Music Transcription by Poisson Point Processes and Sequential Markov chain Monte Carlo

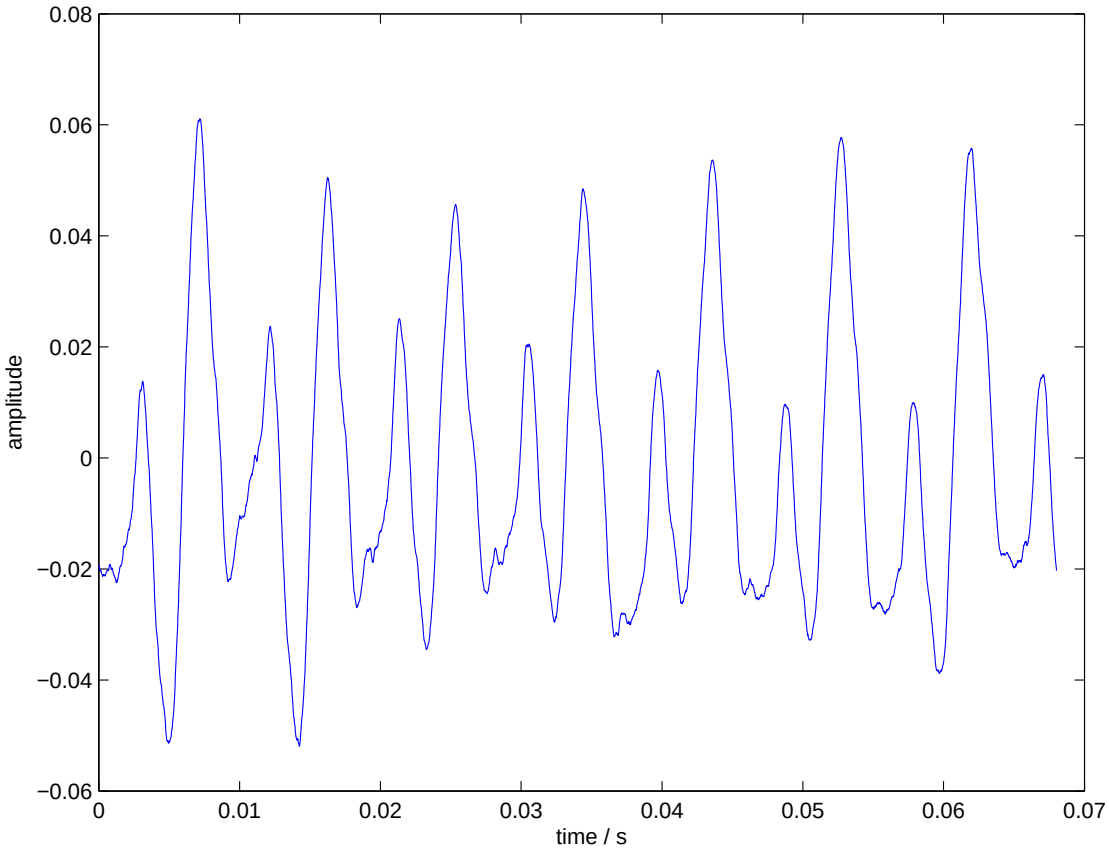
Pete Bunch and Simon Godsill

Cambridge University Engineering Department
Signal Processing & Communications Lab

26th May, 2011

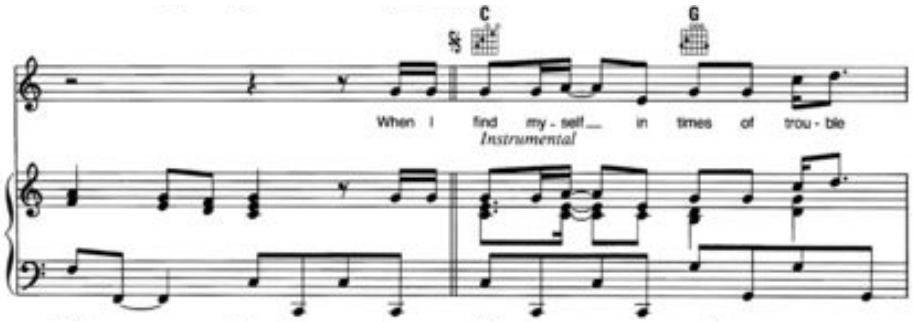
Introduction

Input:



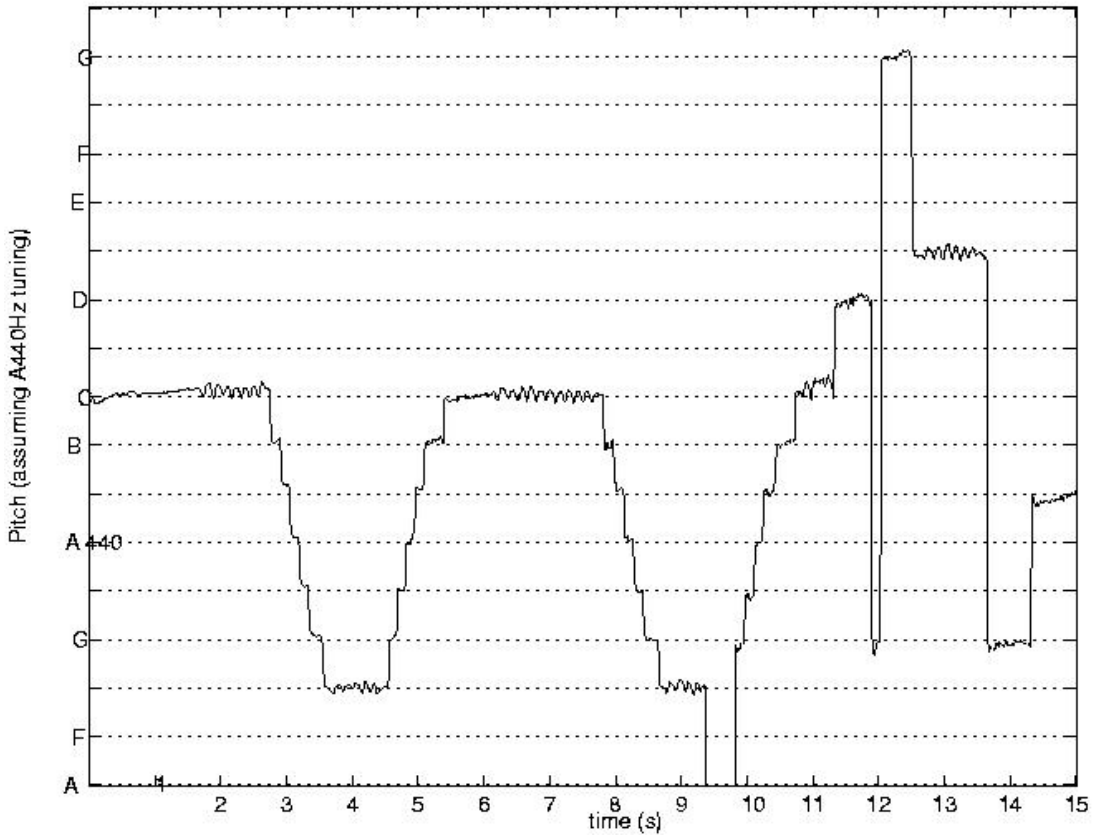
Introduction

Desired Output:



Introduction

Realistic Output:



Overview

- Sequential music transciption
- Poisson point process model for notes
- Modelling of Note evolution
- MCMC schemes
- Results

