BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian Harmonic Models for Musical Signal Analysis

MANUEL DAVY
IRCCyN/CNRS, France
Manuel.Davy@irccyn.ec-nantes.fr

SIMON J. GODSILL
University of Cambridge, UK
sjg@eng.cam.ac.uk

SUMMARY

This paper is concerned with the Bayesian analysis of musical signals. The ultimate aim is to use Bayesian hierarchical structures in order to infer quantities at the highest level, including such things as musical pitch, dynamics, timbre, instrument identity, etc. Analysis of real musical signals is complicated by many things, including the presence of transient sounds, noises and the complex structure of musical pitches in the frequency domain. The problem is truly Bayesian in that there is a wealth of (often subjective) prior knowledge about how musical signals are constructed, which can be exploited in order to achieve more accurate inference about the musical structure. Here we propose developments to an earlier Bayesian model which describes each component 'note' at a given time in terms of a fundamental frequency, partials ('harmonics'), and amplitude. This basic model is modified for greater realism to include non-white residuals, time-varying amplitudes and partials 'detuned' from the natural linear relationship. The unknown parameters of the new model are simulated using a variable dimension MCMC algorithm, leading to a highly sophisticated analysis tool. We discuss how the models and algorithms can be applied for feature extraction, polyphonic music transcription, source separation and restoration of musical sources.

Keywords: MUSICAL ANALYSIS; AUTOMATIC PITCH TRANSCRIPTION; PITCH ESTIMATION; INSTRUMENT CLASSIFICATION; AUDITORY SCENE ANALYSIS.

1. INTRODUCTION

Inference about the high-level information contained in musical audio signals is complex, and requires sophisticated signal processing tools (Bregman (1990)). In this paper, we focus on the automatic interpretation of musical signals. Musical audio is highly structured, both in the time domain and in the frequency domain. In the time domain, tempo specifies the range of possible note transition rates. 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 *et al.* (1983), Fletcher and Rossing (1998).) The frequencies of the partials are approximately integer multiples of the fundamental frequency. 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, 659 Hz $\approx 5/4 \times 523$ Hz and 785 Hz $\approx 3/2 \times 523$ Hz. An additional level of structure is the melody, which gives the frequency dependence of successive notes. A given melody is characterised by the succession of fundamental frequencies at specific time instants. Figure 2 shows a spectrogram analysis for a simple monophonic (single note) flute recording (this may be auditioned at `www-sigproc.eng.cam.ac.uk/~sjg/sounds/flute.wav`). In this both the temporal segmentation and the frequency domain structure are clearly visible on the plot. In polyphonic musical examples, several (possibly many!) such structures are superimposed, and the eye is then typically unable to separate the individual note structures from the spectrogram alone.

The goals of a musical analysis can be manifold, and we seek to make our models general enough that they will fit with the inference requirements at hand. Some important goals and applications, which can all be expressed in probabilistic terms through the use of suitably chosen model structures, include automatic transcription (generation of a musical 'score'), classification and search (*e.g.*, determining which instruments are playing or whether a particular tune is present) and source separation (separation of individual instruments from a polyphonic mix).

It should be clear from this discussion that musical audio provides an ideal structure for Bayesian modelling and inference: there is plenty of prior information available (both subjective and physically-based), there is a natural hierarchical structure (from individual partial frequencies through notes, chords and eventually whole melodies). All of these elements of the structure should ideally be estimated jointly in order to exploit the full power of the information available. This is a formidable task that none has successfully achieved to our knowledge. Most researchers have focused either on very high level, or very low level, modelling alone. Here we attempt partially to bridge the gap between these extremes, by exploring models which directly model musical signals in terms of their component 'notes', while retaining a moderately realistic signal level model. Of course, in future work we would wish to see the whole task performed jointly, and we will expect to see dramatic performance improvements once this is properly achieved.

At the level of musical notes, two principal tasks may be identified for the analysis of musical audio: a segmentation step that identifies note transitions in time, and an estimation step in which the number of notes as well as their fundamental frequencies, their partial structure and other characteristics are estimated at any given time. We focus on the latter, since efficient music segmentation algorithms such as the time-frequency (Laurent and Doncarli (1998)), Support Vector Machines (Davy and Godsill (2002b)) or generalised likelihood ratio (Basseville and Nikiforov (1993)) techniques can be used for this step.

Numerous musical pitch estimation and analysis techniques can be found in the literature. Most apply only to monophonic (single note) recordings and rely on non-parametric signal analysis tools (local autocorrelation function, spectrogram, etc.). We do not have space to reference all approaches here. Certain authors have, however, adopted methods with a statistical modelling flavour, often using iterative procedures to estimate the individual components of a musical signal, see for example De Cheveigne (1993), Virtanen and Klapuri (2001) and De Cheveigne and Kahawara (1999). Bayesian approaches have been surprisingly rare, considering the large quantities of prior information available about musical signals. Notable exceptions include Kashino *et al.*

(1995), Kashino and Murase (1999), who adopt a Bayesian hierarchical structure for high level features in music such as chords, notes, timbre, etc., and Sterian *et al.* (1999) who adopt Bayesian tracking ideas for modelling of time varying frequency partials. Bayesian models for polyphonic music have been proposed in Walmsley *et al.* (1998), Walmsley *et al.* (1999) and it is these that we extend and discuss in this paper.

In this paper, we devise novel Bayesian models for periodic, or nearly periodic, components in a musical signal. The work develops upon models devised for automatic pitch transcription in Walmsley *et al.* (1998), Walmsley *et al.* (1999) in which it is assumed that each musical note may be described by a fundamental frequency and linearly related partials with unknown amplitudes. The number of notes, and also the number of harmonics for each note are generally unknown and so a reversible jump MCMC procedure is adopted for inference in this variable dimension probability space; see Andrieu and Doucet (1999), Godsill and Rayner, Godsill and Rayner (1998b) and Davy *et al.* (2002) for some relevant MCMC work in signal processing and audio. Use of these powerful inference methods allows estimation of pitch, harmonic amplitudes, and the number of notes/harmonics present at each time. The methods of Walmsley *et al.* (1998) and Walmsley *et al.* (1999) have shown promise in highly complex problems with many notes simultaneously present. However, in the presence of non-stationary or ambiguous data, problems are expected in terms of large residual modelling errors and pitch errors (especially errors of +/- one octave). Here we seek to address some of these shortcomings by elaboration of the model to include more flexibility in the modelling of non-stationary data, modelling of non-white residual noise, and also to allow the modelling of inharmonicity (or 'detuning' of individual harmonics relative to the usual linear frequency spacing). As before, a variable dimension MCMC strategy is adopted for inference in the new model, and novel proposal mechanisms are developed for this purpose.

