BAYESIAN HARMONIC MODELS FOR MUSICAL PITCH ESTIMATION AND ANALYSIS

Simon Godsill

Cambridge University
Engineering Department
Trumpington Street, Cambridge CB2 1PZ - UK
sjg@eng.cam.ac.uk

Manuel Davy

IRCCyN - UMR CNRS 6597
1, rue de la Noë – BP92101
44321 Nantes Cedex 3 - FRANCE
Manuel.Davy@irccyn.ec-nantes.fr

ABSTRACT

Estimating the pitch of musical signals is complicated by the presence of partials in addition to the fundamental frequency. In this paper, we propose developments to an earlier Bayesian model which describes each component signal in terms of fundamental frequency, partials ('harmonics'), and amplitude. This basic model is modified for greater realism to include non-white residual spectrum, time-varying amplitudes and partials 'detuned' from the natural linear relationship. The unknown parameters of the new model are simulated using a reversible jump MCMC algorithm, leading to a highly accurate pitch estimator. The models and algorithms can be applied for feature extraction, polyphonic music transcription, source separation and restoration of musical sources.

1. INTRODUCTION

Accessing the high level information contained in general audio signals (i.e., music, environmental noise, speech or a mixture of these) is complex, and requires sophisticated signal processing tools. Many previous studies have emphasized the major interest in audio descriptors, or *audio features*, that summarize the spectral information contained in an audio signal at a given time. Among these features, the *pitch*, which is closely related to the *fundamental frequency*, is of prime importance for applications involving music. Numerous musical pitch estimation techniques can be found in the literature [1–3], and most rely on nonparametric signal analysis tools (local autocorrelation function,spectrogram, etc.). However, characterizing a musical signal with a single pitch value at time t is not sufficient for applications such as music transcription. More complex approaches using banks of filters have been proposed in order to estimate all the frequencies, but their accuracy is not sufficient in complex cases.

In this paper, we devise novel Bayesian models for periodic, or nearly perdiodic, components in a musical signal. The work develops upon models devised for automatic pitch transcription by Walmsley *et al.* [4, 5] 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 [6–10] 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 [4, 5] 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, and also to allow the modelling of inharmonicity (or 'detuning' of individual harmonics relative to the usual linear frequency spacing). Specifically, the modelling contributions of this paper are:

- A continuously variable amplitude process over time for each harmonic
- Modelling of non-white residual error
- More realistic prior modelling of harmonic amplitudes
- Modelling of inharmonicity in partials

As before, a MCMC strategy is adopted for inference in the new model, and novel proposal mechanisms are developed for this purpose.

For a stark example of the importance of these modifications, see Fig. 1, in which the standard harmonic model of [4, 5] is compared with the new (time-varying amplitude) model over the start attack region of a saxophone note. The harmonic modelled data (centre of figure) is visually much improved through use of the time-varying amplitude model. The residual noise after harmonic components have been extracted (right hand figures) is also remarkably improved through use of the new time-varying amplitude model (note different y-axis scalings). This has obvious advantages from a pure modelling perspective, but is likely to impinge significantly on analysis of quantities such as pitch, especially in the polyphonic setting.

Preliminary results are presented for monophonic pitch estimation, showing the effectiveness of the model for real data. We have also extended the methods to the polyphonic setting and will report results in future papers. We anticipate that the methods will find application in instrument recognition, music transcription, source separation and restoration.

The paper is organized as follows. In Section 2, we present the new harmonic model and specify the probabilistic framework. In Section 3, we describe the MCMC algorithm implemented. Finally, simulation results are presented Section 4.

This work was sponsored by the European research project MOUMIR, http://www.moumir.org