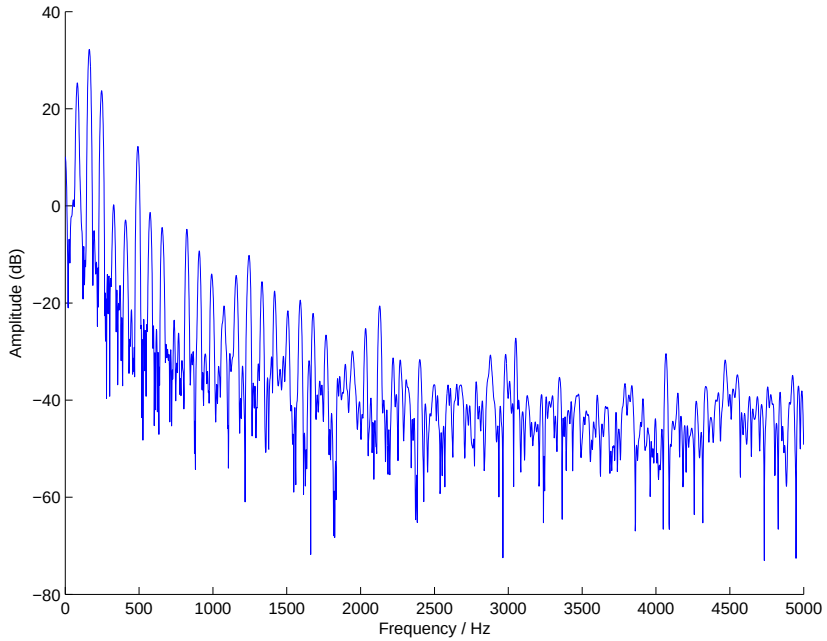
A framework for sequential music transription

- Divide audio data into frames
- Infer notes in each frame, $\boldsymbol{\theta}_t$, given data, $\mathbf{Y}_t$.

$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto P(\mathbf{Y}_t | \boldsymbol{\theta}_t) P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$

The spectrum of a musical note

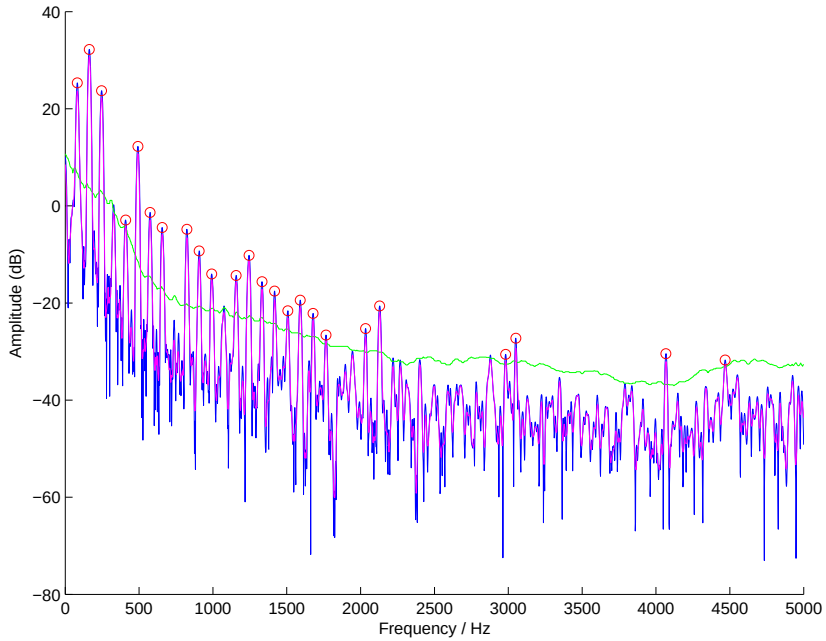
$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto \underbrace{P(\mathbf{Y}_t | \boldsymbol{\theta}_t)} P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$



