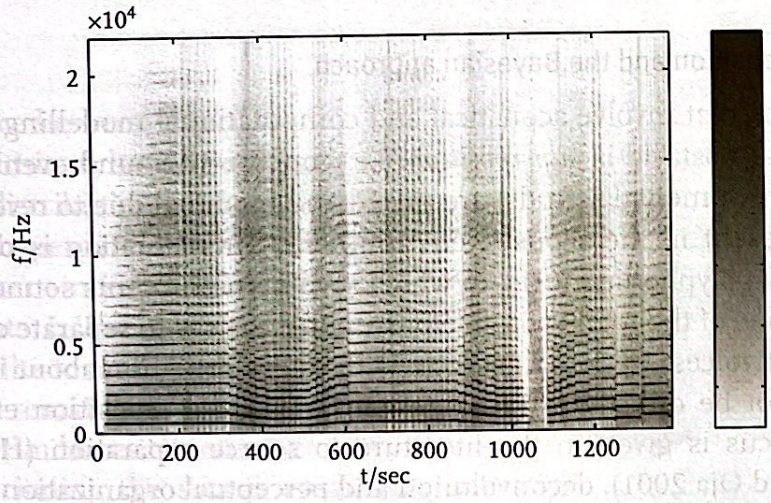

Fig. 25.4 Time – frequency spectrogram representation for the flute recording.

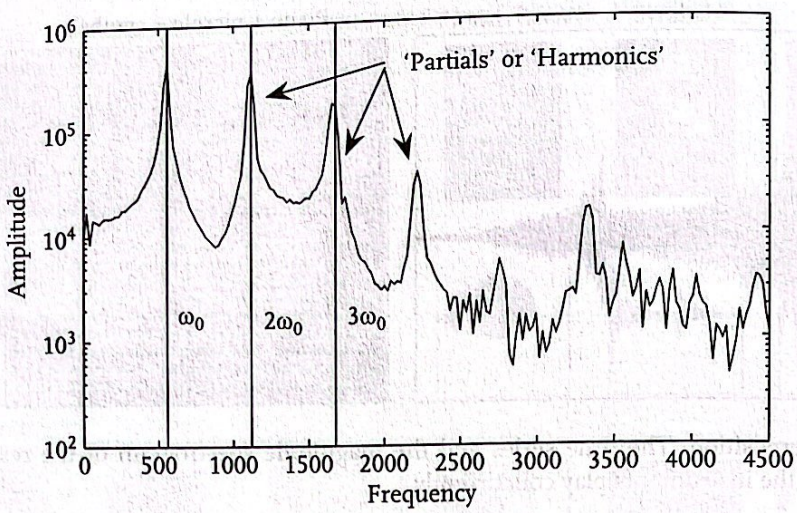

Fig. 25.5 Short-time Fourier analysis of a single frame of data from the flute extract.

the fundamental frequency component, labelled ω_0, and the partial stucture, at frequencies $2\omega_0$, $3\omega_0, \ldots$ of a single musical note in Figure 25.5. It is clear from spectra such as Figure 25.5 that it will be possible to estimate the pitch 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). We will refer to pitch interchangeably with fundamental frequency ω_0, although it should be noted that *perceived* pitch is a more complex function of the fundamental and amplitudes and number of its harmonics. There are many ways to achieve pitch detection, based on sample autocorrelation functions, spectral peak locations, etc. Of course, real musical extracts don't usually arrive in conveniently segmented single-note form or extracts, and much more complex structures need to be considered, as detailed in the sections below.

25.1.3 Superposition and the Bayesian approach

In applications that involve acoustical and computational modelling of sound, a fundamental obstacle is *superposition*, i.e. concurrent sound events (music, speech or environmental sounds) are mixed and modified due to reverberation and noise present in the acoustic environment. This situation is of primary importance in polyphonic music, in which several instruments sound simultaneously and one of the many possible processing goals is to separate or identify the individual voices. In domains such as these, information about individual sources cannot be directly extracted, owing to the superposition effect, and significant focus is given in the literature to source separation (Hyvärinen, Karhunen, and Oja 2001), deconvolution and perceptual organization of sound (Wang and Brown 2006).

25.1.4 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:

$$Y = \sum_{i=1}^{I} s_i \tag{25.1}$$

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 multichannel 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 render 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 $p(Y|\sum_{i=1}^{I} s_i)$ that models an additive noise component in addition to the desired signals.

We typically assume that the individual sources s_i are independent a priori. They are parametrized 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 previously identified.

25.1.4.1 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, Fevotte, and Godsill 2007)

$$p(S|Y) = \frac{1}{P(Y)} \int p(Y|S, \Lambda)p(S|\Theta, \Lambda)p(\Lambda)p(\Theta)d\Lambda d\Theta \qquad (25.2)$$

where, under our deterministic model above in equation (25.1), 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 both cases one has to be especially careful with the interpretation of expectations in models where likelihoods and priors are invariant to source permutations.

25.1.4.2 Polyphonic music transcription

Music transcription refers to extraction of a human readable and interpretable description from a recording of a music performance, see Figure 25.6. In cases where more than a single musical note plays at a given time instant, we term this task *polyphonic music transcription* and we are once again in the superposition regime. The general task of interest is 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

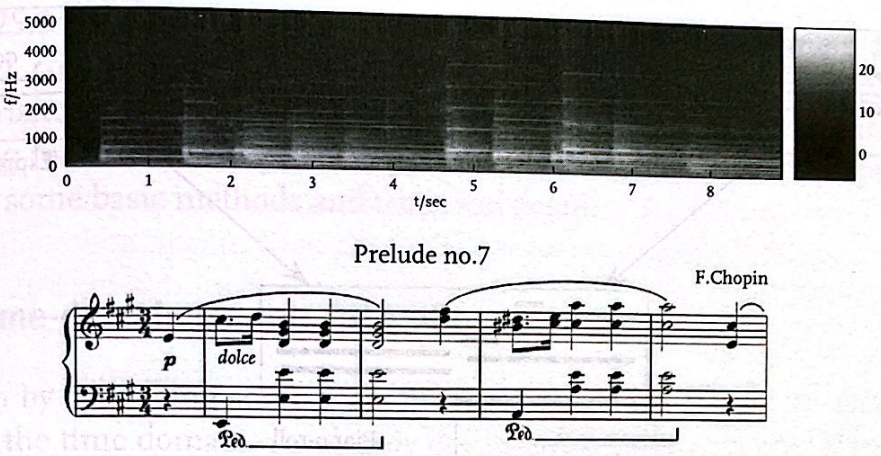

Fig. 25.6 Polyphonic music transcription. The task is to generate a human readable score as shown below, given the acoustic input. The computational problem here is to infer pitch, number of notes, rhythm, tempo, meter, time signature. The inference can be achieved online (filtering) or offline (smoothing), depending upon requirements.

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.

Music transcription is an inference problem in which we wish to find a musical 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 description of their intrinsic properties such as shape, colour or relative position. We view music transcription from the same perspective, where we wish to 'explain' individual samples of a music signal in terms of a collection of musical objects and 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.[1] The performers add their interpretation to music and render the score into a collection of

[1] 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.

