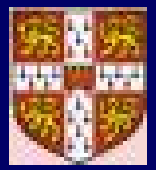
Bayesian harmonic models for musical signal analysis

Simon Godsill and Manuel Davy

June 2, 2002

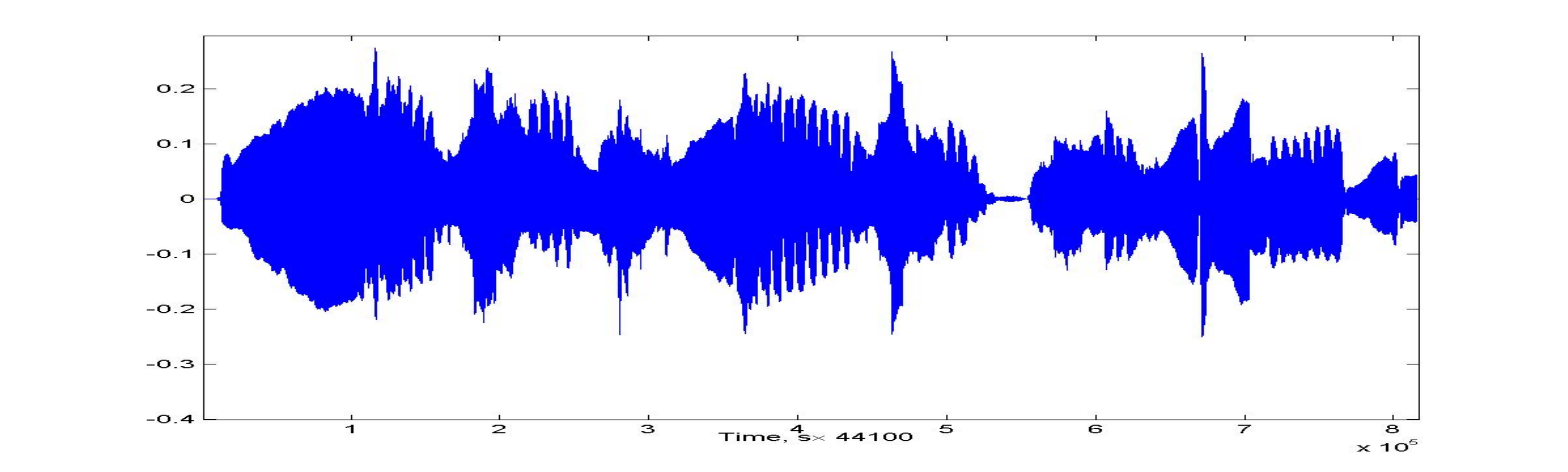
Cambridge University Engineering Department and
IRCCyN – UMR CNRS 6597

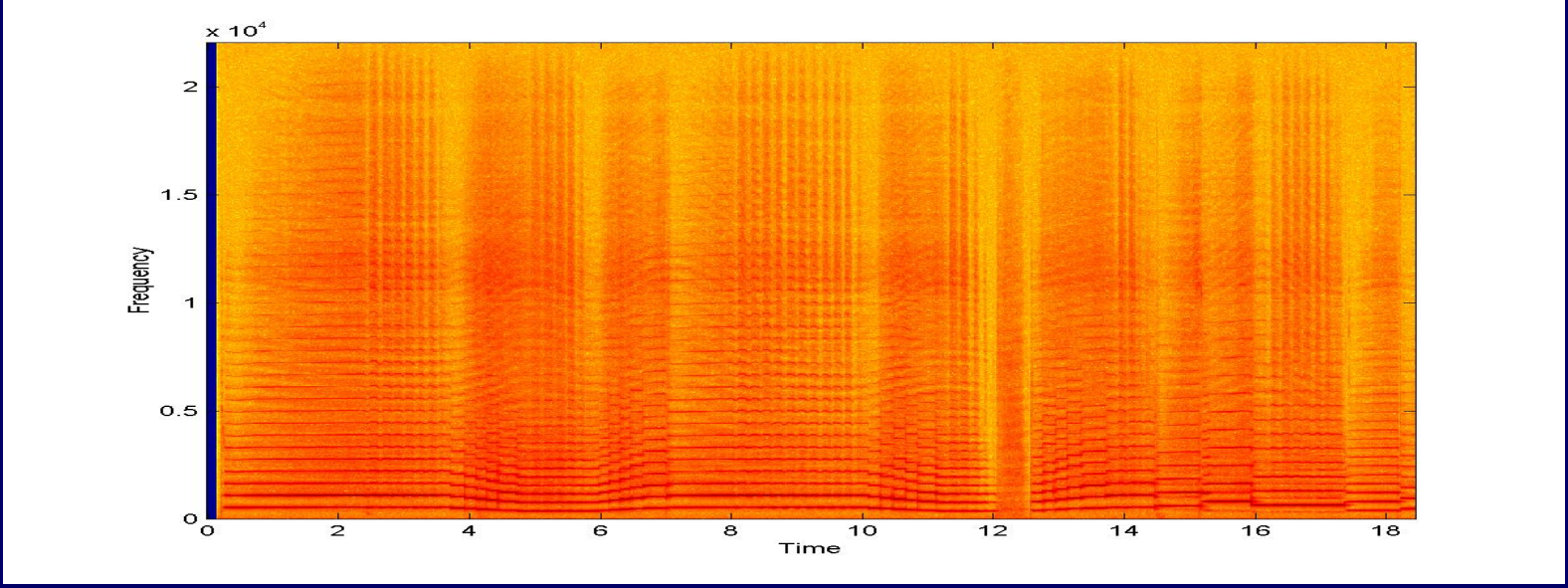


The work of both authors was partially supported by the EU project MOUMIR

Bayesian music models: Motivation

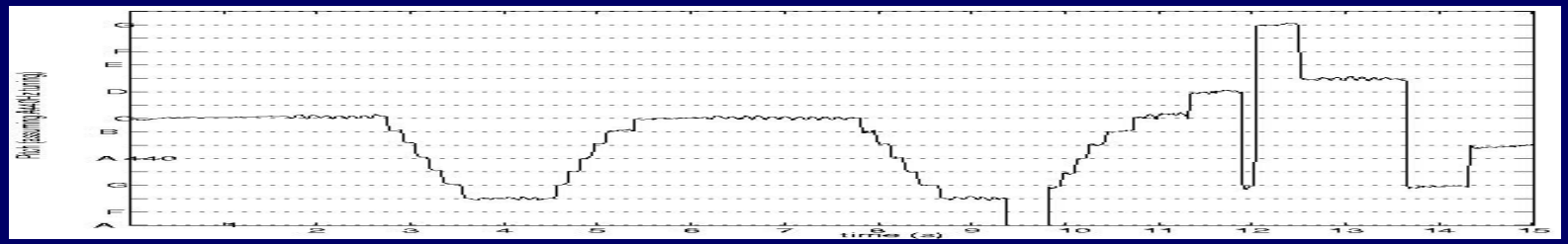
A typical musical signal and its spectrogram (solo flute):





The aim is to perform inference about the high level structures in music from the waveform alone:

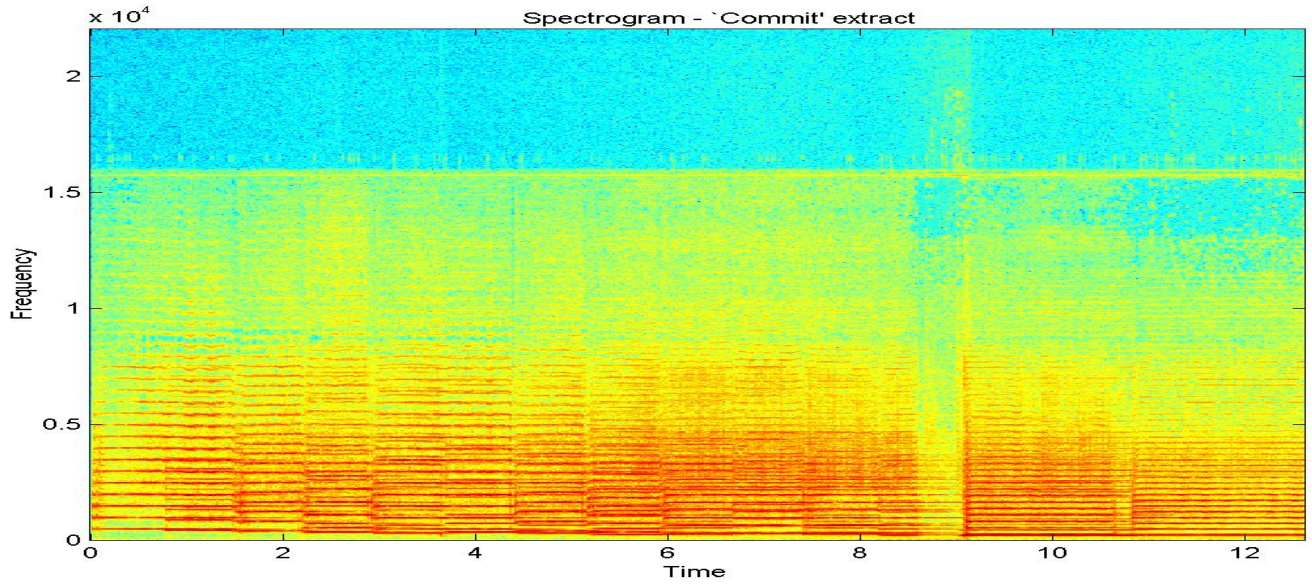
- Musical pitch

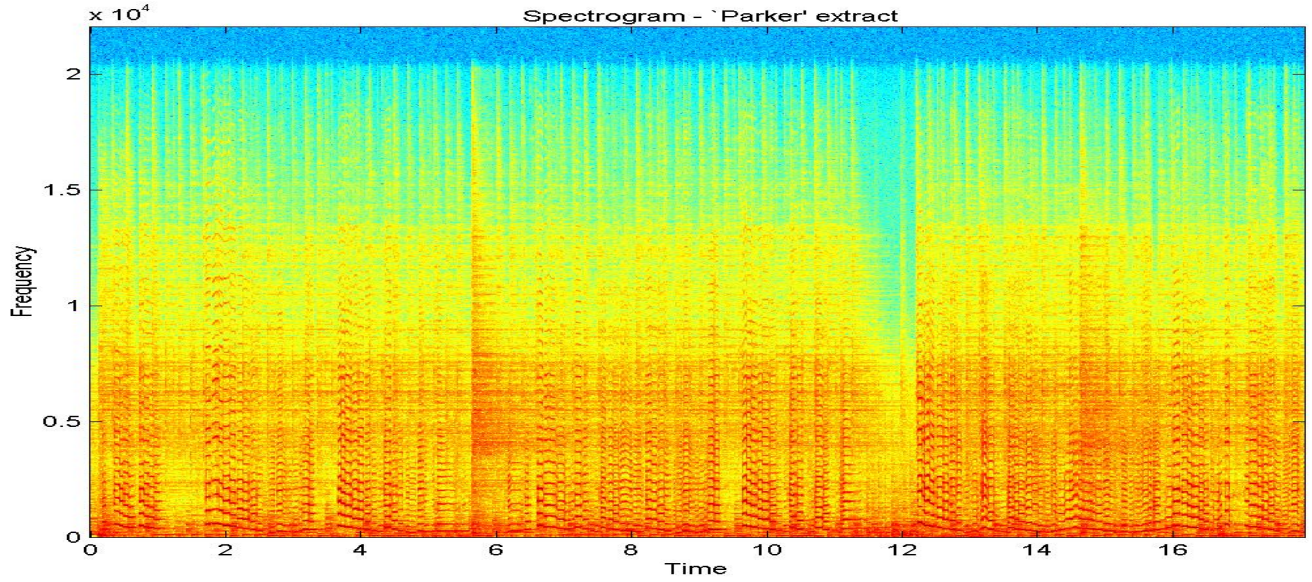


- Instrument classification
- Timbre
- Signal separation
- ...many other possible inferences depending on the application ...

These are all probabilistic questions that are naturally formulated in a Bayesian framework.

The problem is substantially more complex in the many instrument, or 'polyphonic', case:


Spectrogram - `Commit' extract
x 10^4
Frequency
Time
0
0.5
1
1.5
2
0
2
4
6
8
10
12


Spectrogram - `Parker' extract
x 10^4
Frequency
Time
0
0.5
1
1.5
2
0
2
4
6
8
10
12
14
16

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

Overview

1. Music modelling issues
2. Other work
3. Bayesian harmonic models
4. Computations
5. Results
6. Discussion

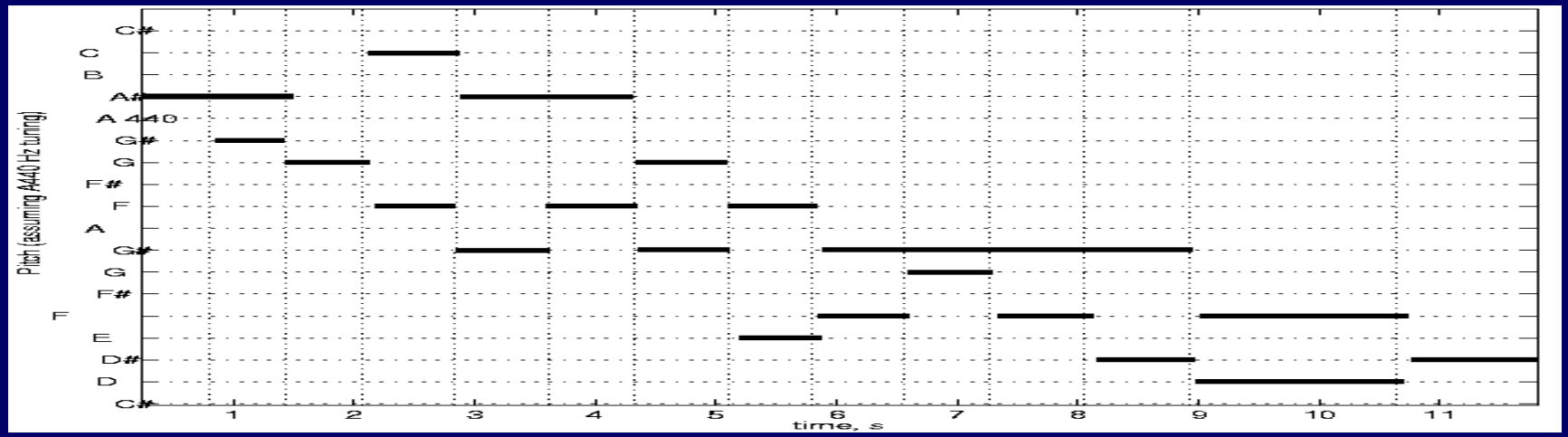
1. Music modelling issues

In my opinion the two key (interconnected) modelling issues for music are:

- Contextual modelling - i.e. the likely sequence of notes (n_t), both over time and within chords at fixed times. This is naturally a Bayesian prior model:

$$P(n_{1:t}) = P(n_{1:t}) = \prod_{\tau=1:t} P(n_\tau | n_{1:\tau-1})$$

[not necessarily Markovian]



- Accurate low level signal modelling. This involves a careful consideration of the physical sound generation mechanism and is ideally based on physical prior models of musical sound generation.