The paper is organized as follows. In Section 2, we present the basic harmonic model for the description of musical signals. Moreover, we specify the probabilistic framework, and give the parameter priors. In Section 3, we discuss estimation objectives and summarise the Bayesian computational method. Simulation results are presented in Section 4, and finally a discussion is given. Given space restrictions it has been impossible to describe in detail the prior models and exact MCMC implementation scheme used. In fact we have implemented several different versions of the models and MCMC algorithms, and the code is still undergoing development as more sophisticated and realistic modelling assumptions are incorporated. A snapshot detaling the implementation which generated the simulation results for this paper can be found as Davy and Godsill (2002a).

2. BAYESIAN MODEL FOR MUSICAL ANALYSIS

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}.

In this section we describe Bayesian models suited to musical audio analysis. We first introduce a robust monophonic (single note) model for music. We then explain

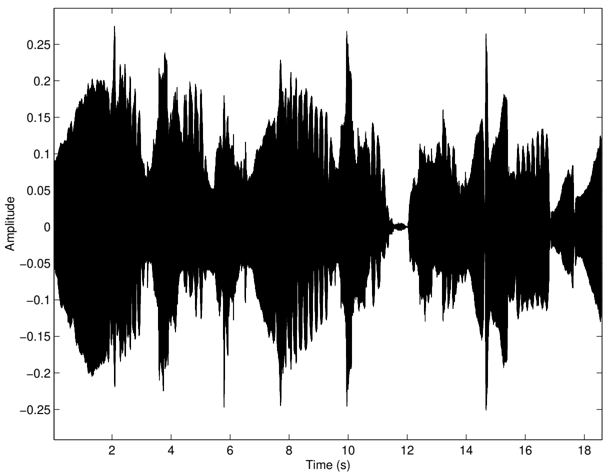

Figure 1. *Waveform: flute extract*

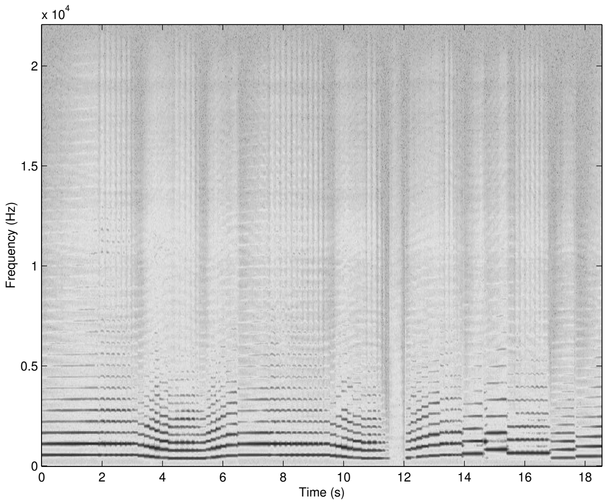

Figure 2. *Spectrogram: flute extract*

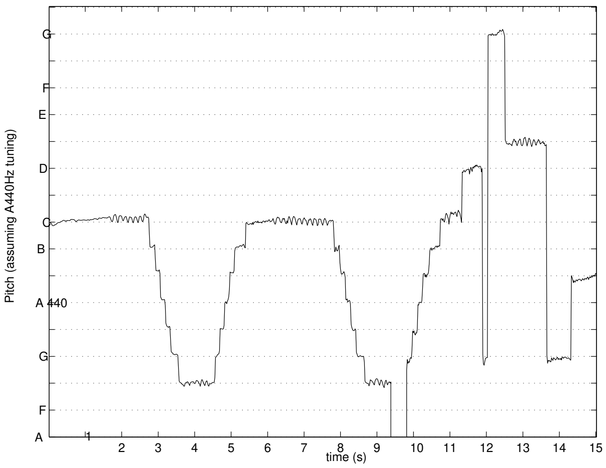

Figure 3. *Pitch estimation from flute extract*

how to expand it to a polyphonic (many note) model, and discuss the salient features of the approach. It is assumed throughout that the musical audio has been segmented such that no note transitions occur for $t \in \{0, 1, 2, ..., N-1\}$.

2.1 *Monophonic case: single note models*

In the *monophonic* case it is assumed that at any given time only one single musical pitch is sounding, *e.g.*, solo trumpet, or solo clarinet, etc. From this simple case we can build more sophisticated *polyphonic* (many note) structures by superposition of several monophonic units.

Physical considerations and empirical observation of spectrograms (see *e.g.*, Figure 2 - it is clear from this that there is a series of spectral 'lines' corresponding to the fundamental and partials of each note) lead to the conclusion that the notes in musical signals are composed of a *fundamental frequency* ω_0 and a set of M *partials*. This classic model, see *e.g.*, Serra (1997), results from the approximate short-term periodicity of musical signals, which enables a Fourier series decomposition. In the simplest cases, this idealised assumption holds well and the following model can be applied for short time segments (as in Walmsley *et al.* (1998) and Walmsley *et al.* (1999)):

$$y_t = \left\{ \sum_{m=1}^{M} \alpha_m \cos(m\omega_0 t) + \beta_m \sin(m\omega_0 t) \right\} + v_t \tag{1}$$

for $t \in \{0, \ldots, N-1\}$. Here, $M > 0$ is the number of partials present, α_m and β_m give the amplitudes of these partials, and v_t is a residual noise component. Note that $\omega_0 \in (0, \pi)$ is here scaled for convenience - its audible frequency is $\frac{\omega_0}{2\pi}\omega_s$.