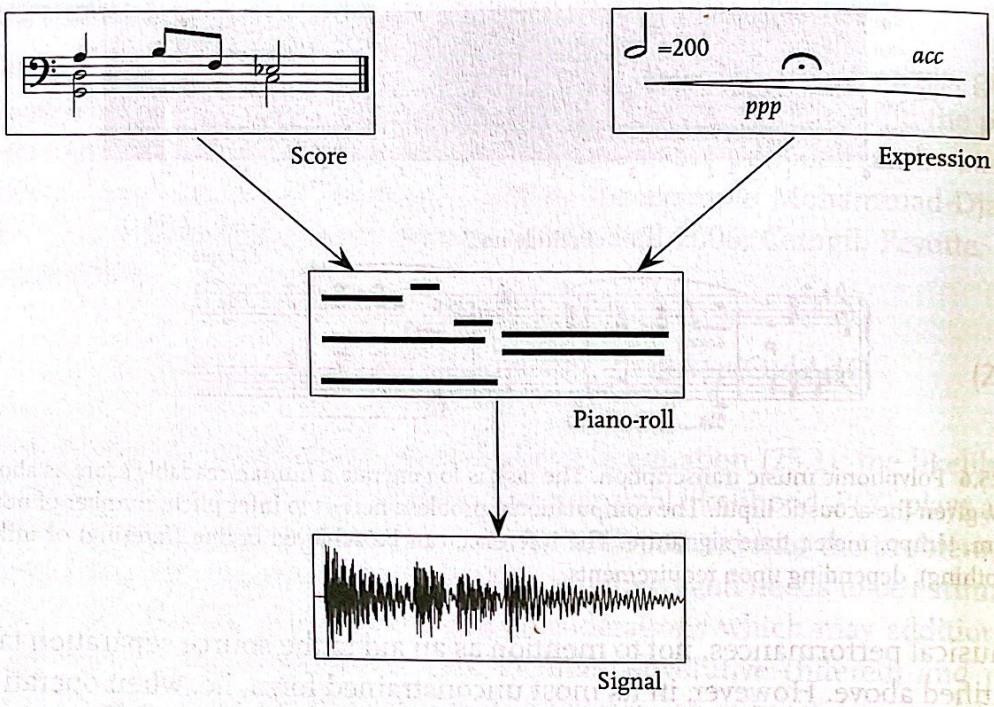

Fig. 25.7 A hierarchical generative model for music transcription. In this model, an unknown score is rendered by a performer into a 'piano roll'. The performer introduces expressive timing deviations and tempo fluctuations. The piano roll is rendered into audio by a synthesis model. The piano roll can 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 chapter can be viewed as inference techniques as applied to subgraphs of this graphical model.

'control signals'. Further down at the physical level, the control signals trigger various musical instruments that synthesize the observed sound signal. We illustrate these generative processes using a hierarchical graphical model (see Figure 25.7), where the arcs represent generative links.

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 but still very challenging task, we may try and invert back only as far as the piano roll.

25.1.5 Organization of the chapter

In Section 25.2, signal models for audio are developed in the time domain, including some examples of their inference for a musical acoustics problem.

Section 25.3 describes models in the frequency transform domain that lead to greater computational tractability. In particular, we describe new dependence structures across time and frequency that allow for very accurate prior modelling for the audio. A final conclusion section is followed by appendices covering some basic methods and technical detail.

25.2 Time-domain models for audio

We begin by describing some basic note and chord models for musical audio, based in the time 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

$$s_{i,t} = \sum_{m=1}^{M_i} a_{m,i} \cos(m\omega_{0,i}t) + \beta_{m,i} \sin(m\omega_{0,i}t) \tag{25.3}$$

for $t \in \{0, \ldots, N-1\}$. Here, $M_i > 0$ is the number of partials present in note i, $\sqrt{a_{m,i}^2 + \beta_{m,i}^2}$ gives the amplitude of a partial and $\tan^{-1}(\beta_{m,i}/a_{m,i})$ 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 $a_{m,i}$, $\beta_{m,i}$, 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 equation 25.1). 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 considerable 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, 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-idealized 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. In particular, it is fairly straightforward to modify the model so that the partial amplitudes α and β may vary with time,

$$s_{i,t} = \sum_{m=1}^{M_i} a_{m,i,t} \cos(m\omega_{0,i}t) + \beta_{m,i,t} \sin(m\omega_{0,i}t) \tag{25.4}$$

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

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

In our work we have adopted 50%-overlapped Hamming windows for the basis functions, see Figure 25.8, 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 $a_{m,i,t}$ in the state-space formulation. 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.

25.2.1 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 $\{\omega_{0,i}, M_i, \alpha_i, \beta_i\}$, where α_i, β_i are the vectors of parameters $a_{m,i}, \beta_{m,i}$,

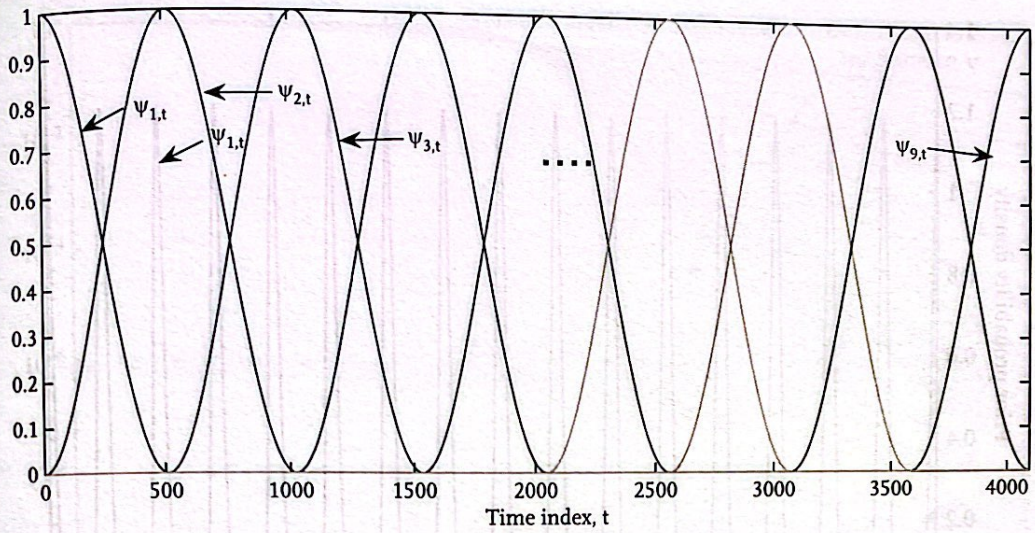

Fig. 25.8 Basis functions $\psi_{i,t}$, $I = 9$, 50% overlapped Hamming windows.

$m = 1, 2, \ldots, M_i$. 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[2] which labels middle C (or 'C4') as note number 60, and all other notes as integers relative to this – the A below this would be 57, for example, and the A above middle C (usually at 440 Hz 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:

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

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

$$\mu(n) = (n - 69)/12 \log(2) + \log(440/\omega_s) \tag{25.5}$$

where once again ω_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 Figure 25.9, capturing the expected clustering of note frequencies at semitone spacings relative to A440.

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

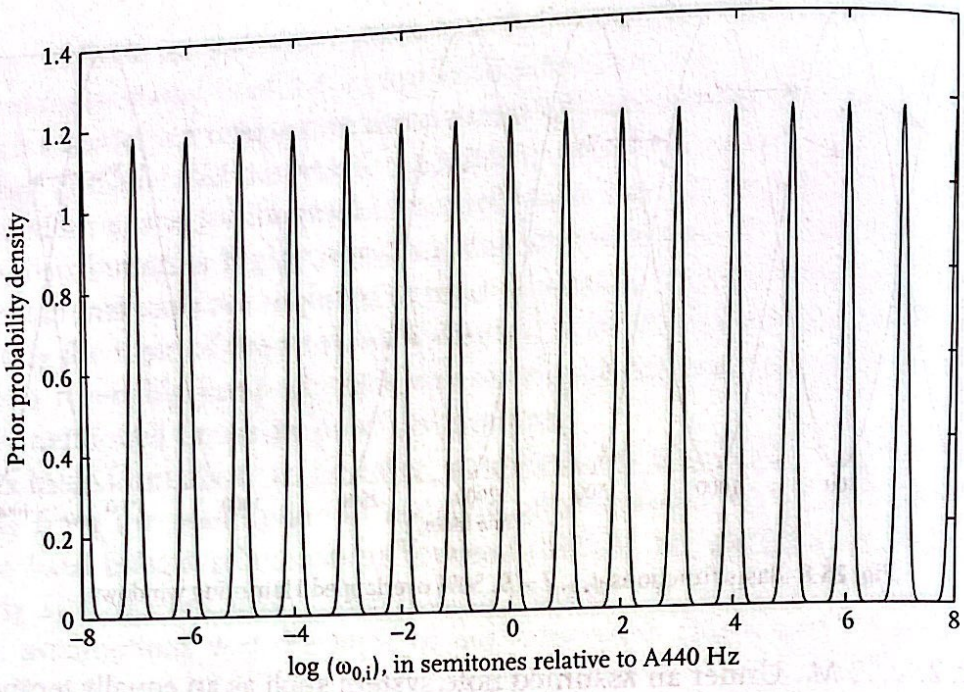

Fig. 25.9 Prior for fundamental frequency $p(\omega_{0,i})$.