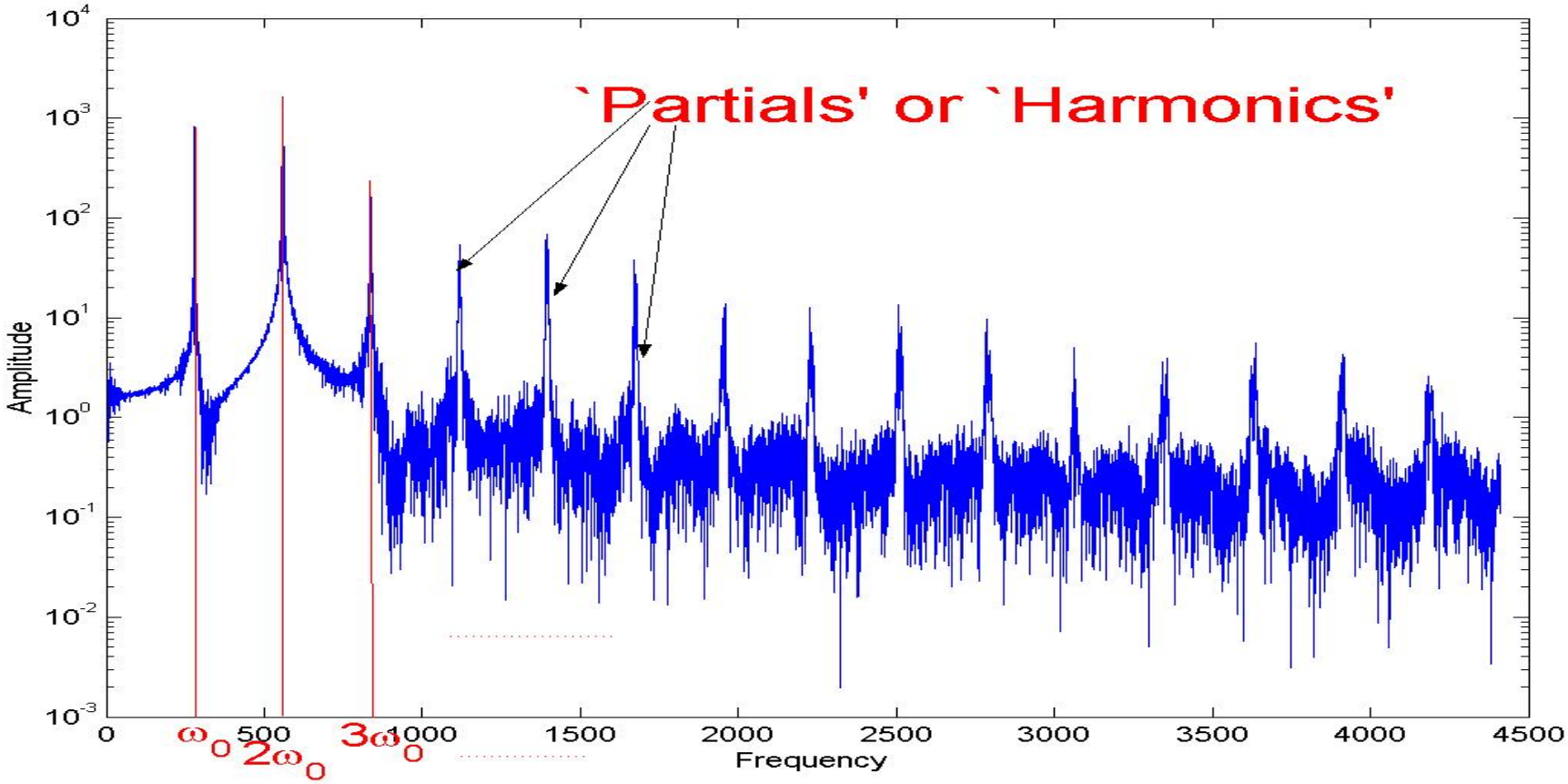
2. Other approaches: Literature review

- There is a large literature on the various aspects of this topic, beginning around 1975 (J.A. Moorer, Stanford). Most approaches are 'non-statistical', and nearly always not model-based. Most only apply to single-note (monophonic) data. Many are applied for specific instructions only.
- Typically features are extracted from the data, such as peaks in the spectrogram. Then grouping into notes is performed and finally some note consistency over time is enforced.
- Statistical approaches are rare, especially Bayesian, and there is no fully Bayesian approach to the problem to our knowledge (Kashino et al. (1993/1995/1999), Sterian et al. (1996-1999), include probabilistic models for aspects of the problem, Klapuri et al. (1996-2002) provide ML-based methods)

- In our view, the problem is sufficiently complex and rich in prior structure that a fully Bayesian approach to both context and detailed signal modelling are likely to yield best results.
- Here we concentrate on accurate Bayesian musical signal modelling, developing on our earlier Bayesian models - Walmsley, Godsill and Rayner (1998,1999).

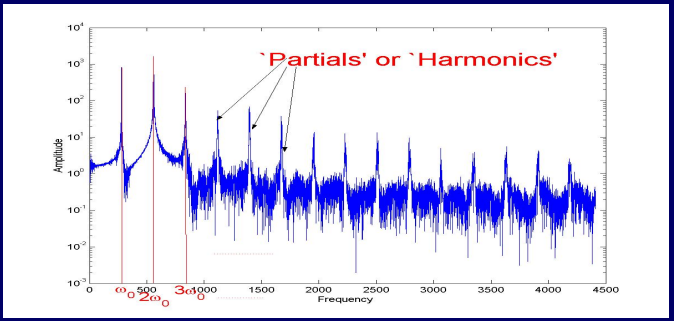
3. Bayesian Harmonic models for music:

$\Rightarrow$ Short-time Fourier Spectrum of a single note (flute)



Harmonic modelling of music: Simplistic model

$\Rightarrow$ Spectrum of a single note (flute)



$\Longrightarrow$ A simple one-note model:

$$y_t = \sum_{m=1}^{M} \alpha_m \cos[m\omega_0 t] + \beta_m \sin[m\omega_0 t] + v_t \qquad (1)$$

with

- ω_0 is the Fundamental frequency
- M is the total number of partials (m is the harmonic number)
- α_m and β_m are the partial amplitudes

The simplistic model is highly idealised, accounting only for a single note in a steady state (periodic) regime.

Amongst other things, it does not account for:

1. Time variations in fundamental frequency ω_0
2. Amplitude variations with time
3. Residual noise
4. Inharmonicity of partials (real instruments do not generate periodic waveforms)

The simplistic model is highly idealised, accounting only for a single note in a steady state (periodic) regime.

Amongst other things, it does not account for:

1. Time variations in fundamental frequency ω_0
2. Amplitude variations with time
3. Residual noise
4. Inharmonicity of partials (real instruments do not generate periodic waveforms)

The simplistic model is highly idealised, accounting only for a single note in a steady state (periodic) regime.

Amongst other things, it does not account for:

1. Time variations in fundamental frequency ω_0
2. Amplitude variations with time
3. Residual noise
4. Inharmonicity of partials (real instruments do not generate periodic waveforms)

The simplistic model is highly idealised, accounting only for a single note in a steady state (periodic) regime.

Amongst other things, it does not account for:

1. Time variations in fundamental frequency ω_0
2. Amplitude variations with time
3. Residual noise
4. Inharmonicity of partials (real instruments do not generate periodic waveforms)

A more general model which captures some of these effects is:

$$y_t = v_t + \sum_{m=1}^{M} \alpha_{m,t} \cos((m + \delta_m)\omega_{0,t}t) \; + \; \beta_{m,t} \sin((m + \delta_m)\omega_{0,t}t)$$

Such a general model is highly intractable and requires a very careful construction of the dynamics of the individual components (regularisation) to avoid frequency/amplitude ambiguities. We can, however, go some way towards the general case without losing tractability altogether, specifically:

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

with

- $\alpha_{m,t}$ projected onto smooth basis functions $\phi_{i,t}$, $\alpha_{m,t} \;=\; \sum_{i=1}^{I} a_{m,i}\phi_{i,t}$
- δ_m is aimed at modelling inharmonicity (de-tuning)
- ω_0 fixed over the (short) time-frame
- v_t is a correlated residual noise, modelled as an autoregressive Gaussian process

  $$v_t \;=\; \gamma_1 v_{t-1} + \gamma_2 v_{t-2} + \ldots + \gamma_p v_{t-p} + \epsilon_t$$

  with $\epsilon_t \sim \mathcal{N}(\epsilon_t; 0, \sigma_\epsilon^2)$

$\Rightarrow$ Extension to several notes

$$y_t = v_t + \sum_{k=1}^{K} \sum_{m=1}^{M_k} \alpha_{k,m,t} \cos\left[(m + \delta_{k,m})\omega_{0,k} t\right] + \beta_{k,m,t} \sin\left[(m + \delta_{k,m})\omega_{0,k} t\right]$$

with K the total number of notes

Probabilistic model: Posterior distribution

$\Rightarrow$ Based in this model, we can we can contruct the posterior distribution

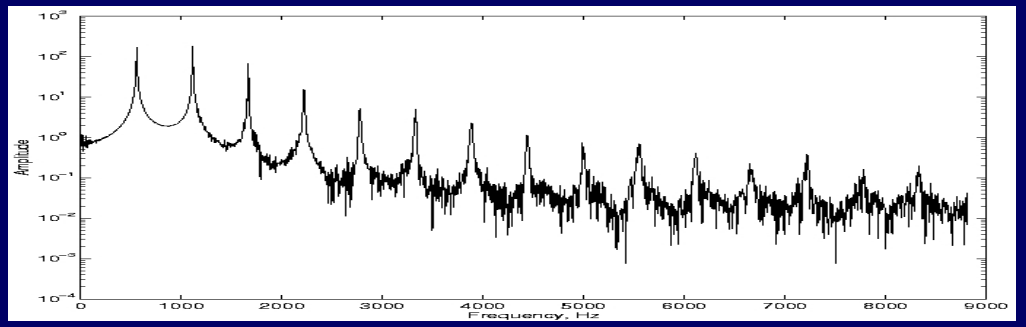
$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \mathbf{M}, K|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \mathbf{M}, K) \\ &\times p(\boldsymbol{\theta}|\boldsymbol{\omega}_0, \boldsymbol{\delta}, \sigma_\epsilon^2, \mathbf{M}, K)\, p(\boldsymbol{\delta}|\boldsymbol{\omega}_0, \mathbf{M}, K) \\ &\times p(\mathbf{M}|\boldsymbol{\omega}_0, K)\, p(\boldsymbol{\alpha})\, p(\boldsymbol{\omega}_0|K)\, p(\sigma_\epsilon^2)\, p(K) \end{aligned}$$

where the priors contain our prior knowledge about the musical parameter values (most are physically interpretable)

- The likelihood function is straightforwardly computed for given parameters.
- Consider the priors...