It turns out that this model is over-idealised for many realistic cases and must be modified in several ways to improve performance. In particular, partials can be expected to exhibit time-varying amplitudes, and the partials can be expected to deviate from the ideal frequency spacing. These two facts can be accommodated in a new model:

$$y_t = \left\{ \sum_{m=1}^{M} \alpha_{m,t} \cos[(m+\delta_m)\omega_0 t] + \beta_{m,t} \sin[(m+\delta_m)\omega_0 t] \right\} + v_t \tag{2}$$

where the partial amplitudes $\alpha_{m,t}$ and $\beta_{m,t}$ can now depend on time, and de-tuning parameters δ_m allow each partial to be offset from its nominal frequency of $m\omega_0$.

Many evolution models are possible for the amplitude processes $\alpha_{m,t}$ and $\beta_{m,t}$, including random walks, autoregressions, etc., and most would be tractable within our Bayesian framework. It is important, however, to regularise the evolution of these components *a priori* in order that ambiguities between true frequency modelling and modelling of the time-varying amplitudes do not occur. We adopt a simple solution which consists of representing the amplitudes $\alpha_{m,t}$ and $\beta_{m,t}$ in terms of smooth basis functions ϕ_i, $i = 1, \ldots, I$ (with I fixed and known) such that

$$\alpha_{m,t} = \sum_{i=1}^{I} a_{m,i}\phi_{i,t} \; ; \quad \beta_{m,t} = \sum_{i=1}^{I} b_{m,i}\phi_{i,t} \tag{3}$$

There are many possible choices for the basis functions, and in practice any sufficiently smooth interpolation functions will do. The number of basis functions $(I+1)$ has to be upper bounded to avoid unidentifiability: low frequencies can actually be modelled by a time-varying amplitude as well as by a sinusoid. It is thus important to limit the number of basis functions such that $\omega_{amp.max} = (I+1)\omega_s/N$ is below the lowest note frequency in a given harmonic signal (*e.g.*, , 20 Hz). Here we have implemented a simple scheme involving raised cosine functions (Hanning windows) with 50% overlap, see Figure 4. Since I is typically chosen much smaller than N, the reparameterisation in terms of basis coefficients $a_{m,i}$ and $a_{m,i}$ is of much lower dimensionality than the original formulation in terms of $\alpha_{m,t}$ and $\beta_{m,t}$. The monophonic note model now becomes:

$$y_t = \left\{ \sum_{m=1}^{M} \sum_{i=1}^{I} a_{m,i}\phi_{i,t} \cos\left[(m+\delta_m)\omega_0 t\right] + b_{m,i}\phi_{i,t} \sin\left[(m+\delta_m)\omega_0 t\right] \right\} + v_t \qquad (4)$$

We note that the model can now be seen as a representation of y_t in terms of a set of *Gabor atoms* (Flandrin (1999)). Here, each atom has a precise time-frequency location $(t_i, [m+\delta_m]\omega_0)$ and an amplitude $(a_{m,t}^2 + b_{m,t}^2)^{1/2}$ where t_i is the temporal centre of ϕ_i. We will refer to each term $\phi_{i,t}\cos[\omega t]$ or $\phi_{i,t}\sin[\omega t]$ as an 'atom' in the subsequent material.

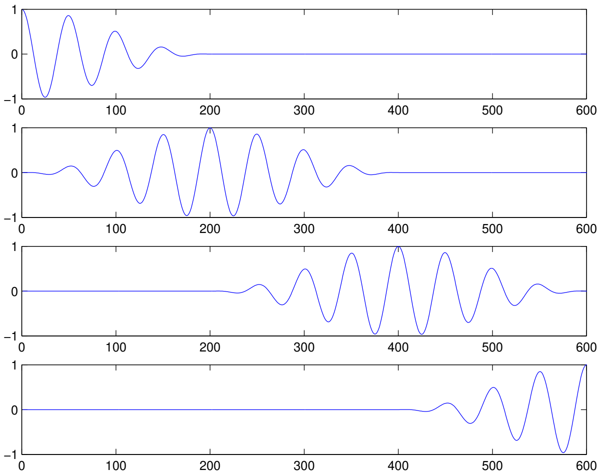

Figure 4. *Typical set of Gabor atoms* $\phi_{i,t}\cos(.)$

2.2 *A polyphonic harmonic model*

The above monophonic model can be easily expanded to the polyphonic case, that is for signals composed of K concurrent notes. A suitable model is:

$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=1}^{I} a_{k,m,i}\phi_{i,t} \cos\left[(m+\delta_{k,m})\omega_{0,k} t\right] + \ b_{k,m,i}\phi_{i,t} \sin\left[(m+\delta_{k,m})\omega_{0,k} t\right] \right\} + v_t \qquad (5)$$

for $t = 0, \ldots, N-1$. Each note, $k = 1, \ldots, K$, now has its own set of parameters, and the notation is extended in an obvious way with an additional subscript 'k'. The (unknown) parameters determining the polyphonic model of Equation (5) are: the total number of notes K, the number of partials for each note $\boldsymbol{M} = [M_1, \ldots, M_K]^T$, the de-tuning parameters $\boldsymbol{\delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K]$ with $\boldsymbol{\delta}_k = [\delta_{k,1}, \ldots, \delta_{k,M_k}]^T$, the fundamental frequencies $\boldsymbol{\omega}_0 = [\omega_{0,1}, \ldots, \omega_{0,K}]^T$ and the amplitudes $\boldsymbol{\theta} = [a_{1,1,1}, b_{1,1,1}, \ldots, a_{K,M_K,I}, b_{K,M_K,I}]^T$. It is assumed that I is prespecified. In vector notation the model is now written as:

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{v} \qquad (6)$$

where $\boldsymbol{y} = [y_0, \ldots, y_{N-1}]^T$, $\boldsymbol{v} = [v_0, \ldots, v_{N-1}]^T$, the matrix $\boldsymbol{D}$ contains the Gabor atoms stacked in columns, and $\boldsymbol{\theta}$ contains the amplitude parameters $a_{k,m,i}$ and $b_{k,m,i}$. The detailed expressions for $\boldsymbol{D}$ and $\boldsymbol{\theta}$ are given in Davy and Godsill (2002a).