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 $\{M_{\min}, \ldots, M_{\max}\}$, with limits truncated to prevent partials at frequencies greater than $\omega_s/2$, the Nyquist rate. Secondly, a prior for the amplitude parameters α_i, β_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 lower 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 to decay with increasing frequency of the partial number m. The general form of this is

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

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

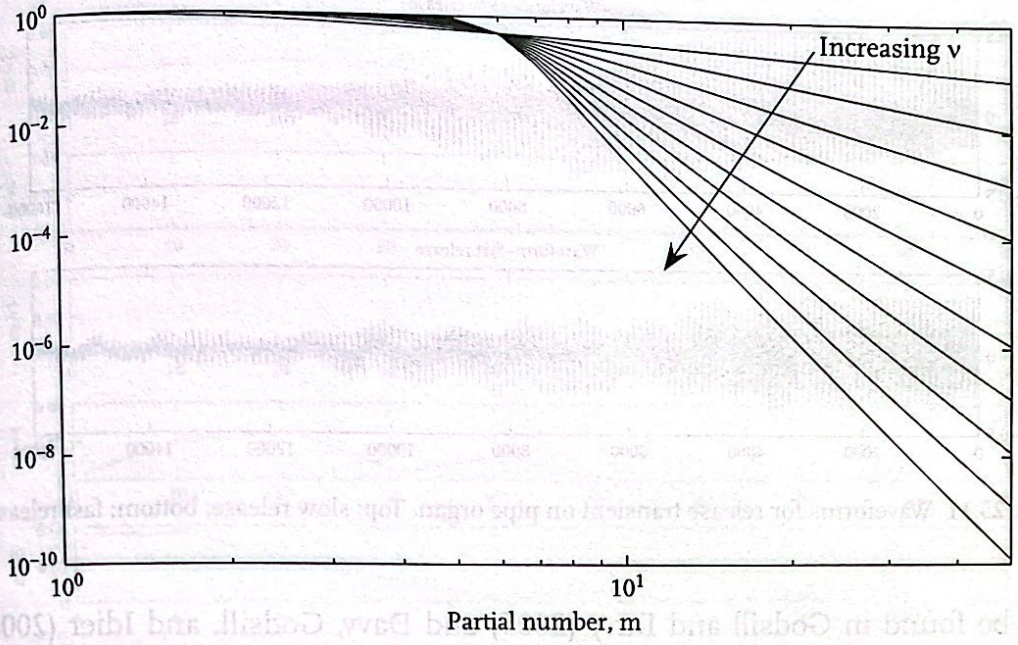

Fig. 25.10 Family of k_m curves (log-log plot), $T = 5$, $\nu = 1, \ldots, 10$.

with $T = 5$, $\nu = 1, 2, \ldots, 10$, Figure 25.10. 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 a and β parameters is readily extended to the time-varying amplitude case of equation (25.4), 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 the individual notes and the number of notes present:

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

where

$$\theta_i = \{n_i, \omega_{0,i}, M_i, a_i, \beta_i, g_i\}.$$

This model can be explored using MCMC methods, in particular the reversible jump MCMC method (Green 1995), and results from this and related models

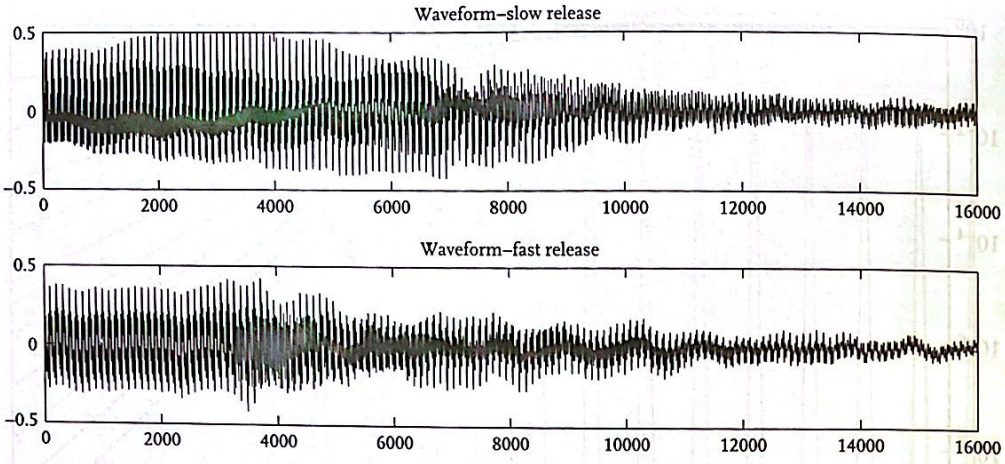

Fig. 25.11 Waveforms for release transient on pipe organ. Top: slow release; bottom: fast release.

can be found in Godsill and Davy (2005) and 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.

25.2.2 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 Figure 25.11. 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 equation (25.4). A fixed length Hamming window of duration 0.093 s was used for the basis functions. The resulting MCMC output can be used in many ways. For example, examination of the expansion coefficients α_i and β_i allows an analysis of how the partials vary with time under each playing condition. In both cases the reversible jump MCMC identifies nine significant partials in the data. In Figures 25.12 and 25.13 we plot the first five ($m = 1, \ldots, 5$) partial energies $a_{m,i}^2 + b_{m,i}^2$ as a function of time.

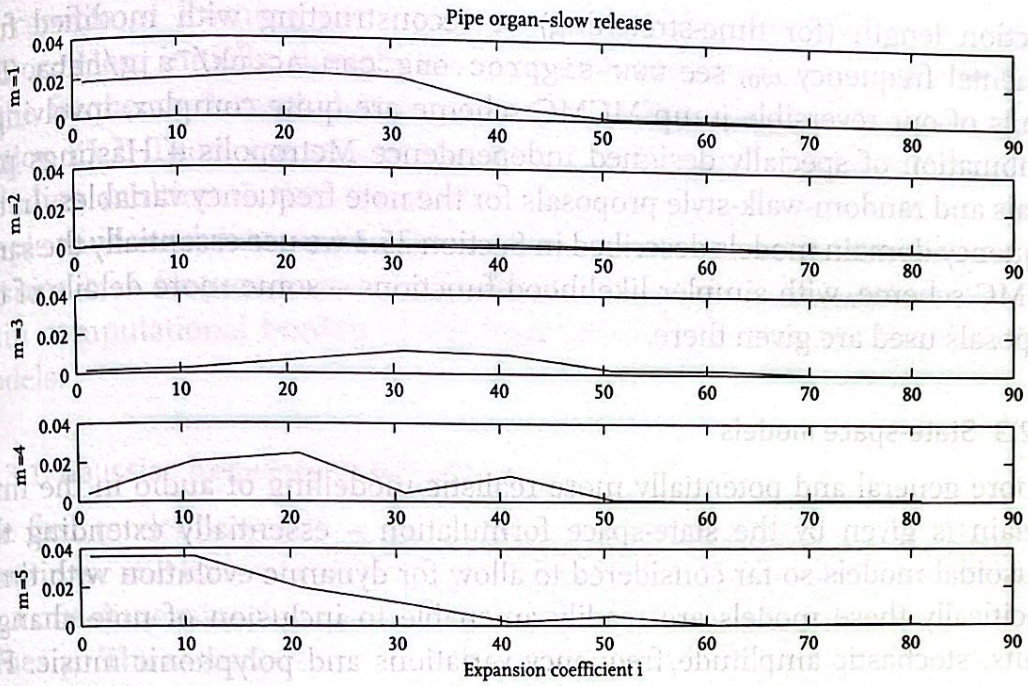

Fig. 25.12 Magnitudes of partials with time: slow release.

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 Hamming window basis

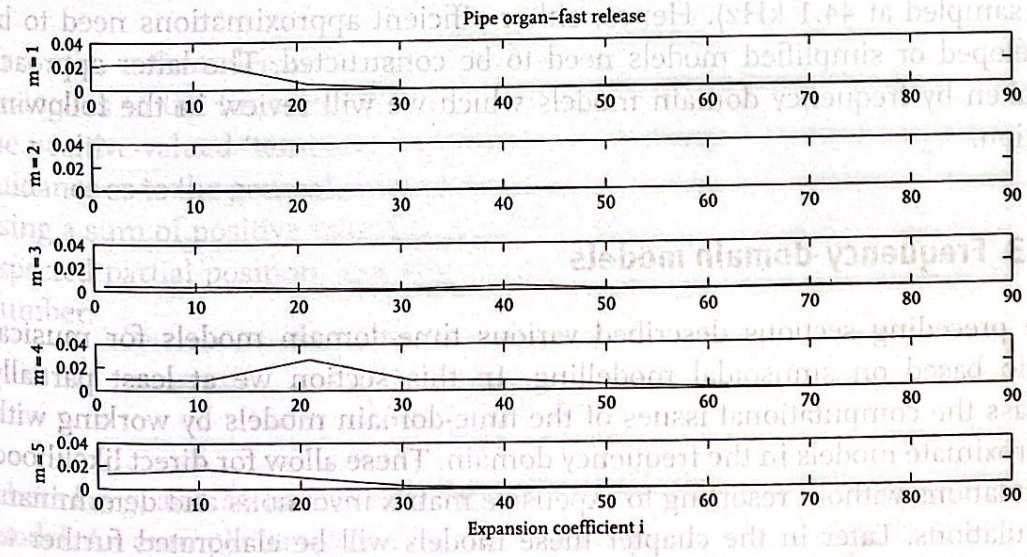

Fig. 25.13 Magnitudes of partials with time: fast release.