Probabilistic model: Prior distributions

The prior distributions are chosen based on physical considerations and intuition:



Some salient features are:

- Amplitudes of partials $\alpha_{m,t}$ and $\beta_{m,t}$: These decay with increasing frequency. They are modelled as independent, zero mean, Gaussians, with tapering variance as m increases (timbre, classification, ...)
- Detuning parameters δ_m: these take small values ($<< 1$) and are modelled as iid zero mean Gaussian with small variance (caveat...)

- Fundamental frequencies $\omega_{0,k}$: the prior should model the expected pitch clustering in frequency. Ideally will involve frequency domain interaction ('chords') and also time domain interaction ('melody'). We have incorporated neither as yet, but will do in future implementations
- Number of harmonics M_k: this is assigned a Poisson distribution - parameters can in principle be learned from real musical instrument sounds
- Number of notes K: again, a Poisson prior, reflecting the expected complexity of the music
- Other parameters: assigned (vague) conjugate distributions

4. Bayesian computations:

- We typically require MC approximations to integrals of the form

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

  where $\mathbf{\Phi} = \{\boldsymbol{\theta}, \boldsymbol{\omega}_0, \boldsymbol{\delta}, \mathbf{M}, K, \boldsymbol{\gamma}, \chi^2, \sigma_\epsilon^2\}$ is the collection of all unknowns in the model, $f(.)$ is a given integrable function with respect to the posterior and $\mathbf{\Omega}$ is the sample space for the posterior distribution.

- Choice of functional will be application dependent.

- For pitch transcription, we may require an MC estimate for $p(\boldsymbol{\omega}_0|\mathbf{y})$.

- For source separation we require estimates of the signals themselves: - note, however, inherent unidentifiability over the labelling of notes. This can be elegantly overcome by assigning notes labels according to the Western scale (A,B,C,...) [for Western music].

- The MCMC algorithm is complex, involving trans-dimensional moves for both number of harmonics M_k (for each note) and number of notes K. This is performed using reversible jump MCMC coupled with Metropolised product space ideas. The critical M-H proposal is for fundamental frequencies $\omega_{0,k}$. This is performed one-by-one using specially constructed local and global proposals, aimed at getting out of various 'traps' and ambiguities:
  - Independence proposal based on short-time spectrum of data
  - Independence proposal based on autocorrelation function of data
  - Gaussian random walk proposal
  - Proposal to frequencies in the set $\{\omega_0/3, \omega_0/2, 2\omega_0/3, 3\omega_0/2, 2\omega_0, 3\omega_0\}$. These powererful proposals eliminate many of the potential octave and fifth ambiguities inherent in music transcription systems.
  - Note - the full simulation is very slow! For longer extracts we fix K and set δ_m's to zero to make processing feasible.

Results: 'Commit' example

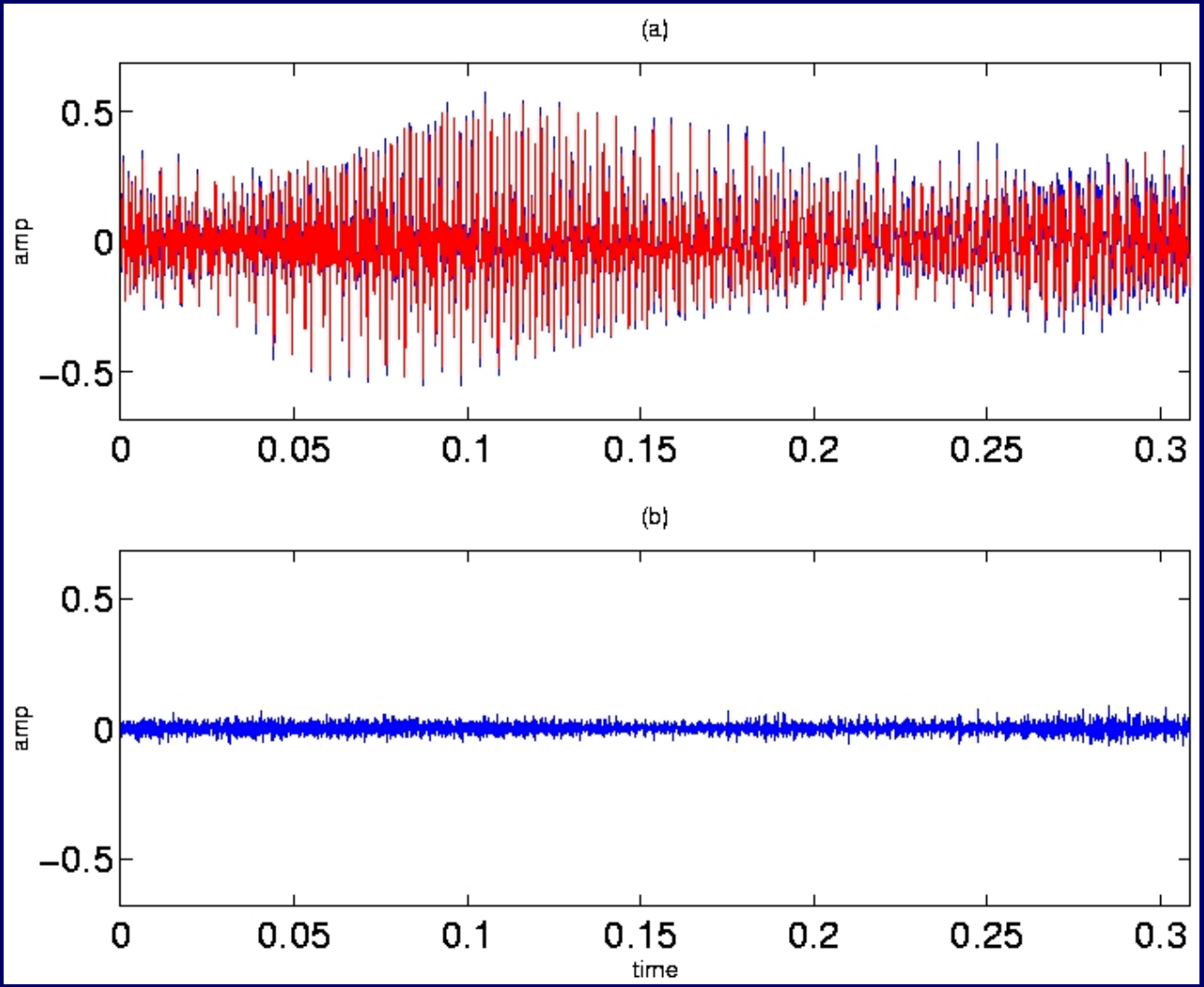
$\Rightarrow$ Two instruments playing: trumpet and saxophone