In practice, musical signals also have non-harmonic components, such as emitted air sounds or aspiration noise. These components, in addition to any background noise, are subsumed in the noise term v_t in Equation (5). It is desirable that the noise v_t models accurately all the possible sources of model errors. A simple and general possibility is the *autoregressive* (AR) model of order p:

$$v_t \;=\; \gamma_1 v_{t-1} + \ldots + \gamma_p v_{t-p} + \epsilon_t \tag{7}$$

where ϵ_t is a zero mean Gaussian white noise of variance σ_ϵ^2. This introduces an additional set of parameters σ_ϵ^2, p and $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^T$.

Given the linear model formulation of Equation (5) and the assumption of i.i.d. Gaussian excitation for the AR process, we immediately obtain the likelihood function:

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \boldsymbol{\gamma}, \sigma_\epsilon^2) = \frac{1}{(2\pi\sigma_\epsilon^2)^{(N-p)/2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\theta})^T \boldsymbol{A}^T \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\theta})\right] \tag{8}$$

The $(N-p) \times N$-dimensional matrix $\boldsymbol{A}$ is constructed by stacking the AR coefficients in rows, with appropriate zero padding, as detailed in Davy and Godsill (2002a).

2.3 *Features of the Model*

The models proposed in earlier subsections, both monophonic and polyphonic, all fall into the general category of the linear model. Under a Gaussian prior for $\boldsymbol{\theta}$ this will facilitate inference in the model by allowing exact simulation of $\boldsymbol{\theta}$ from its full conditional, and marginalisation of $\boldsymbol{\theta}$ from the posterior distribution. This can provide an important dimensionality reduction in the model since $\boldsymbol{\theta}$ is often high-dimensional. This, however, is all standard material, and we see the main interest of our model to be in the specific structure chosen for the $\boldsymbol{D}$-matrix and in the prior distributions of the unknown parameters, both of which are carefully tuned to subjective and objective information about musical signals.

The polyphonic model presented as Equation (5) has several features that distinguish it from other work in the area. Firstly, it directly incorporates the frequency relationship between partials and fundamental frequencies. This is different from typical approaches in the literature to musical pitch transcription which will estimate the frequencies of each line spectral component independently of the others, performing grouping into units such as chords and notes as a post-processing stage (but note that papers such as Gribonval and Bacry (2001) and Klapuri (1999) go some way towards integrating the harmonic structure directly into the model). This misses an opportunity for greater estimation accuracy through direct modelling of the waveform at the level of notes. We retain this feature in our model, based on the earlier model of Equation (1) (Walmsley *et al.* (1998), Walmsley *et al.* (1999)). An extension provided in the model of Equation (5) is the incorporation of detuned harmonics with parameters $\delta_{k,m}$. Many researchers believe this to be an important component of any realistic musical instrument model Fletcher and Rossing (1998) and we plan to investigate this claim through use of the new model. Potential ambiguities can occur if the $\delta_{k,m}$ parameters allow one partial (harmonic) to stray into the frequency range of adjacent harmonics. However, this ambiguity is suppressed here by careful choice of priors that favour $|\,\delta_{k,m}\,| << 1$. Note that this model, which essentially includes random deviations from the natural harmonic positions, could easily be extended to model the more systematic trends in spacing of harmonics observed in *e.g.*, string instruments Fletcher and Rossing (1998).

Secondly, instruments produce notes with time varying amplitude. This time variation is present throughout an individual note, but particularly evident around the start-up region, or *attack*. A constant amplitude model such as Equation (1) is clearly unable to deal with such a case and will lead to misleading parameter inferences. The chosen decomposition of amplitudes in terms of a set of basis functions, Equation (3), ensures smoothness, and reduces the number of parameters in the model considerably.

Finally, the residual noise is an AR process that can model residual noise from instruments as well as general background noise.

Further improvements in modelling could be achieved by allowing the fundamental frequencies ω_0 to vary over the time-frame - however, this leads to a much more intractable model that we have avoided implementing thus far - provided frame sizes are kept short, the frequencies can usually be modelled as constant within a frame. See Walmsley *et al.* (1999) for some progress on models with time-varying frequency.

Even without time-varying frequencies the price of this more flexible model is a large number of unknown parameters. The principal unknowns are, then, the number of notes K, the fundamental frequency, number of harmonics and amplitudes for each note: $\{\omega_{0,k}, M_k, a_{k,m,i}, b_{k,m,i}\}$, the AR parameters $\boldsymbol{\gamma}$ and the variance σ_ϵ^2. This variable-dimension space of parameters is embedded in a Bayesian scheme as outlined in the next section. To our knowledge this problem has never received a fully Bayesian treatment before. As will be seen, the Bayesian priors, in addition to the special structure of $\boldsymbol{D}$, play a key role in defining the model.

2.4 *Bayesian model*

A natural hierarchical prior structure for the musical model is as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \boldsymbol{\gamma}, \chi^2, \sigma_\epsilon^2) =$$

$$p(\boldsymbol{\theta} \mid \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \sigma_\epsilon^2) p(\boldsymbol{\delta} \mid \boldsymbol{\omega}_0, \boldsymbol{M}, K)\, p(\boldsymbol{\omega}_0 \mid \boldsymbol{M}, K) p(\boldsymbol{M} \mid K)\, p(K)\, p(\boldsymbol{\gamma})\, p(\chi^2)\, p(\sigma_\epsilon^2)$$

where χ is introduced later. The form of the prior distributions can be chosen to reflect prior beliefs about particular types of music, or particular instruments, and this is certainly an interesting line of future study. Here we adopt a generic approach in which the priors are designed to match the average character of musical notes. We consider the prior distributions one by one.