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 25.3 we use essentially the same MCMC scheme, with simpler likelihood functions – some more details of the proposals used are given there.

25.2.3 State-space models

A more general and potentially more realistic modelling of audio in the time domain is given by the state-space formulation – essentially extending the sinusoidal models so far considered to allow for dynamic evolution with time. Specifically these models are readily amenable to inclusion of note change-points, stochastic amplitude/frequency variations and polyphonic music. For space reasons we do not include any detailed discussion here but the interested reader is referred to Cemgil, Kappen, and Barber (2006) and Cemgil (2007). Such state-space 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 – in particular the calculation of marginal likelihood terms in the presence of many Gaussian components α_i and β_i – can become very expensive. 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. The latter approach is taken by frequency domain models which we will review in the following section.

25.3 Frequency-domain models

The preceding sections described various time-domain models for musical audio based on sinusoidal modelling. In this section we at least partially bypass the computational issues of the time-domain models 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 factorization 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 25.2. There are several routes to a frequency-domain representation, including multiresolution 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.

25.3.1 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:

$$s_{i,\nu} \sim N_C(0, \lambda_\nu(\theta_i))$$

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 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:

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

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, Figure 25.5 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:

$$\lambda_\nu(\theta_i) = \sum_{m=1}^{M_i} g_i^2 k_m \phi_{\nu - m\omega_{0,i}} \tag{25.6}$$

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 Figure 25.14, in which a Gaussian pulse shape has been utilized.

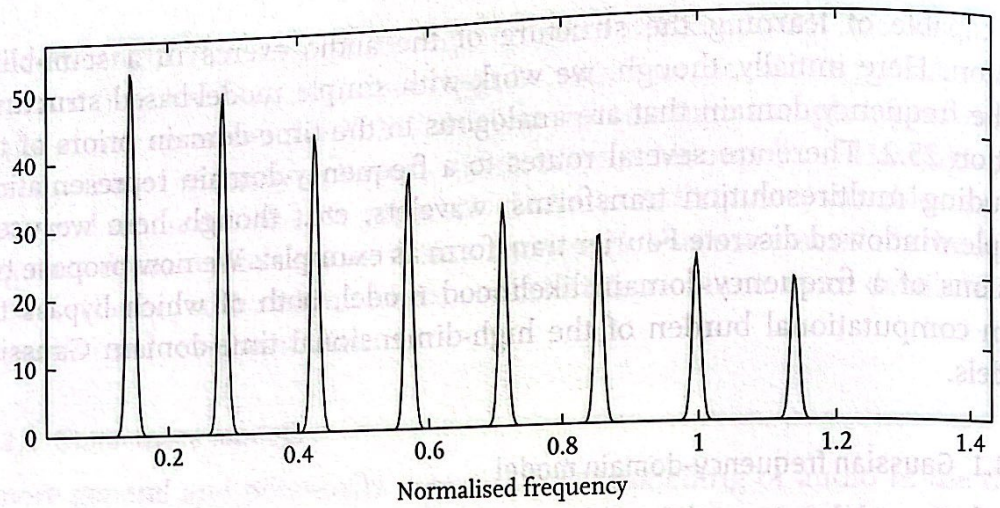

Fig. 25.14 Template function $\lambda_\nu(\theta_i)$ with $M_i = 8$, $\omega_{0,i} = 0.71$, Gaussian pulse shape.

25.3.2 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 Figure 25.5). 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 $\{p_j\}_{j=1}^J$, $\nu_{\min} < p_j < \nu_{\max}$. Then the likelihood may be readily computed using:

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

where $Z(\Theta) = \int_{\nu_{\min}}^{\nu_{\max}} (S_{v,\nu} + \sum_{i=1}^{I} \lambda_\nu(\theta_i)) d\nu$ is the normalizing 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, equation (25.6). A modified form of this likelihood function was successfully applied for chord detection problems by Peeling, Li, and Godsill (2007).

25.3.3 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 0.1 s of data, see Figure 25.15. 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 25.2.1 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 25.3.1, the MCMC burn-in for the note number vector $n = [n_1, n_2, \ldots, n_I]$ is shown in Figure 25.16. 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 specialized Metropolis – Hastings moves which centre 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

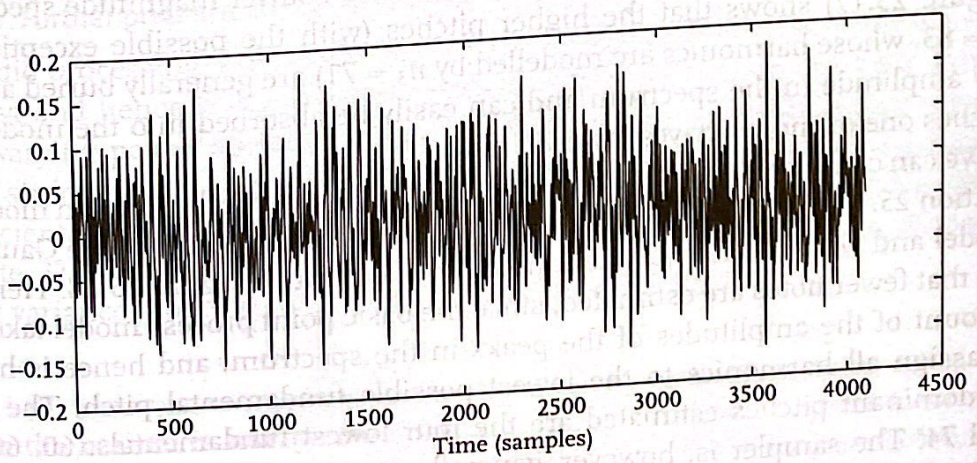

Fig. 25.15 Audio waveform – single chord data.

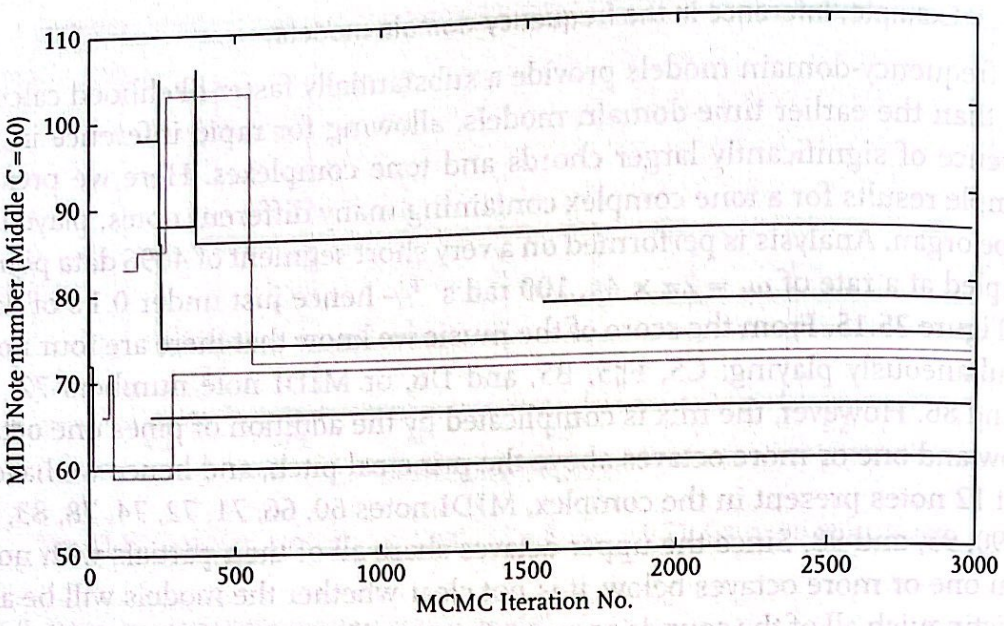

Fig. 25.16 Evolution of the note number vector with iteration number – single chord data. Gaussian frequency-domain model.

faster than for the earlier proposed time-domain models, particularly in terms of the model order sampling, which was here initialized at $I = 1$, i.e. one single note present at the start of the chain. Specialized 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 seven out of the (at least) 12 possible pitches present in the music: 60, 66, 71, 72, 74, 78, 86. The remaining five 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 (Figure 25.17) 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 25.3.2. The MCMC was run under identical conditions to the Gaussian model and we plot the equivalent note index output in Figure 25.18. 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.