Results: 'Commit' example

$\Rightarrow$ Two instruments playing: trumpet and saxophone

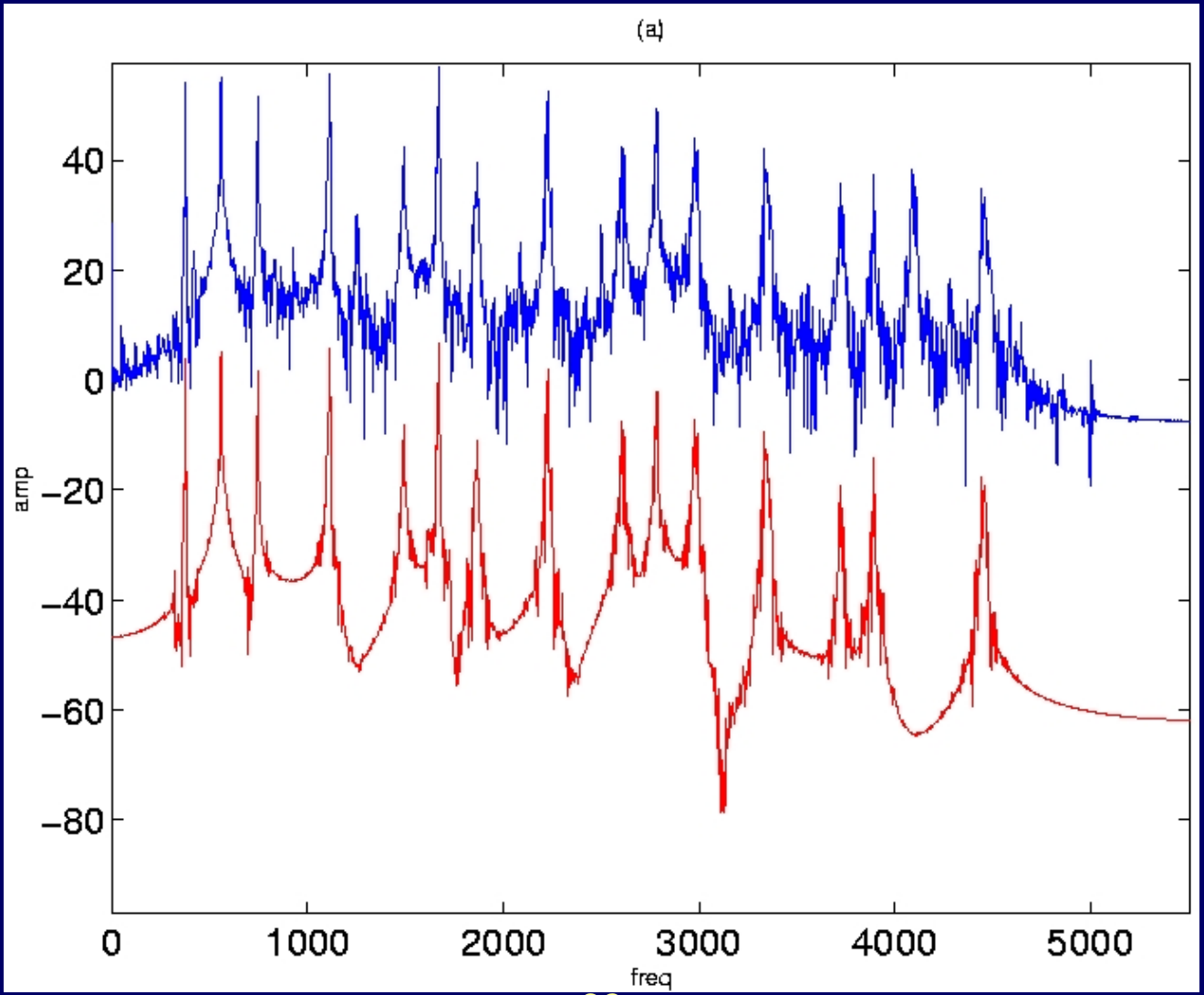
Results: 'Commit' example

⇒ Comparison of time series (reconstructed and original)



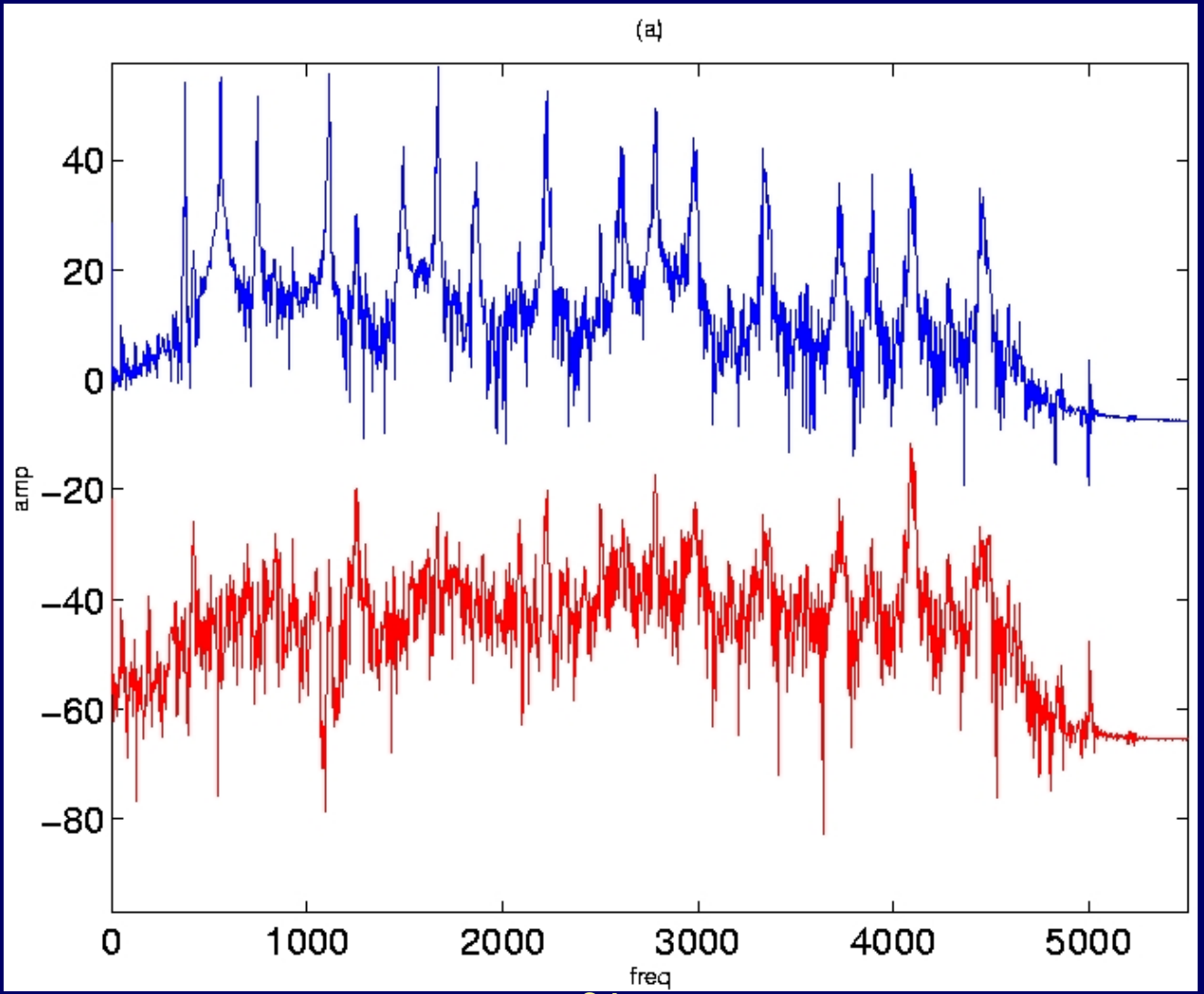
Results: 'Commit' example

⇒ Comparison of spectra (reconstructed and original)