Prior for $\boldsymbol{\theta}$*.* The amplitudes of the partials determine the characteristic 'timbre' of a musical note. Hence it is important to model these accurately in applications such as source separation and musical instrument classification. We adopt a zero-mean Gaussian prior for these parameters. This matches well the variability observed when the same note is played under slightly different conditions or on different instruments. It is also reasonable to assume the scale of the amplitudes is related to the scale of the AR residual process, since the AR residual models principally non-harmonic noises produced by the instruments. Thus we adopt a zero-mean Gaussian prior with covariance matrix $\sigma_\epsilon^2 \Sigma_\theta$. Choice of the matrix Σ_θ will then determine the properties of the prior. We have implemented a number of possibilities here, and it is clear that the prior would ideally be instrument-specific. However, it is possible to build in a certain amount of physical prior knowledge without limiting to very narrow classes of instrument. See for example Figure 5. This displays a single short-time Fourier magnitude spectrum for the flute extract in Figure 1. The spectrum is computed from approximately the first 0.25s of the music where no note transitions occur. The fundamental frequency and partials are clearly visible, exhibiting a slow decay in amplitude with decreasing frequency. This general observation applies to most acoustical sounds and so can be usefully incorporated in a prior. One successful implementation sets Σ_θ as diagonal with diagonal elements equal to χ^2/m^α, where m is the number of the partial, α is experimentally determined ($\alpha = 2$ is a good match to many signals we have analysed), and χ^2 is an unknown scale parameter that is sampled in the MCMC scheme with an inverted gamma prior. This form of covariance matrix assumes joint prior independence of all notes, partials and atoms. While this functions well in practice, it may be argued that dependence should be modelled between partials and atoms within a particular note. There are many ways this could be achieved and we leave this as a future topic of research. As an alternative to these physically based priors, we have also implemented with some success the well known G-prior which has been found to be effective in similar contexts (Andrieu and Doucet (1999), Walmsley *et al.* (1999)). A full investigation of the relative merits of these choices is again left as a topic of future work.

Prior for $\boldsymbol{M}$ *and* K*.* The number of partials is an instrument- and realisation-specific quantity. The precise distribution can be learned from training examples with different instruments. The general feature is that a particular instrument has a mean number of partials with a certain spread about this value. In order to model this, we have adopted an independent Poisson prior for each M_k, truncated to user-specified maximum and minimum limits. Similarly, the number of notes, K, has a truncated Poisson prior. These are specified vaguely for general musical extracts, but can be tuned more precisely when a particular instrument is known to be present.

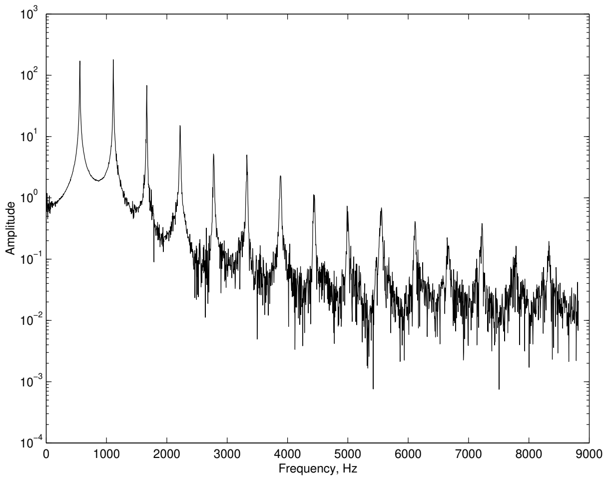

Figure 5. *Flute extract*

Prior for $\boldsymbol{\delta}$*.* A third key parameter is the vector of detuning factors, $\boldsymbol{\delta}$, which is aimed at modelling slight harmonic de-tuning among the partials. Its value is expected to be close to zero, and here we assume no dependence between the δ parameters of different partials, although for certain instruments theory would dictate the general form of the δs (Fletcher and Rossing (1998)). Here a zero mean independent Gaussian is assumed, with variance σ_δ^2, fixed to a small value in order to favor small de-tuning parameters. The distribution can additionally be truncated in order that adjacent partials within a single note do not cross over one another in the frequency domain.

Prior for $\boldsymbol{\omega}_0$. For some instruments such as piano or organ, notes are tuned to a fixed grid of frequencies (one for each key of the instrument). In other instruments, the player will usually tune notes to be close to the fixed grid of note frequencies, see *e.g.*, the pitch transcription of the flute extract in Figure 3. A convenient and informative prior would thus favour these fixed frequencies above others. Moreover, when several notes are played at the same time (a chord), there exist simple relations between the fundamental frequencies. The prior density $p(\boldsymbol{\omega}_0 \,|\, \boldsymbol{M}, K)$ should reflect this prior information. However, these considerations only apply to tonal Western music, and for the sake of generality we here adopt a uniform prior over some region of interest, $\boldsymbol{\omega}_0 \in [\omega_{min}, \omega_{max}]^M$. Again, there are plenty of interesting possibilities for informative priors in future investigations.

Remaining parameters. Briefly, regarding the noise v_t, $p(\sigma_\epsilon^2)$ is inverted gamma with parameters α_ϵ and β_ϵ, $p(\boldsymbol{\gamma})$ is zero-mean Gaussian, with diagonal covariance matrix $\boldsymbol{I}_p$. p is fixed in our implementations thus far, but we note that standard samplers are readily available should this parameter become important (Godsill 2001, Troughton and Godsill (1998), Troughton and Godsill (2001), Vermaak *et al.* (1999)).

Posterior distribution. Given the above prior structure, the posterior distribution may be computed. In particular, $\boldsymbol{\theta}$ and σ_ϵ are marginalised using standard linear model computations (Bernardo and Smith (1994), West and Harrison (1997)) to give a reduced posterior $p(\boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \boldsymbol{\gamma}, \sigma_\epsilon^2, \chi \,|\, \boldsymbol{y})$. Full conditionals are also readily available for $\boldsymbol{\theta}$, σ_ϵ, $\boldsymbol{\gamma}$ and χ, owing to the conjugate prior structures chosen. These distributions are all employed in the variable dimension MCMC algorithms for computation, as summarised in the next section.