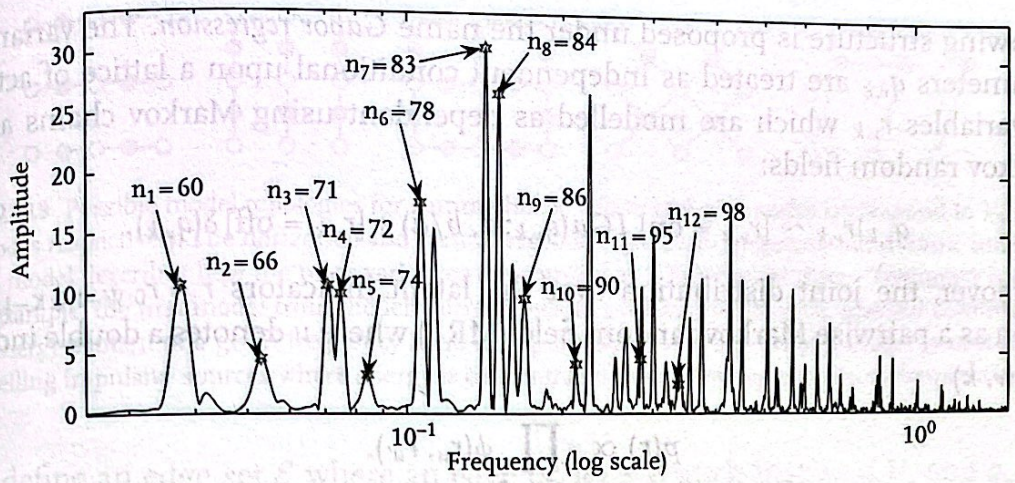

Fig. 25.17 Discrete Fourier magnitude spectrum for 12-note chord. True note positions marked with a pentagram.

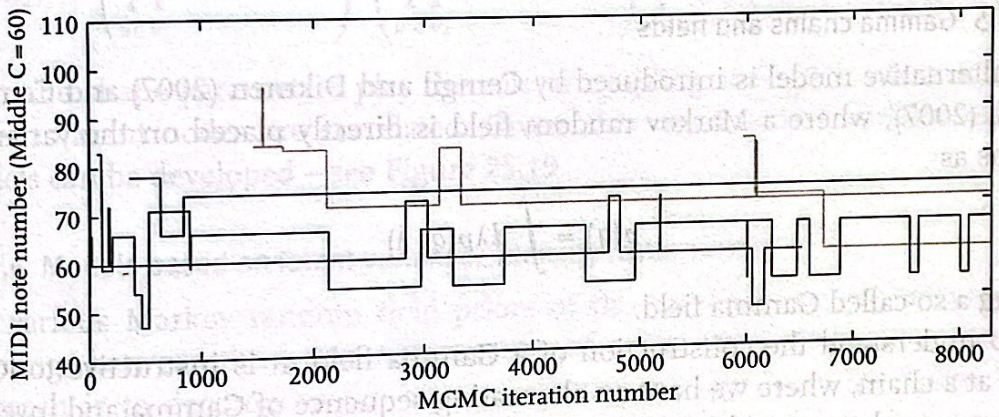

Fig. 25.18 Evolution of the note number vector with iteration number – single chord data. Poisson frequency-domain model.

25.3.4 Further prior structures for transform domain representations

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

$$s_{\nu,k}|q_{\nu,k} \sim N(s_{\nu,k}; 0, q_{\nu,k})$$

where the normal is interpreted either as complex Gaussian or real Gaussian depending on the transform used – the Fourier representation is complex, the discrete sine/cosine representation is real. In Wolfe, Godsill, and Ng (2004), the

following structure is proposed under the name *Gabor regression*. The variance parameters $q_{\nu,k}$ are treated as independent conditional upon a lattice of activity variables $r_{\nu,k}$ which are modelled as dependent using Markov chains and Markov random fields:

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

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

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

Several MRF constructions are considered, including Markov chains across time or frequency and Ising-type models.

25.3.5 Gamma chains and fields

An alternative model is introduced by Cemgil and Dikmen (2007) and Cemgil *et al.* (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

$$q_u|\lambda_u \sim IGa(q_u; a_q, a_q\lambda) \qquad \lambda_{u+1}|q_u \sim Ga(\lambda_{u+1}; a_\lambda, q_u/a_\lambda).$$

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

$$p(q_u|\lambda_u, \lambda_{u+1}) \propto IGa(q_u; a_q + a_\lambda, a_q\lambda_u + a_\lambda\lambda_{u+1})$$
$$p(\lambda_u|q_{u-1}, q_u) \propto Ga\left(\lambda_u; a_\lambda + a_q, a_\lambda q_{u-1}^{-1} + a_q q_u^{-1}\right).$$

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 type of *stochastic volatility* model useful for characterization of non-stationary behaviour observed in, for example, financial 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 Ga and IGa respectively.

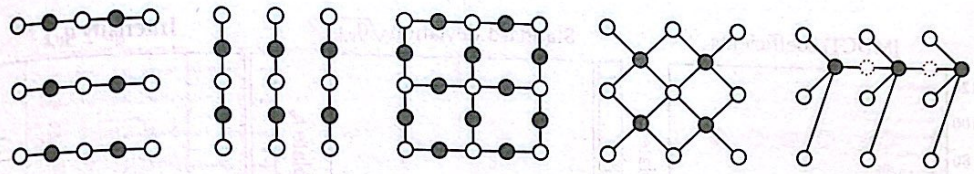

Fig. 25.19 Possible model topologies for gamma fields. White and grey nodes correspond to $\mathcal{V}_q$ and $\mathcal{V}_\lambda$ nodes respectively. The horizontal and vertical axis corresponds to frequency ν and frame index k. Each model describes how the prior variances are coupled as a function of time – frequency index. For example, the first model from the left corresponds to a source model with 'spectral continuity', the 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.

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 factorization

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

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 – see Figure 25.19.

25.3.6 Models based on latent variance/intensity factorization

The various Markov random field priors of the previous section introduced couplings between the latent variances $q_{\nu,k}$. Another alternative and powerful approach is to decompose the latent variances as a product. We define the following hierarchical model (see Figure 25.21)

$$\begin{aligned} s_{\nu,k} &\sim N(s_{\nu,k}; 0, q_{\nu,k}) & q_{\nu,k} &= t_\nu v_k \\ t_\nu &\sim IGa\left(t_\nu; a_\nu^t, a_\nu^t b_\nu^t\right) & v_k &\sim IGa\left(v_k; a_k^v, a_k^v b_k^v\right). \end{aligned} \tag{25.7}$$

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 index this template is modulated by v_ν, to adjust the overall volume. An example is given in Figure 25.20 to represent a piano sound. The template gives the harmonic structure of the pitch and the excitation characterises the time varying energy.

A simple factorial model that uses the Gamma chain prior models introduced in Section 25.3.5 is constructed as follows:

$$x_{\nu,k} = \sum_i s_{\nu,i,k} \quad s_{\nu,i,k} \sim N(s_{\nu,i,k}; 0, q_{\nu,i,k}) \quad Q = \{q_{\nu,i,k}\} \sim p(Q|\Theta^t). \tag{25.8}$$

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

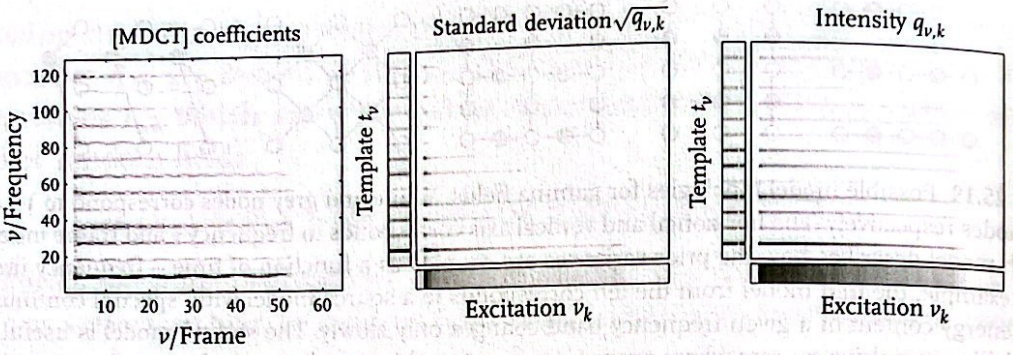

Fig. 25.20 (Left) the spectrogram of a piano $|s_{v,k}|^2$. (Middle) estimated templates and excitations using the conditionally Gaussian model defined in equation (25.7), where $q_{v,k}$ is the latent variance. (Right) estimated templates and excitations using the conditionally Poisson model defined in the next section.