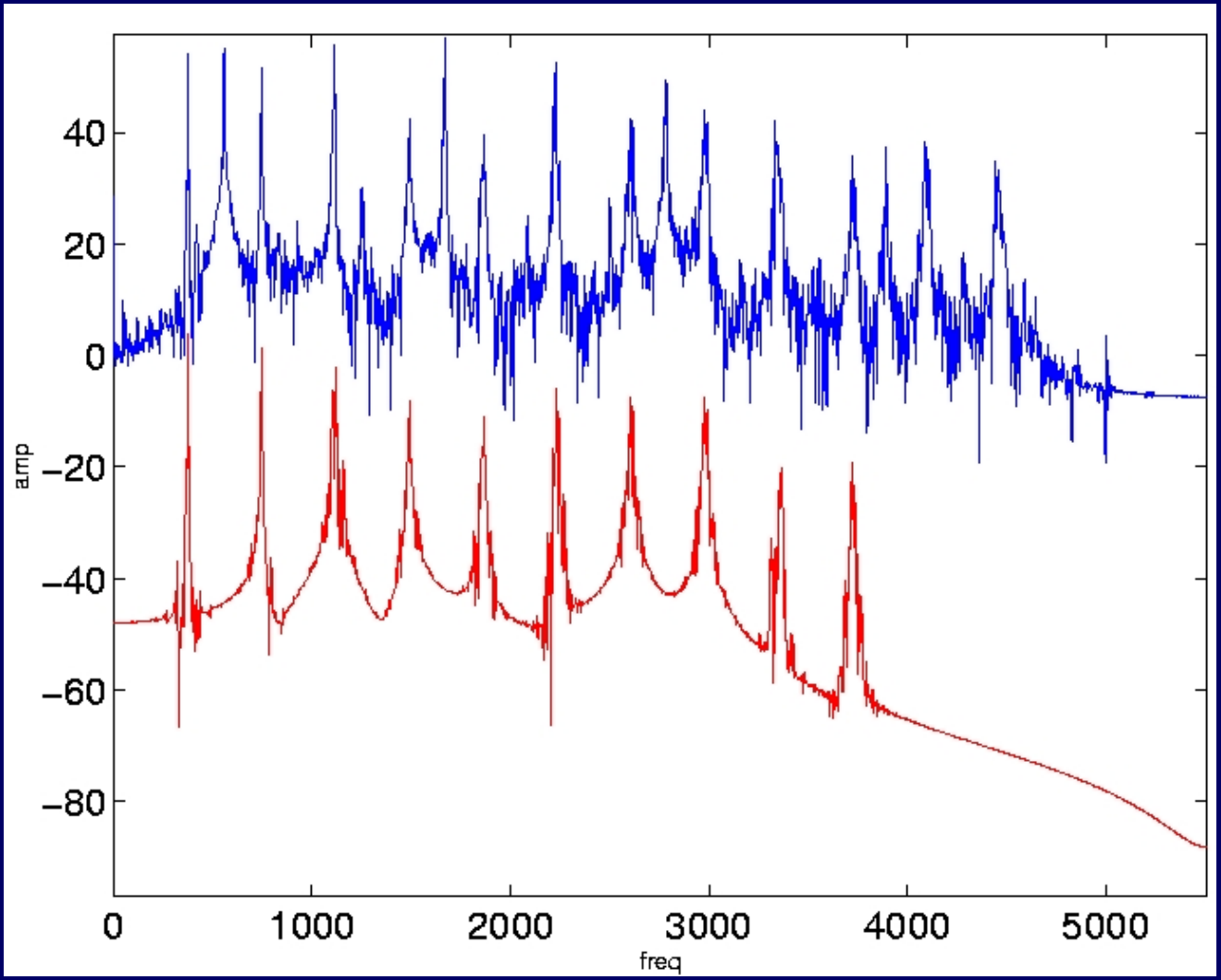
Results: 'Commit' example

⇒ Comparison of spectra (error signal and original)



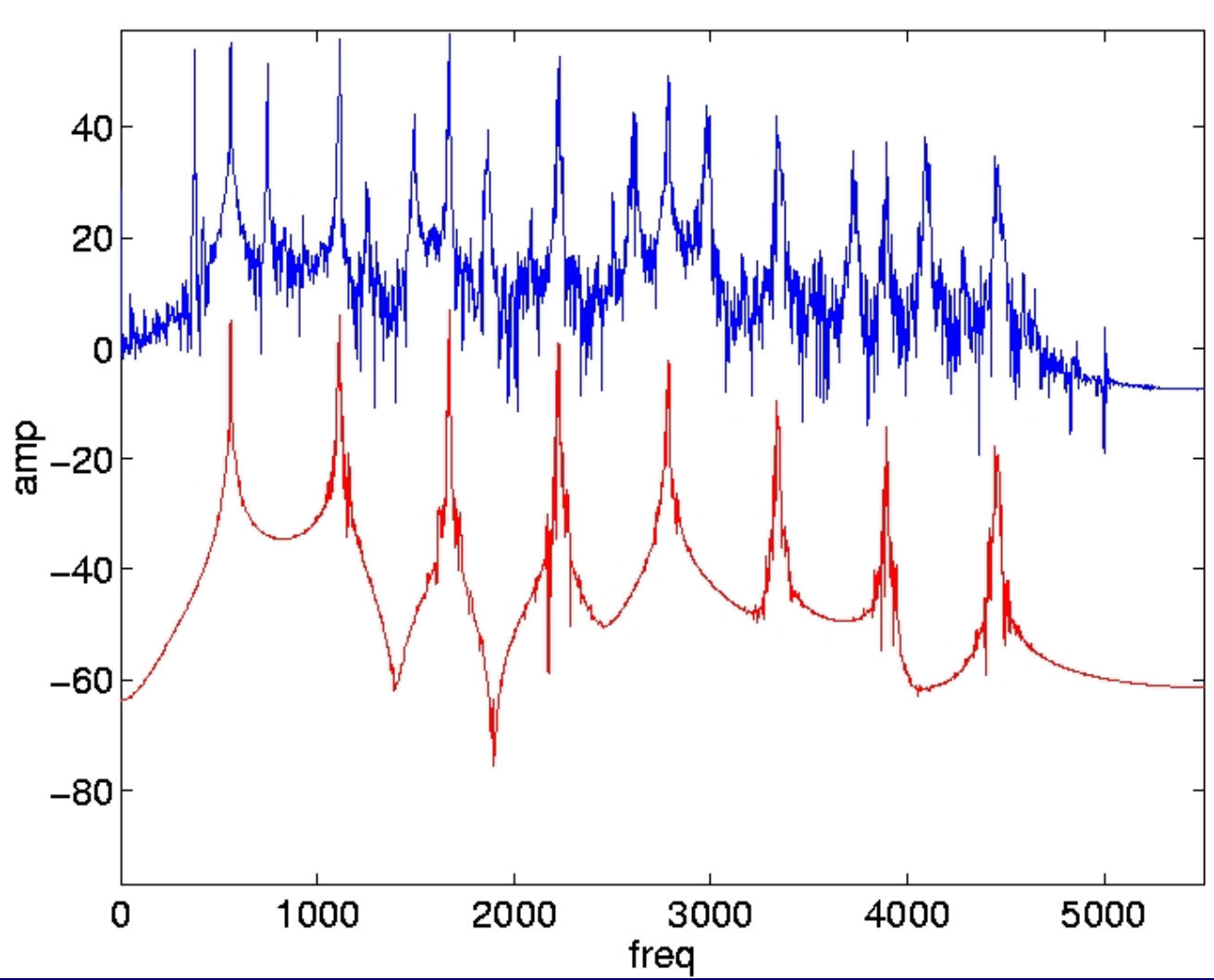
Results: 'Commit' example

⇒ Comparison of spectra (Note 1 and original)



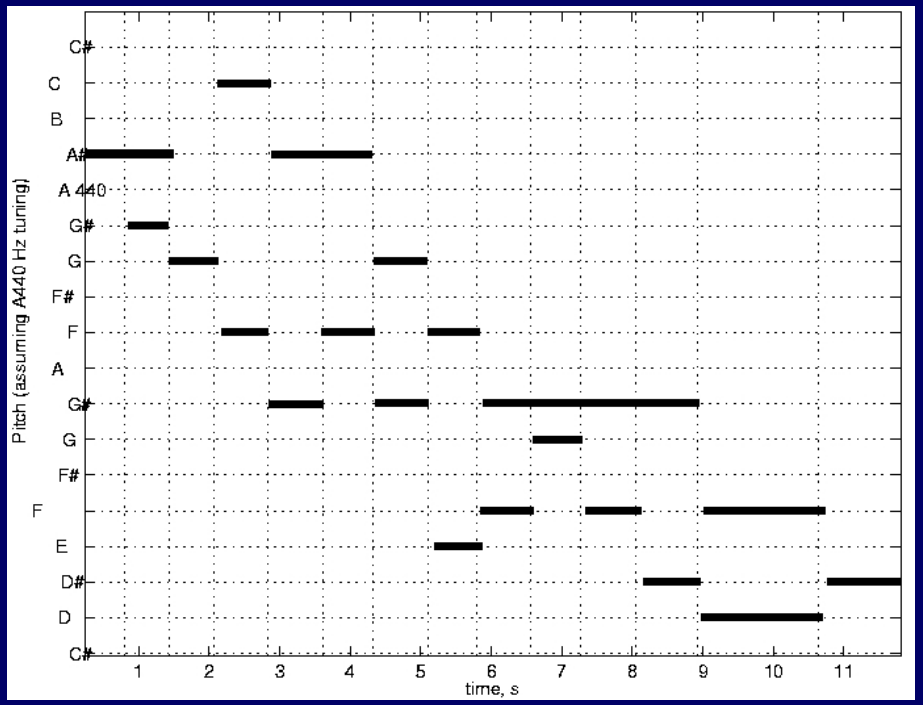
Results: 'Commit' example

⇒ Comparison of spectra (Note 2 and original)