3. BAYESIAN COMPUTATIONS

The precise objectives of inference in these models is very much application driven. In the case of pitch estimation, for example, point estimates will typically be required for the fundamental frequencies $\boldsymbol{\omega}_0$. The posterior distribution of these parameters will also give a useful estimate of the uncertainty and multimodality in the posterior. In source separation and instrument classification tasks, estimates and posterior distributions of the partial amplitiudes, the detuning parameters, fundamental frequency and number of harmonics are required for each note. There are however potential pitfalls in making these estimates, as the individual note ordering is not constrained in our model and it is quite conceivable that two notes are swapped during the MCMC simulation. An ordering by, say, increasing fundamental frequency or energy would help somewhat, although note swapping could still occur when large jumps in parameter values are made (*e.g.*, a change in fundamental frequency by an octave or more). We have avoided these ambiguities in this work by estimating functionals that do not depend upon the note labelling. These are computed as Monte Carlo approximations to posterior means:

$$I(f) \;=\; \int_\Omega f(\Phi)\, p(d\Phi \,|\, \boldsymbol{y}) \;\approx\; \frac{1}{L}\sum_{l=1}^{L} f(\widetilde{\Phi}^{(l)}) \tag{9}$$

where $\Phi = \{\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \boldsymbol{\gamma}, \chi^2, \sigma_\epsilon^2\}$ is the collection of all unknowns in the model, $f(.)$ is a given integrable function with respect to the posterior and Ω is the sample space for the posterior distribution. (The integral $I(f)$ over the discrete parameters should be seen as a discrete sum over all the possible values of these discrete parameters.) $\widetilde{\Phi}^{(l)}$ are (possibly dependent) Monte Carlo samples drawn from $p(\Phi \mid \boldsymbol{y})$.

A suitable family of functions to estimate are spectrogram-like representations. The short-time energy spectrum of the ith component in the model Equation (5) is defined as

$$g_{i,\boldsymbol{\theta},\boldsymbol{\omega}_0,\boldsymbol{\delta},K}(\omega) \;=\; \left| \sum_{k=1}^{K} \sum_{m=1}^{\boldsymbol{M}_k} a_{k,m,i} \Phi_{i,(m+\delta_{k,m})\omega_{0,k}}(\omega) + b_{k,m,i} \Phi_{i,(m+\delta_{k,m})\omega_{0,k}-\frac{\pi}{2}}(\omega) \right|^2 \tag{10}$$

where $\Phi_{i,\omega'}(\omega)$ is the Fourier transform of $\phi_{i,t}\cos(\omega' t)$

The construction of the spectrogram-like representation consists in computing for $i = 1, \ldots, I$ and various values of ω in $[0, \pi]$

$$\widehat{g}_{i,\boldsymbol{\theta},\boldsymbol{\omega}_0,\boldsymbol{\delta},K}(\omega) \;=\; \frac{1}{L} \sum_{l=1}^{L} g_{i,\widetilde{\boldsymbol{\theta}}^{(l)},\widetilde{\boldsymbol{\omega}}_0^{(l)},\widetilde{\boldsymbol{\delta}}^{(l)},\widetilde{K}^{(l)}}(\omega) \tag{11}$$

As stated above, these computations require that a set of samples $\{\widetilde{\boldsymbol{\theta}}^{(l)}, \widetilde{\boldsymbol{\omega}}_0^{(l)}, \widetilde{\boldsymbol{\delta}}^{(l)}, \widetilde{\boldsymbol{M}}^{(l)}, \widetilde{K}^{(l)}, \widetilde{\boldsymbol{\gamma}}^{(l)}, \widetilde{\sigma}_\epsilon^{2\,(l)}\}$, $l = 1, \ldots, L$ is available from the posterior $p(\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{M}, K, \boldsymbol{\gamma}, \sigma_\epsilon^2 \mid \boldsymbol{y})$. The reversible jump MCMC algorithm for achieving this is summarised in the following paragraphs.

The simulation algorithm is a variable dimension MCMC procedure, using reversible jumps to achieve model space moves for both the number of harmonics in each note and the number of notes. Other parameters are updated using Metropolis-within Gibbs sampling moves for non-standard conditionals and Gibbs sampling where the conditionals are standard. The posterior distribution in this problem is highly multimodal and strongly peaked. This is partly as a result of ambiguities inherent in the model, and implies that the MCMC algorithms have to be carefully constructed in order to avoid getting stuck in local traps of the distribution. In fact much of the innovative work in this project has been concerned with generation of effective proposal distributions for fast exploration of the parameter space. As well as standard random walk moves, these include independence proposals based upon the sample autocorrelation function and spectrum of the data, and octave/fifth-jumping moves aimed at moving rapidly between local maxima (these moves update the number of partials as well, preserving the spectral structure, and this leads to improved acceptance rates). The reversible jump proposals allow for several partials to be added/removed at once from a note, and notes may be split and merged in meaningful ways. Full details can be found in Davy and Godsill (2002a).

4. SIMULATION RESULTS

Results are presented based on two implementations. The first is a full implementation of the models as described in the paper with detuned partials and unknown numbers of partials M_k and number of notes K. The full details of this sampler can be found in Davy and Godsill (2002a). This first sampler is used to demonstrate the effectiveness of the model in analysising short data sets containing isolated notes and chords. The second sampler is a reduced version of the first, in which the partials are not detuned (i.e. $\boldsymbol{\delta}$ is fixed to zero), and the number of notes K is specified *a priori* (but M_k is sampled using a reversible jump procedure). As a result, though less robust and general, the second sampler takes far less computational time, both per iteration and in terms of number of iterations to convergence. It is thus used for a rapid frame-based analysis of long data sets containing many different pitches and note transitions. In this way we hope to demonstrate both the potential of the full model and also the possibilities of realistic analysis for long monophonic and polyphonic musical extracts.

4.1 *Full sampler*

The first example is a 2-note mixture of a saxophone and trumpet, playing with fundamental frequencies 349 Hz (F) and 523 Hz (C), assuming A440 Hz tuning. This short extract is taken from example 'Commit' at 2.5s, see reduced sampler section. Note that the number of notes, as well as the number of partials for each notes were unspecified for the MCMC simulations. After fitting the model with MCMC to the data, the fitting error is almost perfect when viewed in the time domain. Figure 6 displays the spectra of the two estimated notes as well as the spectrum of the error signal. As can be seen, both the number of notes and the number of partials were accurately estimated (note however that the 10th partial of the note F was missed, but this had no major consequence for the inference). In addition, the fundamental frequencies are correctly estimated.