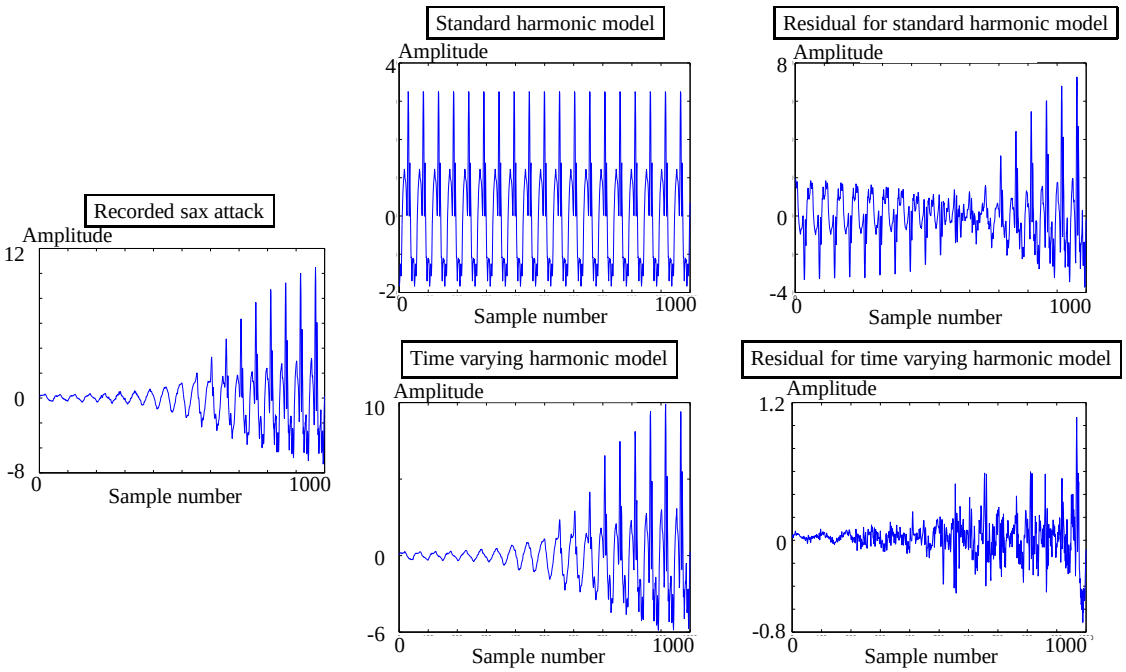

Fig. 1. Note attack from extract 'sax' (left) modelled in two different ways: firstly, with a constant amplidude model (upper figures), secondly with time-varying amplitudes (lower figures).

2. BAYESIAN MODEL

For simplicity, we assume that the original musical signal has been segmented into individual notes by using e.g. the technique presented in [11]. Here we present the model in a single note setting, noting that the polyphonic case can be obtained by a superposition of several such notes in a similar manner to [4, 5]. The model is as follows:

$$y[t] = \sum_{m=1}^{M} \{a_m[t] w_m \cos[(m+\delta_m)\,\omega_0\,t] + b_m[t] w_m \sin[(m+\delta_m)\,\omega_0\,t]\} + v[t] \quad (1)$$

where $t = 0, \ldots, N-1$ is the discrete time index. In Eq. (1), the unknowns are the amplitudes $a_m[t]$ and $b_m[t]$, the fundamental frequency ω_0, the number of partials $M > 0$, and the de-tuning parameters δ_m, denoted $\boldsymbol{\delta}_M = [\delta[1], \ldots, \delta[M]]^T$. The error $v[t]$ is modelled by a P-order Gaussian AR process:

$$v[t] = \alpha_1 v[t-1] + \ldots + \alpha_P v[t-P] + \epsilon[t]$$

where $\epsilon[t]$ is a zero mean Gaussian white noise with variance σ_ϵ^2. The weights w_m are tuned to the average decay of musical partials with increasing frequency, so that the amplitudes $(a_m^2[t] + b_m^2[t])^{1/2}$ are all on a similar scale, see Subsection 2.2. Note that the form of w_m defines our prior knowledge about the expected rate of decay of partials with increasing frequency. We believe this introduces an added element of realism into the model compared with earlier harmonic approaches.

Many evolution models are possible for amplitude processes $a_m[t]$ and $b_m[t]$, including random walks, autoregressions, etc., and many would be tractable within our Bayesian framework. In our specific implementation, in order to reduce the dimensionality of the model and induce smoothness in the amplitude evolution with time, the amplitudes are projected onto basis functions ϕ_i, $i = 0, \ldots, I$ (with I fixed and known) such that:

$$a_m[t] = \sum_{i=0}^{I} a_{m,i}\phi_i[t]\ ; \quad b_m[t] = \sum_{i=0}^{I} b_{m,i}\phi_i[t]$$

The functions ϕ_i are obtained by translating in time a prototype function $\phi[t]$ (typically a spline, Hamming window, etc.):

$$\phi_i[t] = \phi[t - i\Delta_t]$$

where Δ_t is the time offset between adjacent translations. In our implementation we have employed a length Q Hanning window for ϕ with $\Delta_t = Q/2$ (50% overlap).