- Short-time FFT
- Reduce frame "data" to a set of peak frequencies

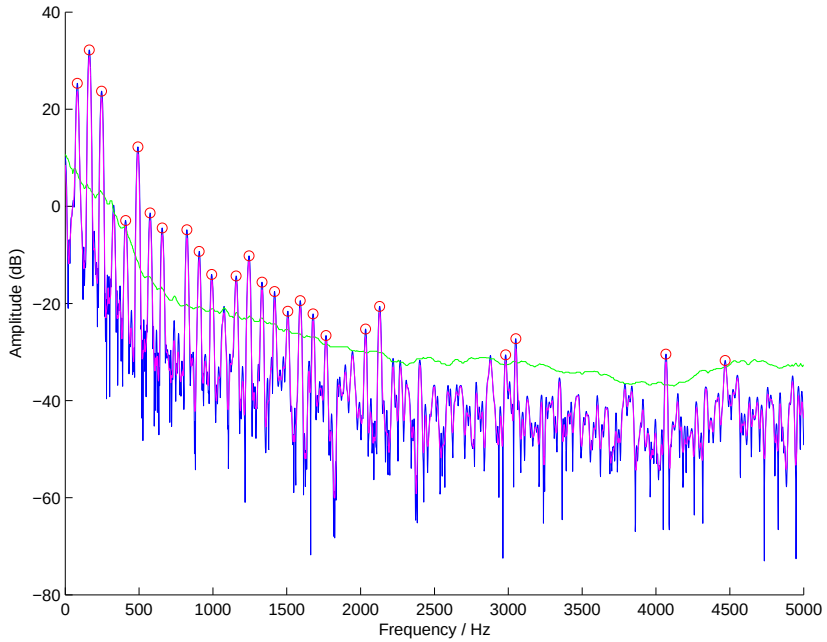
The spectrum of a musical note

$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto \underbrace{P(\mathbf{Y}_t | \boldsymbol{\theta}_t)} P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$