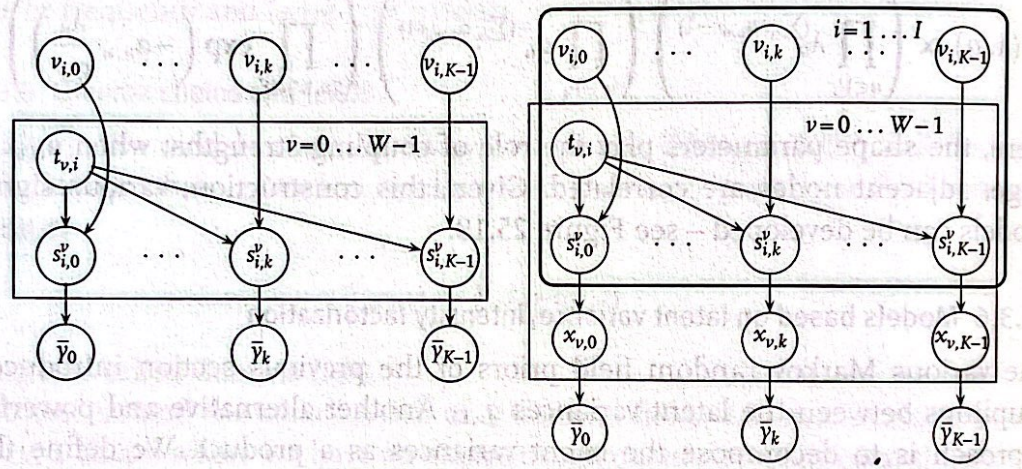

Fig. 25.21 (Left) latent variance/intensity models in product form (equation 25.7). Hyperparameters are not shown. (Right) factorial version of the same model, used for polyphonic estimation as used in Section 25.3.7.1.

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

$$p(s_{v,i,k}|X, Q) = N(s_{v,i,k}; \kappa_{v,i,k}x_{v,k}, q_{v,i,k}(1 - \kappa_{v,i,k}))$$
$$\kappa_{v,i,k} = q_{v,i,k} / \sum_{i'} q_{v,i',k}.$$

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 $\prod_v p(q_{v,i=1,1:K})$. This model simply assumes

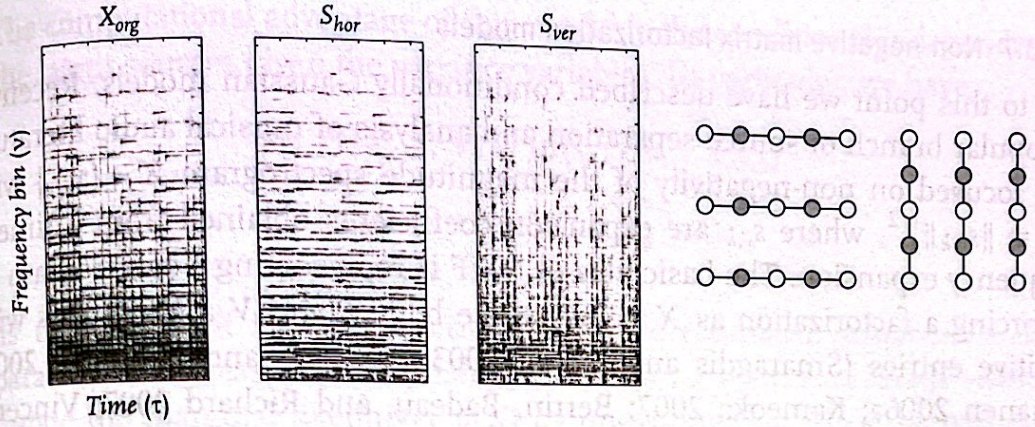

Fig. 25.22 Single channel source separation example, left to right, log-MDCT coefficients of the original signal and reconstruction with horizontal and vertical IGMRF models.

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 Figure 25.1). 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 25.22, 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 factorization models described in equation 25.7 have also straightforward factorial extensions

$$x_{\nu,k} = \sum_i s_{\nu,i,k}$$

$$s_{\nu,i,k} \sim N(s_{\nu,i,k}; 0, q_{\nu,i,k}) \qquad q_{\nu,i,k} = t_{\nu,i} v_{i,k} \tag{25.9}$$

$$T = \{t_{\nu,i}\} \sim p(T|\Theta^t) \qquad V = \{v_{i,k}\} \sim p(V|\Theta^v). \tag{25.10}$$

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

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

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.

25.3.7 Non-negative matrix factorization models

Up to this point 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 of NMF is representing a spectrogram by enforcing a factorization as $X \approx TV$ where both T and V are matrices with positive entries (Smaragdis and Brown 2003; Abdallah and Plumbley 2006; Virtanen 2006a; 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 minimized:

$$(T, V)^* = \min_{T,V} D(X||TV) \tag{25.11}$$

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

$$D(X||\Lambda) = \sum_{\nu,\tau} \left(x_{\nu,\tau} \log \frac{x_{\nu,\tau}}{\lambda_{\nu,\tau}} - x_{\nu,\tau} + \lambda_{\nu,\tau} \right). \tag{25.12}$$

Using Jensen's inequality (Cover and Thomas 1991) and concavity of $\log x$, it can be shown that $D(\cdot)$ is non-negative and $D(X||\Lambda) = 0$ if and only if $X = \Lambda$. The objective in (25.11) could be minimized by any suitable optimization algorithm. Lee and Seung (2000) have proposed an efficient variational bound minimization algorithm that has attractive convergence properties. that has been since successfully applied to various applications in signal analysis and source separation.

It can also be shown that the minimization algorithm is in fact an EM algorithm with data augmentation (Cemgil 2008). More precisely, it can be shown that minimizing D w.r.t. T and V is equivalent finding the ML solution of the following hierarchical model

$$x_{\nu,k} = \sum_i s_{\nu,i,k}$$

$$s_{\nu,i,k} \sim Po(s_{\nu,i,k}; 0, \lambda_{\nu,i,k}) \qquad \lambda_{\nu,i,k} = t_{\nu,i} v_{i,k} \tag{25.13}$$

$$t_{\nu,i} \sim Ga\left(t_{\nu,i}; a^t_{\nu,i}, b^t_{\nu,i}/a^t_{\nu,i}\right) \qquad v_{i,k} \sim Ga\left(v_{i,k}; a^v_{i,k}, b^v_{i,k}/a^v_{i,k}\right). \tag{25.14}$$

Note that this model is quite distinct from the Poisson point model used in Section 25.3.2 since it models each time – frequency coefficient as a Poisson random variable, while the previous approach models detected peaks in the spectrum as a spatial point process.

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

$$p(s_{v,i,k}|X, T, V) = Bi(s_{v,i,k}; x_{v,k}, \kappa_{v,i,k})$$

$$\kappa_{v,i,k} = \lambda_{v,i,k} / \sum_{i'} \lambda_{v,i',k}$$

This means that if the latent variances can be estimated somehow, source separation can be easily accomplished as $\mathsf{E}(s)_{Bi(s;x,\kappa)} = \kappa x$. It is also possible to estimate the marginal likelihood $p(X)$ by integrating out all of the templates and excitations. This can be done via Gibbs sampling or more efficiently using a variational approach that we outline in Appendix A.