Under this general formulation, the model of Eq. (1) becomes:

$$y[t] = \sum_{m=1}^{M} w_m \sum_{i=0}^{I} \{a_{m,i}\cos[(m+\delta_m)\omega_0\,t] + b_{m,i}\sin[(m+\delta_m)\omega_0\,t]\}\,\phi[t-i\Delta_t] + v[t] \quad (2)$$

$$= \mathbf{d}[t]^T\boldsymbol{\theta} + v[t] \quad (3)$$

where $\boldsymbol{\theta} = [a_{1,0}\, b_{1,0}\, a_{2,0}\, b_{2,0} \ldots a_{M,I}\, b_{M,I}]^T$ and the elements of $\mathbf{d}[t]$ are constructed correspondingly to satisfy Eq. (2). Now, defining $\mathbf{y} = [y[0], \ldots, y[N-1]]^T$, $\mathbf{v} = [v[0], \ldots, v[N-1]]^T$ and $\mathbf{D} = [\mathbf{d}[0] \ldots \mathbf{d}[N-1]]^T$, Eq. (3) can be rewritten in the form of the general linear model as

$$\mathbf{y} = \mathbf{D}\boldsymbol{\theta} + \mathbf{v} \quad (4)$$

Here $\mathbf{D}\boldsymbol{\theta}$ corresponds to a basis function expansion of the data in terms of windowed $\sin()$ and $\cos()$ functions (the columns of $\mathbf{D}$). For example, using a Hanning window of length $Q = 400$ with 50% overlap and $I = 3$ leads, for one particular harmonic frequency, to the cosine basis functions shown in Fig. 2.

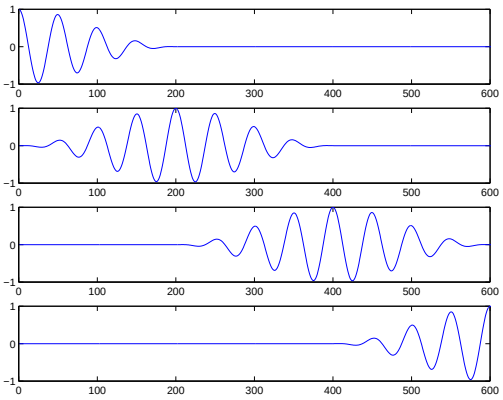

Fig. 2. Typical set of basis functions $\phi_i \cos(\cdot)$

2.1. Likelihood function

Given the linear model formulation and the assumption of i.i.d. Gaussian excitation for the AR process, we immediately obtain the likelihood function (see [4, 8, 9]):

$$p(\mathbf{y}|\boldsymbol{\theta}, \omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, \sigma_\epsilon^2, M) \;=\; \frac{1}{(2\pi\sigma_\epsilon^2)^{(N-P)/2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\theta})^T\mathbf{A}^T\mathbf{A}(\mathbf{y}-\mathbf{D}\boldsymbol{\theta})\right]$$

where $\mathbf{A}$ is the matrix

$$\mathbf{A} \;=\; \begin{bmatrix} -\alpha_P & \dots & -\alpha_1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\alpha_P & \dots & -\alpha_1 & 1 \end{bmatrix}$$

2.2. Parameter priors

The model structure selected leads naturally to the following prior structure:

$$p(\boldsymbol{\theta}, \omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, \sigma_\epsilon^2, M) \;=\; p(\boldsymbol{\theta}|\omega_0, \boldsymbol{\delta}_M, \sigma_\epsilon^2, M) \times p(\boldsymbol{\delta}_M|\omega_0, M)\, p(M|\omega_0)\, p(\boldsymbol{\alpha})\, p(\omega_0)\, p(\sigma_\epsilon^2)$$