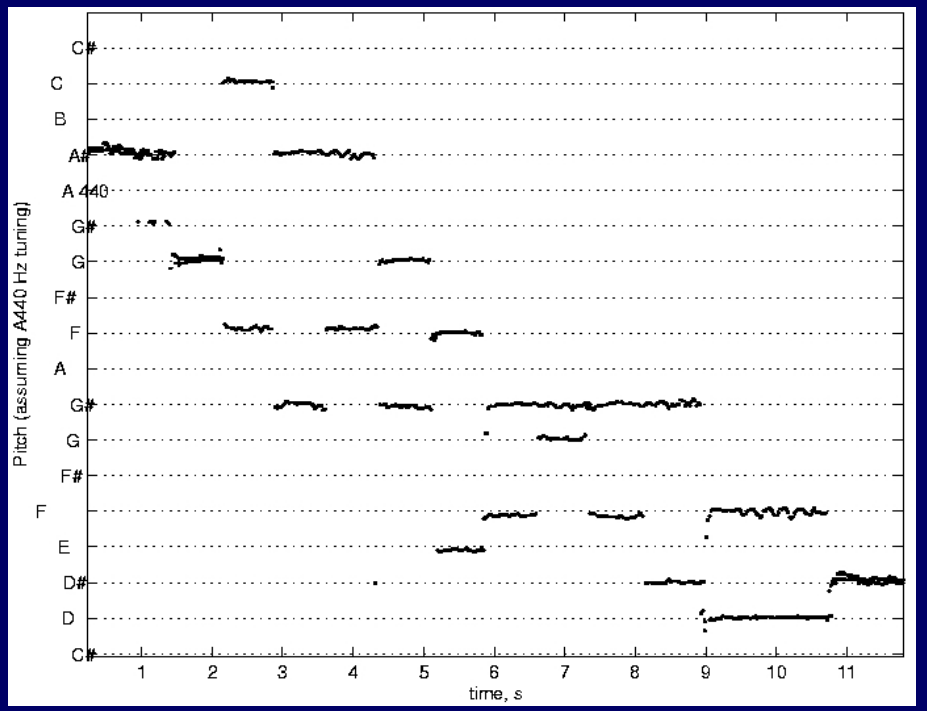


With a reduced model (fixed K and $\delta_{m,k} = 0$) we track the pitch for the entire extract:

Ground truth



Pitch estimation



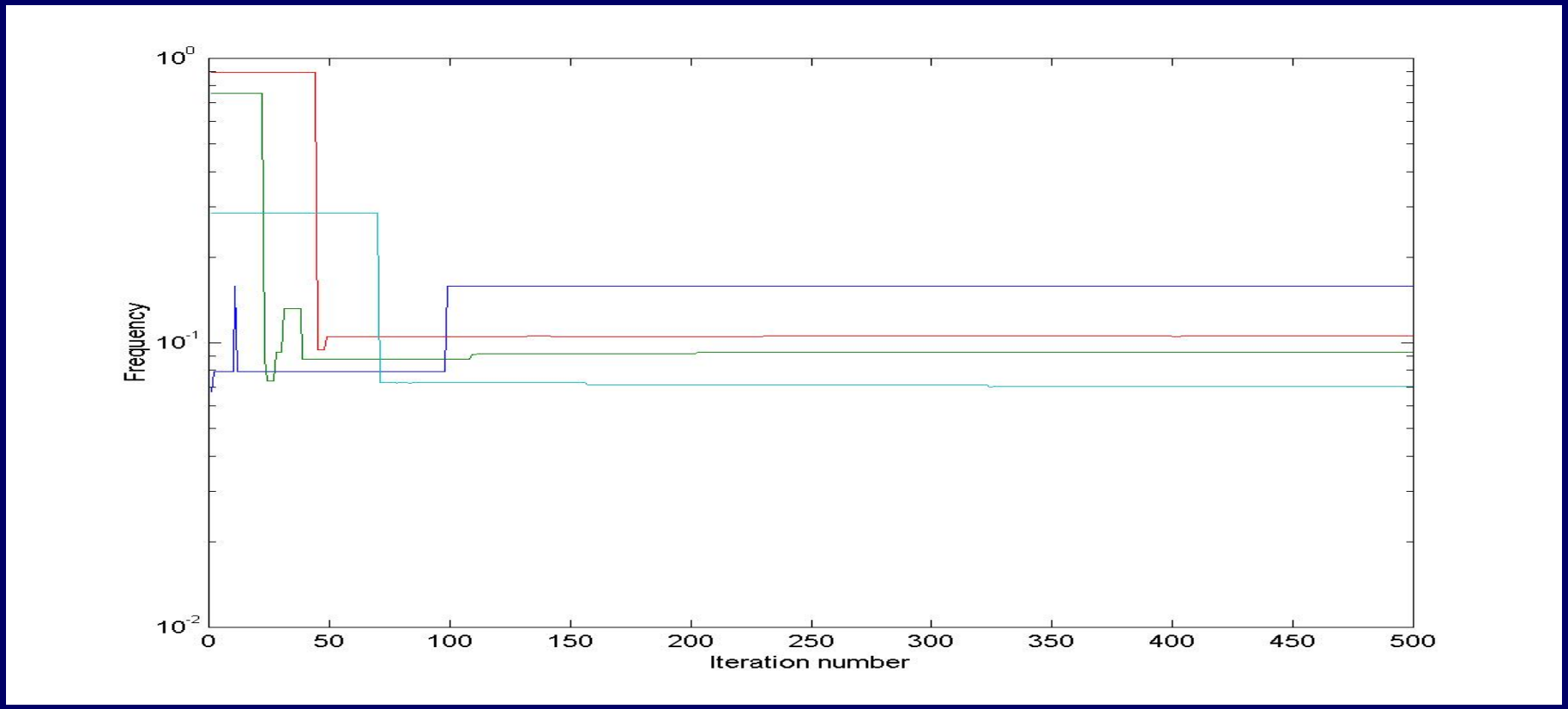
Results: Four note example

Sound examples (a new MCMC convergence diagnostic!?)