25.3.7.1 *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 $a^t_{v,i} = a^t_i$ and $b^t_{v,i}$ (see equation 25.14), via maximization of the variational bound given in equation 25.20. Here, the observations are matrices X_i; a spectrogram computed given each note $i = 1 \ldots I$. In Figure 25.23, we show the estimated scale parameters $b^t_{v,i}$ as

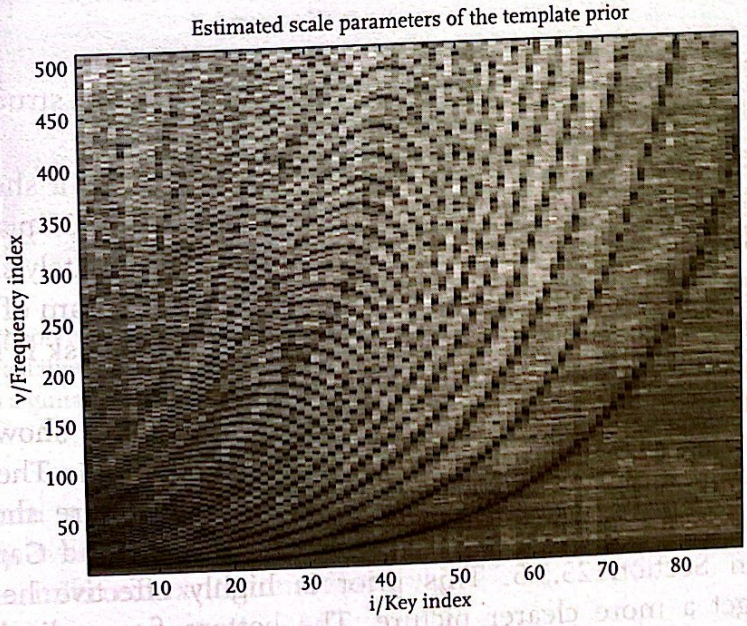

Fig. 25.23 Estimated template hyperparameters $b^t_{v,i}$.

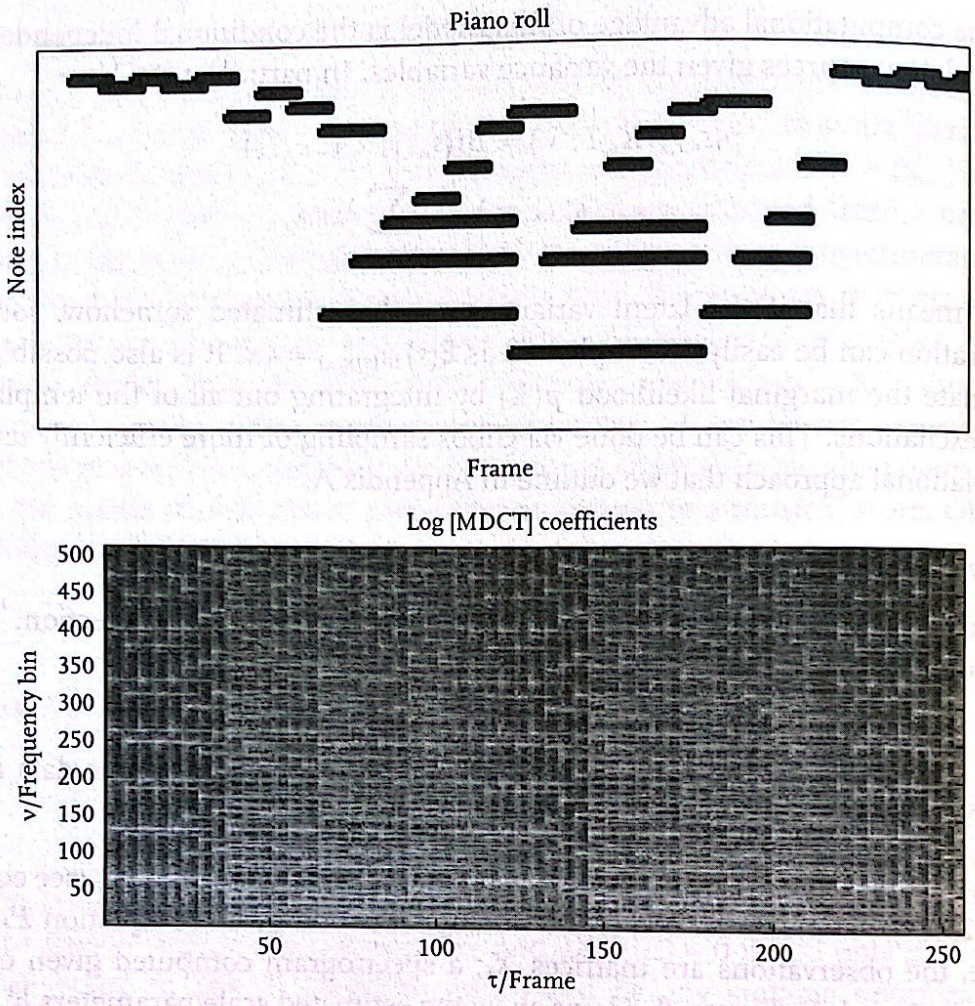

Fig. 25.24 The ground truth piano roll and the spectrum of the polyphonic data.

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 25.24. The pitch detection task is infering the excitations given the hyperparameters and the spectrogram.

The results are shown in Figure 25.25. The top figure shows the excitations estimated give the prior shown in equation 25.14. 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 25.3.5. This prior is highly effective here and we are able to get a more clearer picture. The bottom figure displays results

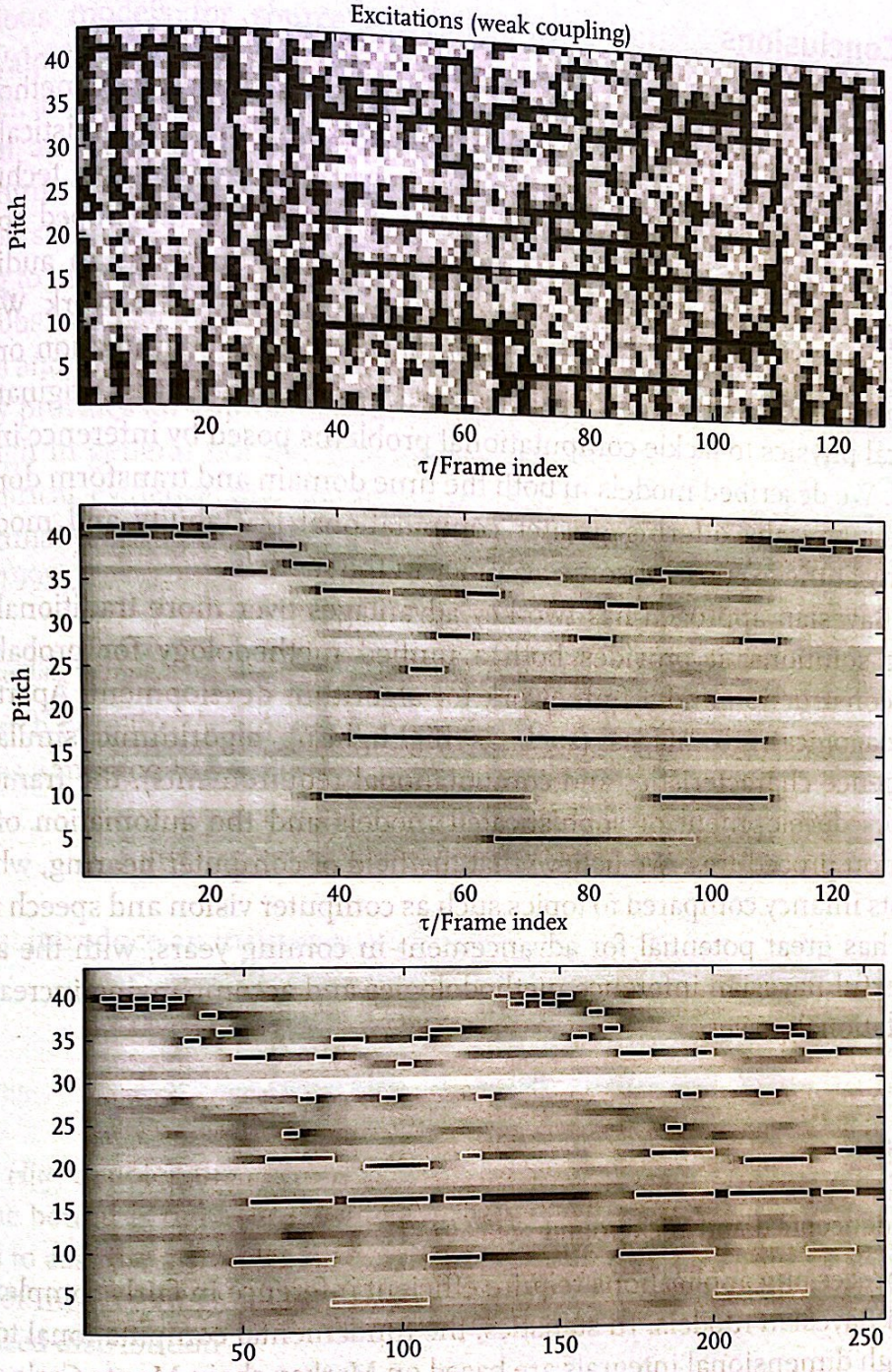

Fig. 25.25 Polyphonic pitch detection. Estimated expected excitations. (Top) uncoupled excitations. (Middle) tied excitations using a Gamma chain, ground truth shown in white. (Bottom) excitations estimated from a guitar using the hyperparameters estimated from a piano – ground truth shown in black.