The form for each of these distributions should reflect prior beliefs about *generic* musical signals without unfairly guiding the posterior towards inferences based upon examining the *particular* extract under consideration. Briefly, $p(\sigma_\epsilon^2)$ is inverted gamma with parameters α_ϵ and β_ϵ, $p(\omega_0)$ is the truncated Jeffreys distribution $p(\omega_0) \;=\; 1/\omega_0 \mathbb{I}_{[0,\pi]}$, $p(M|\omega_0)$ is a Poisson distribution with parameter Λ truncated to $[1, \min(M_{\max}, \pi/\omega_0]$, $p(\boldsymbol{\alpha})$ is zero-mean Gaussian with covariance matrix $\boldsymbol{\Sigma}_\alpha$, and $p(\boldsymbol{\delta}_M|\omega_0, M)$ is the zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}_\delta \;=\; \sigma_\delta^2\mathbf{I}_M$ (where $\mathbf{I}_M$ is the identity matrix of size M), restricted such that $m\omega_0 + \delta_m \in [0, \pi]$. Finally, the amplitudes prior is a zero-mean Gaussian distribution with covariance matrix $\sigma_\epsilon^2\boldsymbol{\Sigma}_\theta$, since we expect harmonic amplitudes to scale relative to the residual noise energy. Various forms have been coded up and investigated for $\boldsymbol{\Sigma}_\theta$, including $\xi^2\mathbf{I}$, the identity matrix (independent amplitude components for the harmonics), and the g-prior, which has some convenient properties for model selection ([4, 5, 7] have all employed g-priors in related sinusoidal modelling contexts):

$$\boldsymbol{\Sigma}_\theta \;=\; \xi^2\left(\mathbf{D}^T\mathbf{A}^T\mathbf{A}\mathbf{D}\right)^{-1}$$

We assume the g-prior for the remainder of this paper and will present a comparative analysis of various alternatives in future work.

2.3. Posterior distribution

Under the assumed prior structure above, it is straightforward to integrate out the amplitude parameter $\boldsymbol{\theta}$ and the noise variance σ_ϵ^2, to obtain the following posterior distribution, which is defined over the prior support for all the parameters (see last section):

$$p(\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, M|\mathbf{y}) \propto (\mathbf{y}^T\mathbf{P}_M\mathbf{y} + 2\beta_\epsilon)^{-N/2-\alpha_\epsilon} \times \frac{1}{\omega_0}\frac{1}{(M-1)!}\left[\frac{\Lambda}{(1+\xi^2)^{I+1}\sqrt{2\pi}\sigma_\delta}\right]^M \times \exp\left(-\frac{1}{2\sigma_\delta^2}\sum_{m=1}^{M}\delta_m^2\right) \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\alpha}\right) \tag{5}$$

Where $\mathbf{P}_M \;=\; \mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{A}\mathbf{D}\mathbf{S}_M^T\mathbf{D}^T\mathbf{A}^T\mathbf{A}$ with $\mathbf{S}_M^{-1} \;=\; \mathbf{D}^T\mathbf{A}^T\mathbf{A}\mathbf{D} + \boldsymbol{\Sigma}_\theta^{-1}$. Other conditional distributions used in our algorithms are as follows (with $\boldsymbol{\mu} = \mathbf{S}_M\mathbf{D}^T\mathbf{A}^T\mathbf{A}\mathbf{y}$):

$$p(\sigma_\epsilon^2|\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, M, \mathbf{y}) \;=\; \mathcal{IG}(\frac{N}{2} + \alpha_\epsilon, \frac{\mathbf{y}^T\mathbf{P}_M\mathbf{y}}{2} + \beta_\epsilon) \tag{6}$$

$$p(\boldsymbol{\theta}|\sigma_\epsilon^2, \omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, M, \mathbf{y}) \;=\; \mathcal{N}(\boldsymbol{\mu}, \sigma_\epsilon^2\mathbf{S}_M) \tag{7}$$

3. BAYESIAN COMPUTATION

Estimating the parameters in the model of Eq. (4) requires to compute multidimensional integrals of a function f of the form:

$$\int_\Omega f(\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma_\epsilon^2)p(\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma_\epsilon^2|y)\, d\omega_0 d\boldsymbol{\delta}_M d\boldsymbol{\alpha} d\boldsymbol{\theta} d\sigma_\epsilon^2$$