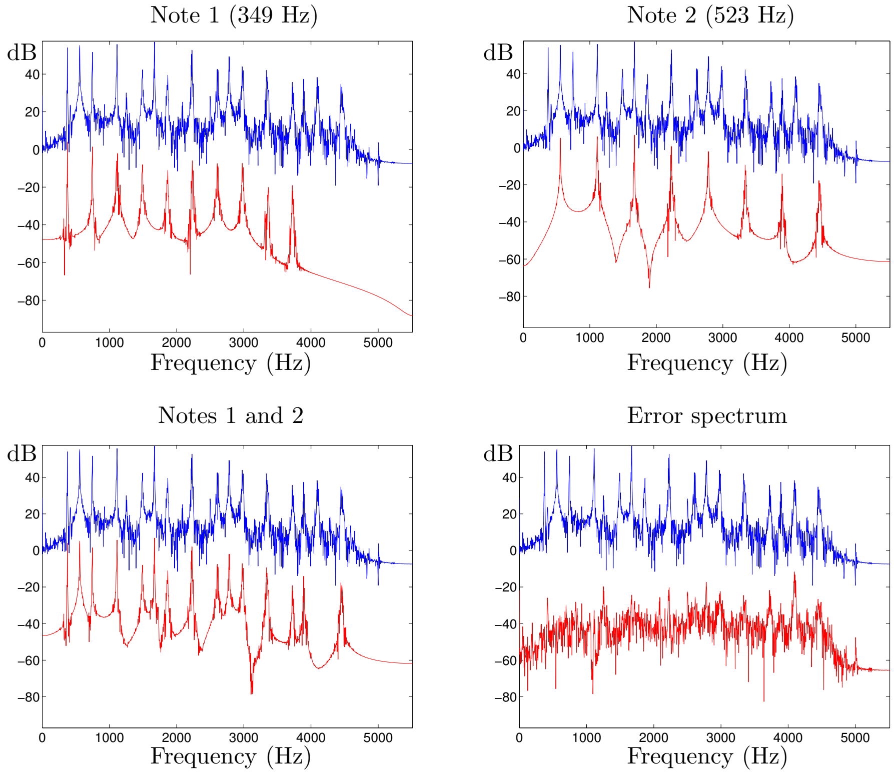

Figure 6. *Comparison of the original spectrum with MCMC-estimated spectra. In order to improve clarity, the original spectrum (top graph in each panel) has been translated by adding 50dB to its amplitude in the four panels. (Top row) Estimated spectra of the individual notes. (Bottom Left) Spectrum of notes 1 and 2 superimposed. (Bottom Right) Spectrum of the residual error.Flute extract*

We have also plotted the spectrogram-like representation of the estimated notes, see Figure 7. As can be seen, the notes are very concentrated around the 'true frequencies'

349 Hz, and 523 Hz, which shows that the posterior distribution is well concentrated around the true frequency values. Moreover, the spectrogram of the original data has also been computed (for equal comparison, we used $\phi_{i,t}$ as windowing function with 50% overlap). The frequencies and amplitudes of the line components in the two representations are very similar, which demonstrates again the accuracy of the approach. An audio animation of the MCMC procedure during convergence can be listened to at `http://www-sigproc.eng.cam.ac.uk/ ~md283/harmonic.html`.

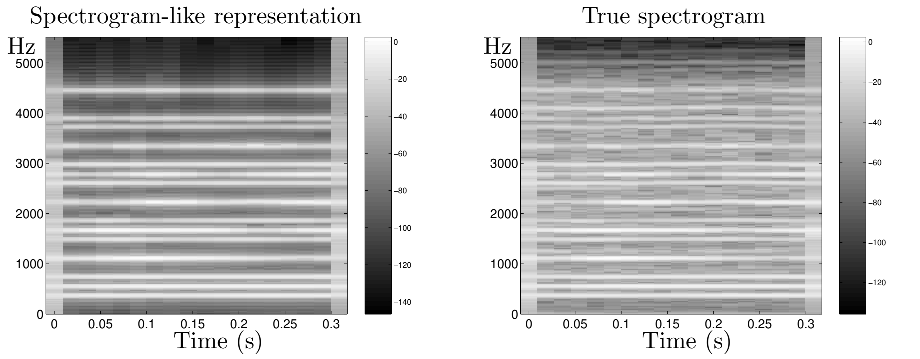

Figure 7. *(Left) Spectrogram-like representation inferred from the MCMC samples using Equation (11). (Right) Spectrogram of the original time series, computed using the windows $\phi_{i,t}$.*

We have also plotted the spectrogram-like representation of the estimated notes, see Figure 7. As can be seen, the notes are very concentrated around the 'true frequencies' 349 Hz, and 523 Hz, which shows that the posterior distribution is well concentrated around the true frequency values. Moreover, the spectrogram of the original data has also been computed (for equal comparison, we used $\phi_{i,t}$ as windowing function with 50% overlap). The frequencies and amplitudes of the line components in the two representations are very similar, which demonstrates again the accuracy of the approach. An audio animation of the MCMC procedure during convergence can be listened to at `http://www-sigproc.eng.cam.ac.uk/ ~md283/harmonic.html`.

4.2 *Reduced sampler*

The reduced sampler is applied frame-wise to the data, as in the restoration processing of Godsill and Rayner (1998). The waveform was arbitrarily segmented into blocks of duration 0.1s with 50% overlap and the reduced MCMC sampler applied in turn to each block. First a short solo flute extract considered here is the opening of Debussy's Syrinx, downsampled to a 22050 Hz sampling rate. This is monophonic throughout and hence processed with $K = 1$ throughout. The pitch estimates obtained are shown in Figure 3, corresponding to the waveform and spectrogram in the figures above it. Estimated pitches are plotted logarithmically with grid lines showing semitone steps relative to A440Hz. The estimated pitch corresponds exactly to a manual transcription of the recording with the exception of the brief low G around 12s. Close listening around 12s shows that the flute plays a low distortion undertone in addition to the scored pitch at this point, and the algorithm is clearly modelling this undertone. The 'drop-out' between 9s and 10s corresponds to a short period of silence. Informal examination of spectrograms indicates that the reversible jump algorithm for determining the number of harmonics is very successful. This demonstrates the high reliability and accuracy of the models for monophonic pitch estimation. Next the example 'Commit' (see full sampler above) is processed in its entirity. There are two notes playing throughout, so $K = 2$ is used. Figure 8 displays the pitch estimation results for the extract. Note that the number of notes is known, and was provided for inference. Comparison with ground truth shows the pitch estimation accuracy in the presence of polyphonic music.