- Input extract repeated 70 times
- MCMC output during the convergence period

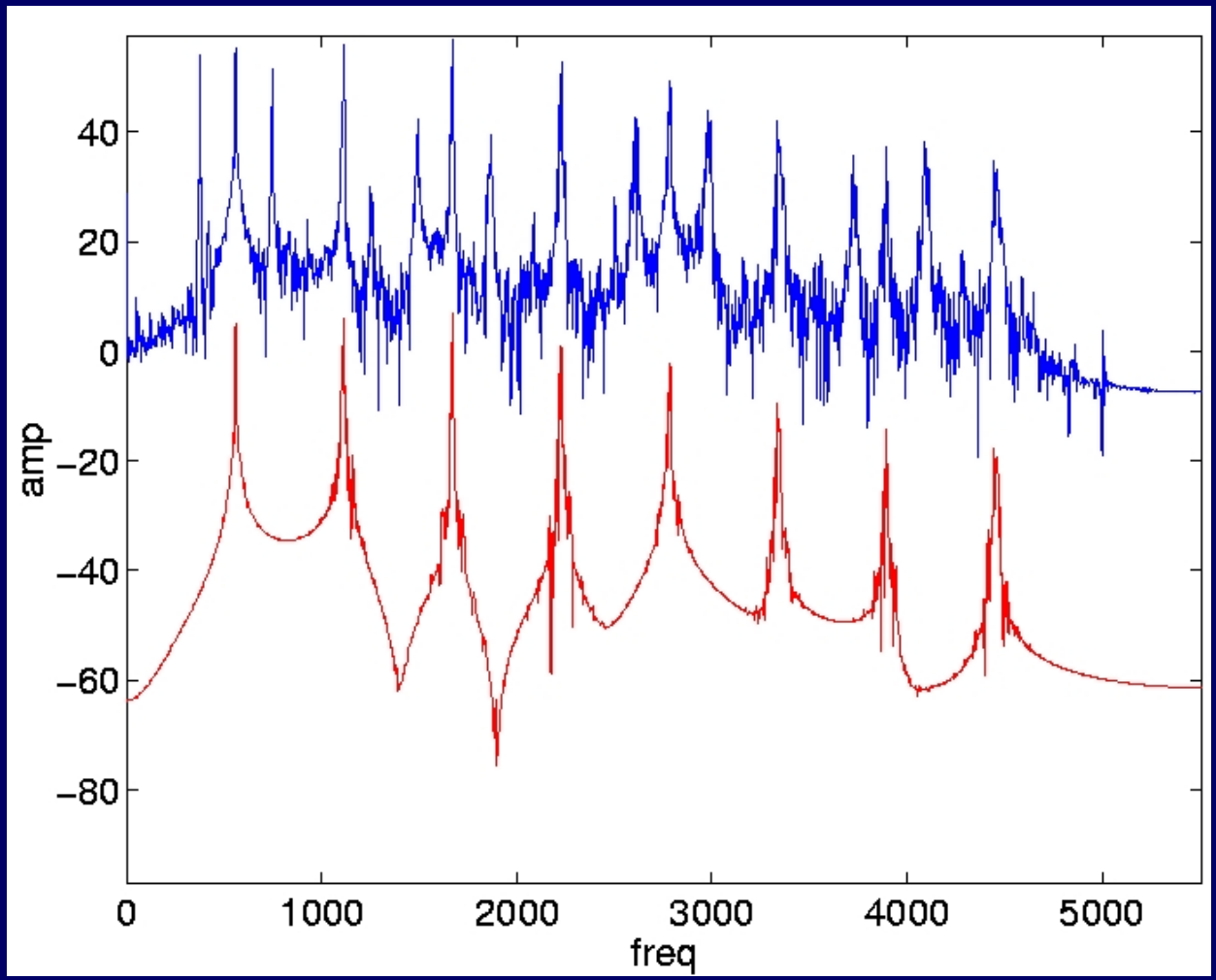
Results: Four note example

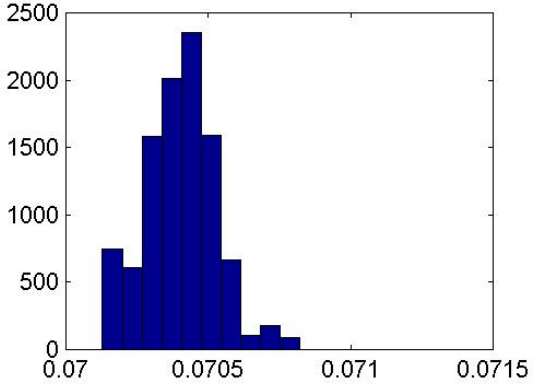
⇒ Convergence of the four frequencies

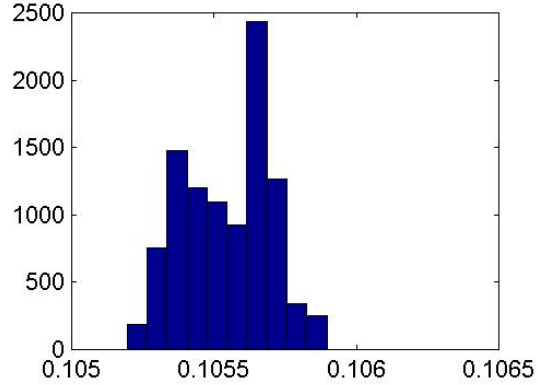


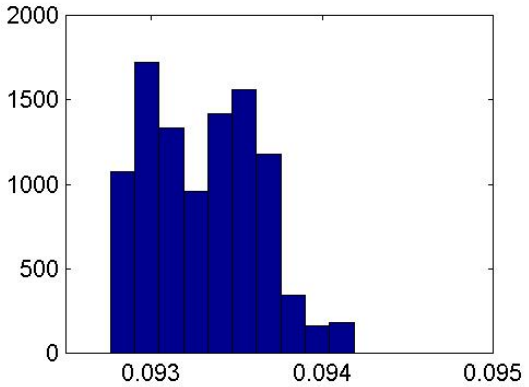
Results: 'Commit' example

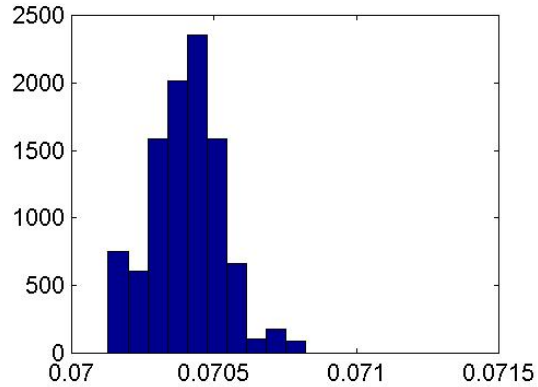
$\Rightarrow$ Posterior densities for four frequencies




2500
2000
1500
1000
500
0
0.07
0.0705
0.071
0.0715

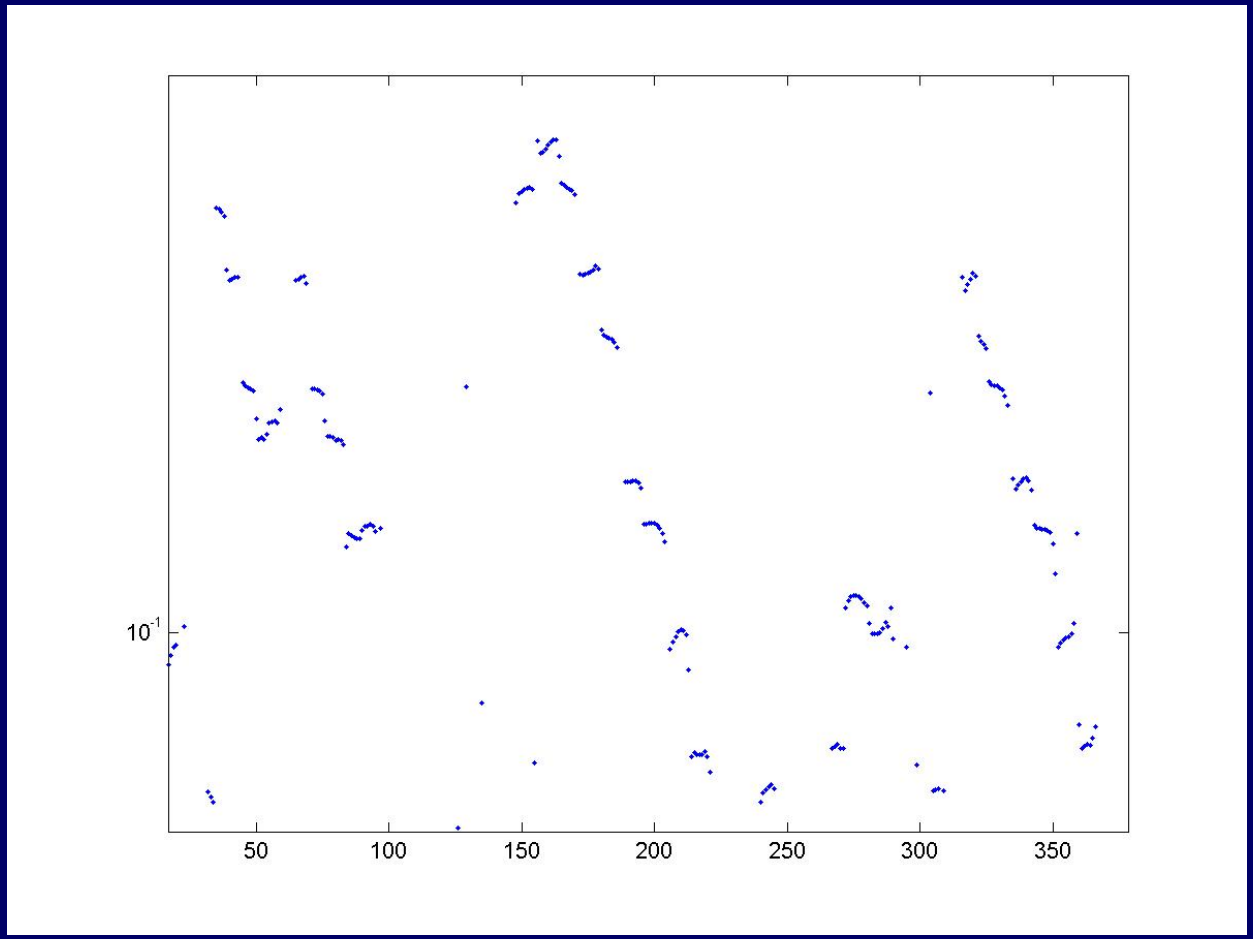

2500
2000
1500
1000
500
0
0.105
0.1055
0.106
0.1065


2000
1500
1000
500
0
0.093
0.094
0.095


2500
2000
1500
1000
500
0
0.07
0.0705
0.071
0.0715

Parker extract and source separation:

⇒ Attempt to perform source separation based on a monophonic model limited to saxophone frequency range



6: Discussion:

- Bayesian harmonic models can provide the fundamental building blocks for automatic transcription systems
- To get really good performance they need to be included in a larger hierarchical scheme that incorporates context/instrument specific information
- Computations are too slow at the moment for routine application - better MCMC algorithms, non-MCMC approximations?