Standard numerical techniques are generally inaccurate, and we apply MCMC techniques in order to create a Markov Chain (MC) $\{\widetilde{\omega}_0^{(l)}, \widetilde{\boldsymbol{\delta}}_M^{(l)}, \widetilde{\boldsymbol{\alpha}}^{(l)}\}$ whose stationary density is $p(\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}|y)$. The estimates are then computed by the following Monte Carlo average (written for ω_0):

$$\widehat{\omega}_0 \;=\; \frac{1}{L}\sum_{l=1}^{L}\widetilde{\omega}_0^{(l)} \text{ corresponding to } f(\omega_0, \boldsymbol{\delta}_M, \boldsymbol{\alpha}) = \omega_0 \tag{8}$$

In order to produce the MC samples, we propose to implement the following Metropolis-Hastings (MH) MCMC algorithm, inspired from [7] (the proposal distributions are details below):

MCMC algorithm for harmonic models

1. Initialization.
   - Sample $\widetilde{\omega}_0^{(0)} \sim \mathcal{U}[0, \pi/M]$, Sample $\widetilde{\boldsymbol{\alpha}}^{(0)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$
   - Sample $\widetilde{\boldsymbol{\delta}}_M^{(0)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$ and set $l \leftarrow 1$
2. While $l < L$, do
   - With probability λ_1 perform a MH step with proposal distribution $q(\omega_0, \boldsymbol{\delta}_M|\mathbf{y})$, and with probability $1-\lambda_1$, perform a one-variable-at-a-time MH step with Normal random walk proposal distribution w.r.t the target distribution $p(\omega_0, \boldsymbol{\delta}_M, \widetilde{\boldsymbol{\alpha}}^{(l-1)}|\mathbf{y})$ given Eq. (5)

- With probability λ_2 perform a MH step with the uniform proposal distribution $\mathcal{U}([\alpha_{\min};\alpha_{\max}])$ and with probability $1-\lambda_2$, perform a MH step with a Normal random walk proposal distribution w.r.t the target distribution $p(\widetilde{\omega}_0^{(l)}, \widetilde{\boldsymbol{\delta}}_M^{(l)}, \boldsymbol{\alpha}|\mathbf{y})$ given Eq. (5)
- Sample $\widetilde{\sigma}_\epsilon^{2\,(l)} \sim p(\sigma_\epsilon^2|\widetilde{\omega}_0^{(l)}, \widetilde{\boldsymbol{\delta}}_M^{(l)}, \widetilde{\boldsymbol{\alpha}}^{(l)}, M, \mathbf{y})$ given Eq. (6)
- Sample $\widetilde{\boldsymbol{\theta}}^{(l)} \sim p(\boldsymbol{\theta}|\widetilde{\sigma}_\epsilon^{2\,(l)}, \widetilde{\omega}_0^{(l)}, \widetilde{\boldsymbol{\delta}}_M^{(l)}, \widetilde{\boldsymbol{\alpha}}^{(l)}, M, \mathbf{y})$ given Eq. (7)
- Set $l \leftarrow l+1$

---

where the proposal distribution $q(\omega_0, \boldsymbol{\delta}_M|\mathbf{y})$ is chosen in order to ensure a high acceptance reta, namely

$$q(\omega_0, \boldsymbol{\delta}_M|\mathbf{y}) = \left[\prod_{m=1}^{M} q(\delta_m|\mathbf{y}, \omega_0)\right] q(\omega_0|\mathbf{y}) \qquad (9)$$

where $q(\omega_0|\mathbf{y}) \propto \mathcal{F}^y_{[0;\pi/M]}(\omega_0)$ and $\mathcal{F}^y_{[0;\pi/M]}$ denotes the discrete spectrum of y restricted to $\omega_0 \in [0;\pi/M]$. Moreover, $q(\delta_M|\mathbf{y}) = \mathcal{N}(0;\sigma_\delta^2)$.