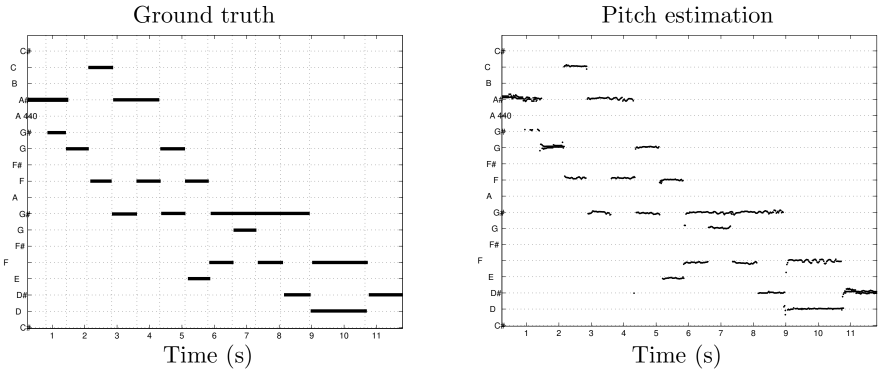

Figure 8. *Ground truth (manual transcription) and pitch extraction for Commit extract, assuming A440 Hz tuning.*

5. DISCUSSION

We have presented rather limited results here. The methods have in fact been tested on a range of real audio material and found to be robust provided that there are no more than 3 notes playing simultaneously, in which case ambiguities can cause errors in the fundamental frequency estimation. This result is similar to those reported by other authors using other techniques, such as Virtanen and Klapuri (2001) and Kashino *et al.* (1995). However, these other methods integrate contextual information or more specific instrument-based knowledge into the processing, while we have specified prior distributions at a very generic level and have not integrated temporal information from surrounding data or the relationships that exist between notes at a particular time. These aspects can all be encoded within a Bayesian framework and we anticipate that future incorporation of ideas such as these into our problem will lead to significant enhancements in performance. Computation remains a concern, however, as the distributions involved are highly multimodal and intractable to more efficient analysis.

ACKNOWLEDGEMENTS

The work of both authors was partially supported by the EU project Models for Multimedia Information Retrieval. We would also like to thank Steve Hainsworth, Prof. Peter Rayner, Patrick Wolfe and Dr Kunio Kashino for helpful discussions relating to this paper.

REFERENCES

Andrieu, C. and Doucet, A. (1999). Joint Bayesian Detection and Estimation of Noisy Sinusoids via Reversible Jump MCMC. *IEEE Trans. Signal Processing* **47**, 2667–2676.

Basseville, M. and Nikiforov, I. (1993). *Detection of abrupt changes: theory and application.* Englewood Cliffs, NJ: Prentice-Hall.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* New York: Wiley

Bregman, A. *Auditory scene analysis.* Cambridge, MA: The MIT Press.

Davy, M., Doncarli, C. and Tourneret, J. Y. (2002). Classification of chirp signals using hierarchical Bayesian learning and MCMC methods. *IEEE Trans. Signal Processing* **50**, 377–388.

Davy, M. and Godsill, S. (2002a). Bayesian harmonic for musical pitch estimation and analysis. *Tech. Rep.*, CUED/F-INFENG/TR431 Engineering Department, University of Cambridge, UK.

Davy, M. and Godsill, S. (2002b). Detection of abrupt spectral changes using support vector machines. *Proc. IEEE ICASSP*.

De Cheveigne, A. (1993). Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model for auditory processing. *J. Acoustical Society of America* **93**, 3271–3290.

De Cheveigne, A. and Kawahara, H. (1999). Multiple period estimation and pitch perception model. *Speech Communication* **27**, 175–185.

Flandrin, P. (1999). *Time-Frequency/Time-Scale Analysis*. New York: Academic Press

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

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comp. Graph. Statist.* **10**, 230–248.

Godsill, S. J. and Rayner, P. J. W. (1998a). *Digital Audio Restoration: A Statistical Model-Based Approach*. Berlin: Springer

Godsill, S. J. and Rayner, P. J. W. (1998b). Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. *IEEE Trans. on Speech and Audio Processing* **6**, 352–372.

Gribonval, R. and Bacry, E. (2001). Harmonic decomposition of audio signals with matching pursuit. *Tech. Rep.*, IRISA-INRIA.

Kashino, K.and Murase, H. (1999). A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication* **27**, 337–349.

Kashino, K., Nakadai, K., Kinoshita, T. and Tanaka, H. (1995). Organisation of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism. *Proc. 14th Int. joint conf. on artificial intelligence*.

Laurent, H. and Doncarli, C. (1998). Stationarity index for abrupt changes detection in the time-frequency plane. *IEEE Signal Processing Letters* **5**, 43–45.

McIntyre, M., Schumacher, R. and Woodhouse, J. (1993). On the oscillations of musical instruments. *J. Acoustical Society of America* **74**, 1325–1345.

Serra, X. (1997). Musical Sound Modeling With Sinusoids plus Noise. In *Musical Signal Processing*, C. Roads and S.T. Pope and Aldo Piccialli and G. De Poli (eds.), Swets and Zeitlinger.

A. Sterian, M.H. Simoni and G.H. Wakefield (1999), Model-based Musical Transcription, Proc. International Computer Music Conference.

Troughton, P. and Godsill, S. J. (1998). A reversible jump sampler for autoregressive time series. *Proc. IEEE ICASSP*.

Troughton, P. and Godsill, S. J. (2001). MCMC methods for restoration of nonlinearly distorted autoregressive signals. *Signal Processing* **81**, 83–97.

Vermaak, J., Andrieu, C., Doucet, A. and Godsill, S. J. (1999). Bayesian model selection of autoregressive processes. *J. Time Series Analysis* .

Virtanen, T. and Klapuri, A. (2001). Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. *Proc. IEEE Workshop on Audio and Acoustics 2001*.

Walmsley, P., Godsill, S. J. and Rayner, P. J. W. (1998). Multidimensional optimisation of harmonic signals. *Proc. EUSIPCO*.

Walmsley, P., Godsill, S. J. and Rayner, P. J. W. (1999).Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. *Proc. IEEE Workshop on Audio and Acoustics 2001*.

West, M. and Harrison, J. *Bayesian Forecasting and Dynamic Models*. New York: Springer