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 inferred excitations show significant overlap with the original score.

25.4 Conclusions

In this chapter we have described recently proposed 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. We have reviewed and developed some of these statistical models and described how various problems in audio and music processing can be cast into the Bayesian inference framework. 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 described 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 unified methodology for probabilistic 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 the 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 increases in computational power.

Appendix

A. Broader context and background

Audio processing applications require efficient inference in fairly complex hierarchical Bayesian models. In statistics, the fundamental computational tools to such high dimensional integrals are based on Markov chain Monte Carlo strategies such as the Gibbs sampler (Gilks, Richardson, and Spiegelhalter 1996). The main advantage of MCMC is its generality, robustness and attractive theoretical properties. However, the method comes at the price of heavy computational burden which may render it impractical for data intensive applications.

An alternative approach for computing the required integrals is based on deterministic fixed point iterations (Variational Bayes – Structured Mean field; Ghahramani and Beal 2000; Wainwright and Jordan 2003; Bishop 2006). This set of methods have direct links with the well-known expectation-maximization (EM) type of algorithms. Variational methods have been extensively applied

to various models for source separation by a number of authors (Attias 1999; Valpola 2000; Girolami 2001; Miskin and Mackay 2001; Hojen-Sorensen, Winther, and Hansen 2002; Winther and Petersen 2006).

From an algorithmic point of view, the VB method can be viewed as a 'deterministic' counterpart of the Gibbs sampler. Especially for models where a Gibbs sampler is easy to construct (e.g. in models with conjugate priors leading to known full conditionals) the VB method is equally easy to apply. Like the Gibbs sampler, the framework facilitates generalization to more complex models and to automation of code generation procedure. Moreover, the method directly provides an approximation (a lower bound) to the marginal likelihood. Although in general not much is known about how tight the bound is, there is empirical evidence that for many models the bound can provide a good approximation to an estimate obtained from Gibbs sampling via Chib's method (Chib 1995).

A.1 *Bounding marginal likelihood via variational Bayes*

We sketch here the Variational Bayes (VB) (Ghahramani and Beal 2000; Bishop 2006) as a method to bound the marginal loglikelihood

$$\mathcal{L}_X(\Theta) \equiv \log p(X|\Theta) = \log \int dT dV p(X, T, V|\Theta). \tag{25.15}$$

We first introduce an instrumental distribution $q(T, V)$.

$$\mathcal{L}_X(\Theta) \geq \int dT, dV q \log \frac{p(X, T, V|\Theta)}{q} \tag{25.16}$$

$$= \mathsf{E}(\log p(X, V, T|\Theta))_q + H[q] \equiv \mathcal{B}_{VB}[q]. \tag{25.17}$$

Here, $H[q]$ denotes the entropy of q. From the general theory of EM we know that the bound is tight for the exact posterior $q(T, V) = p(T, V|X, \Theta)$. The VB idea is to assume a simpler form for the instrumental distribution by ignoring some of the couplings present in the exact posterior. A natural candidate is a factorized distribution

$$q(T, V) = q(T)q(V) \equiv \prod_{\alpha \in C} q_\alpha.$$

In the last equation, we have formally written the q distribution as a product over variables from disjoint clusters $\alpha \in C$ and $C = \{\{T\}, \{V\}\}$ denotes the set of disjoint clusters. Since in general the family of q distributions won't include the exact posterior density, 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 optimize the bound. Although the best q distribution respecting the factorization is not available in closed form, it turns out that a local optimum

can be attained by the following fixed point iteration:

$$q_a^{(n+1)} \propto \exp\left(\mathsf{E}(\log p(X, T, V|\Theta))_{q_{\neg a}^{(n)}}\right) \tag{25.18}$$

where $q_{\neg a} = q/q_a$. 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).

B. Variational Bayesian NMF

In this section we derive a variational Bayes algorithm for the NMF model described in equations (25.13) and (25.14). The marginal likelihood is given as

$$\mathcal{L}_X(\Theta) \equiv \log p(X|\Theta) \geq \sum_S \int d(T, V) q \log \frac{p(X, S, T, V|\Theta)}{q} \tag{25.19}$$

$$= \mathsf{E}(\log p(X, S, V, T|\Theta))_q + H[q] \equiv \mathcal{B}_{VB}[q] \tag{25.20}$$

where q is defined as

$$q(S, T, V) = q(S)q(T)q(V)$$

$$= \left(\prod_{\nu,\tau} q(\nu, 1:I, \tau)\right)\left(\prod_{\nu,i} q(t_{\nu,i})\right)\left(\prod_{i,\tau} q(v_{i,\tau})\right) \equiv \prod_{a\in\mathcal{C}} q_a.$$

Here, $a \in C = \{\{S\}, \{T\}, \{V\}\}$ denotes a set of disjoint clusters. A local optimum can be attained by the following fixed point iteration:

$$q_a^{(n+1)} \propto \exp\left(\mathsf{E}(\log p(X, S, T, V|\Theta))_{q_{\neg a}^{(n)}}\right) \tag{25.21}$$

where $q_{\neg a} = q/q_a$.

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

$$q(\nu, 1:I, \tau) = \mathcal{M}(\nu, 1:I, \tau; x_{\nu,\tau}, p_{\nu,1:I,\tau}) \tag{25.22}$$

$$p_{\nu,i,\tau} = \exp(\mathsf{E}(\log t_{\nu,i}) + \mathsf{E}(\log v_{i,\tau}))/\sum_i \exp(\mathsf{E}(\log t_{\nu,i}) + \mathsf{E}(\log v_{i,\tau})) \tag{25.23}$$

$$q(v_{i,\tau}) = Ga\left(v_{i,\tau}; \alpha^v_{i,\tau}, \beta^v_{i,\tau}\right) \tag{25.24}$$

$$\alpha^v_{i,\tau} = a^v_{i,\tau} + \sum_\nu m_{\nu,\tau}\mathsf{E}(\nu, i, \tau) \quad \beta^v_{i,\tau} = \left(\frac{a^v_{i,\tau}}{b^v_{i,\tau}} + \sum_\nu m_{\nu,\tau}\mathsf{E}(t_{\nu,i})\right)^{-1}. \tag{25.25}$$

The variational parameters of $q(t_{v,i}) = Ga\left(t_{v,i}; \alpha^t_{v,i}, \beta^t_{v,i}\right)$ are found similarly. The hyperparameters can be optimized by maximizing the variational bound $\mathcal{B}_{VB}[q]$. While this does not guarantee to increase the true marginal likelihood, it leads in this application to quite practical and fast algorithms and is very easy to implement (Cemgil 2008).

For the same model, it is also straightforward to implement a Gibbs sampler. A comparison showed that both algorithms give qualitatively very similar results, both for inference as well as model order selection (Cemgil 2008). We find the variational approach somewhat more practical as it can be expressed as simple matrix operations, where both the fixed point equations as well as the bound can be compactly and efficiently implemented using matrix computation software. In contrast, our Gibbs sampler is computationally more demanding and the calculation of marginal likelihood is somewhat more tricky. With our implementation of both algorithms the variational method is faster by a factor of around 13.

In terms of computational requirements, the variational procedure has several advantages. First, one circumvents sampling from multinomial variables, which is the main computational bottleneck with a Gibbs sampler in this model. Whilst efficient algorithms are developed for multinomial sampling (Davis 1993), the procedure is time consuming when the number of latent sources I is large. In contrast, the variational method computes the expected sufficient statistics via elementary matrix operations. Another advantage is hyperparameter estimation. In principle, it is possible to maximize the marginal likelihood via a Monte Carlo EM procedure (Tanner 1996; Quintana, Liu, and del Pino 1999), yet this potentially requires many more iterations of the Gibbs sampler. In contrast, the evaluation of the derivatives of the lower bound is straightforward and can be implemented without much additional computational cost.

Acknowledgements

We would like to thank Andrew Feldhaus for carefully proofreading the manuscript.

References

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

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11, 803–851.

Bertin, N., Badeau, R. and Richard, G. (2007). Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP), Honolulu.

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