In the case where M is unknown, it can be sampled by considering the following moves:

- **Addition of a partial** (with probability $a_{\widetilde{M}^{(l)}}$). This move consists of setting $\widetilde{M}^{(l)} \leftarrow \widetilde{M}^{(l-1)}+1$ and sampling $\widetilde{\delta}_{\widetilde{M}^{(l)}} \sim \mathcal{N}(0;\sigma_\delta^2)$;
- **Removal of a partial** (with probability $r_{\widetilde{M}^{(l)}}$). This move consists of setting $\widetilde{M}^{(l)} \leftarrow \widetilde{M}^{(l-1)} - 1$ and removing the highest order partial;
- **Parameters update** (with probability $u_{\widetilde{M}^{(l)}}$). This move consists of updating the parameters $\{\widetilde{\omega}_0^{(l-1)}, \widetilde{\boldsymbol{\delta}}_{M^{(l-1)}}^{(l-1)}, \widetilde{\boldsymbol{\alpha}}^{(l-1)}\}$ by using the algorithm presented above.

where ($a_M + r_M + u_M = 1$), with $a_{M_{\max}} = 0$ and $r_1 = 0$. These probabilities are more precisely:

$$a_M = 0.5 \min\left\{1, \frac{p(M+1)}{p(M)}\right\} \; d_{M+1} = 0.5 \min\left\{1, \frac{p(M)}{p(M+1)}\right\}$$

4. SIMULATION RESULTS

Tests have been carried out on both monophonic and polyphonic musical extracts. Here we report summary results from analysis of a short solo flute extract (the opening of Debussy's Syrinx) downsampled to 22050 Hz sample rate. This demonstrates the high reliability and accuracy of the models for pitch estimation. The waveform was arbitrarily segmented into blocks of duration 0.1s with 50% overlap and the monophonic algorithm applied in turn to each block. The pitch estimates obtained are shown in Fig. 3. 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 short period of silence. Informal examination of spectrograms indicated that the reversible jump algorithm for determining the number of harmonics was very successful.

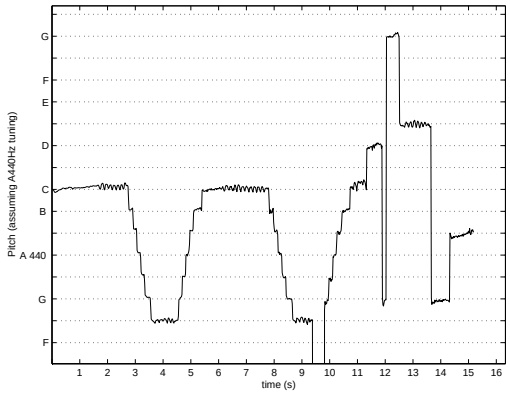

Fig. 3. Pitch estimation from flute extract

5. DISCUSSION

In this paper we have described new models and algorithms for harmonic analysis of musical audio. The methods have been demonstrated reliably in operation with monophonic material. We have extended the models to the polyphonic setting using reversible jump MCMC to determine automatically the numbers of notes and harmonics playing at any given time, and will report detailed algorithms and results in future work.

6. REFERENCES

[1] R.C. Maher, "Fundamental frequency estimation of musical signals using a two-way mismath procedure," *J. Acoust. Soc. of Am.*, vol. 95, no. 4, pp. 2254–2263, April 1994.

[2] A. Klapuri, "Pitch estimation using multiple independent time-frequency windows," 1999.

[3] J. C. rown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes in the fourier transform," *J. Acoust. Soc. Am.*, vol. 94, no. 2, pp. 662–667, 1993.

[4] P.J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Multidimensional optimisation of harmonic signals," in *Proc. European Conference on Signal Processing*, Sept. 1998.

[5] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," in *Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State*, Mohonk, NY State, Oct. 1999.

[6] S. J. Godsill and P. J. W. Rayner, "Robust noise reduction for speech and audio signals," in *Proc. IEEE ICASSP-96*, May 1996.

[7] C. Andrieu and A. Doucet, "Joint Bayesian Detection and Estimation of Noisy Sinusoids via Reversible Jump MCMC," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2667–2676, 1999.

[8] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model-Based Approach*, Berlin: Springer, ISBN 3 540 76222 1, Sept. 1998.

[9] S. J. Godsill and P. J. W. Rayner, "Statistical reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler," *IEEE Trans. Speech, Audio Proc.*, vol. 6, no. 4, pp. 352–372, July 1999.

[10] M. Davy, C. Doncarli, and J. Y. Tourneret, "Classification of chirp signals using hierarchical bayesian learning and mcmc methods," *IEEE Trans. Signal Processing*, 2001, to appear.

[11] M. Davy and S.J. Godsill, "Detection of abrupt spectral changes using support vector machines. An application to audio signal segmentation," in *Proc. IEEE ICASSP-02*, 2002, submitted.