- Short-time FFT
- Reduce frame "data" to a set of peak frequencies

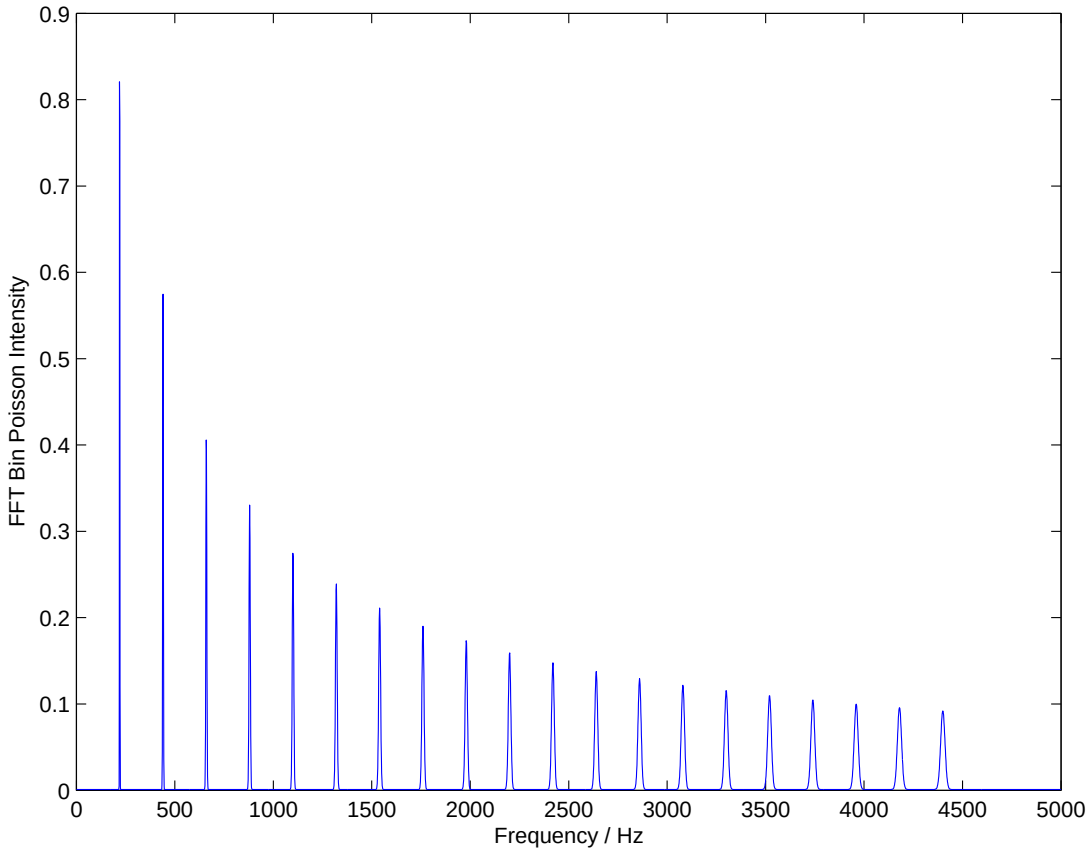
A model for frame likelihoods



- Each note may generate a peak at $n \times f_0$
- Some "clutter" peaks detected
- Data association problem arises
- Combinatorial complexity in number of notes/partials

The Poisson point process assumption

- For a given partial, peaks are generated by a Poisson Process
- Non-homogeneous with intensity peaked at expected frequency
- Intensities add for each harmonic and each note



The Poisson model

$$\lambda(f) = \sum_{j=1}^{N} \lambda_j(f) + \lambda_c \tag{1}$$

$$\lambda_j(f) = \sum_{h=1}^{H_j} \frac{A}{\sqrt{2\pi\sigma_{h,j}^2}} \exp\left\{ -\frac{(f - f_{h,j})^2}{2\sigma_{h,j}^2} \right\} \tag{2}$$

- $f_{h,j}$ is the frequency of the h^{th} partial of the j^{th} note
- Gaussian component at each expected partial frequency

P. Peeling, C. Li, and S. Godsill, "Poisson point process modeling for polyphonic music transcription," Journal of the Acoustical Society of America Express Letters, vol. 121, no. 4, pp. EL168–EL175, 2007.

The Poisson likelihood

- Integrate intensity over each FFT bin, k, to give expectation, μ_k.
- Let $y_k = 1$ if a peak is present, 0 otherwise

One bin:

$$P(y_k|\mu_k) = \begin{cases} e^{-\mu_k}, & y_k = 0 \\ 1 - e^{-\mu_k}, & y_k = 1 \end{cases} \tag{3}$$

Whole frame:

$$P(\mathbf{Y}|\mu) = \prod_{k=1}^{K} [y_k(1 - e^{-\mu_k}) + (1 - y_k)e^{-\mu_k}] \tag{4}$$

Inharmonicity

Partial frequencies do not occur exactly at integer multiples of the fundamental frequency.

$$f_h = f_0 h \sqrt{1 + Bh^2}$$

where

- f_h is the frequency of partial number h
- f_0 is the fundamental frequency of the note
- B is the inharmonicity parameter for the note (of the order 10^{-4})

S. Godsill and M. Davy, "Bayesian computational models for inharmonicity in musical instruments," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, 2005, pp. 283–286.

Parameters for inference

$\boldsymbol{\theta}$ is comprised of:

- N, the number of notes
- $f_{0,j}$, the fundamental frequency of each note
- H_j, the number of partials of each note
- B_j, the inharmonicity of each note

Dynamic note modeling

$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto P(\mathbf{Y}_t | \boldsymbol{\theta}_t) \underbrace{P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})} P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$

- Higher probability of notes remaining at the same frequency, rather than changing
- Higher probability of number of notes remaining the same
- Much more sophisticated models possible

MCMC-particles algorithm

$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto P(\mathbf{Y}_t | \boldsymbol{\theta}_t) P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$

- High-dimensional problem - low acceptance rates if we change all parameters at once
- Metropolis-Hastings. Two types of move: $\boldsymbol{\theta}_t$ and $\{\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t\}$

$\boldsymbol{\theta}_t$ moves:

- Metropolis-within-Gibbs - change one note at a time

S.K. Pang, S.J. Godsill, J. Li, and F. Septier, "Sequential inference for dynamically evolving groups of objects," in *Inference and Learning in Dynamic Models (To Appear)*, Barber, Cemgil, and Chiappa, Eds. CUP, 2011.

MCMC-particles algorithm

$$P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t}) \propto P(\mathbf{Y}_t | \boldsymbol{\theta}_t) P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$$

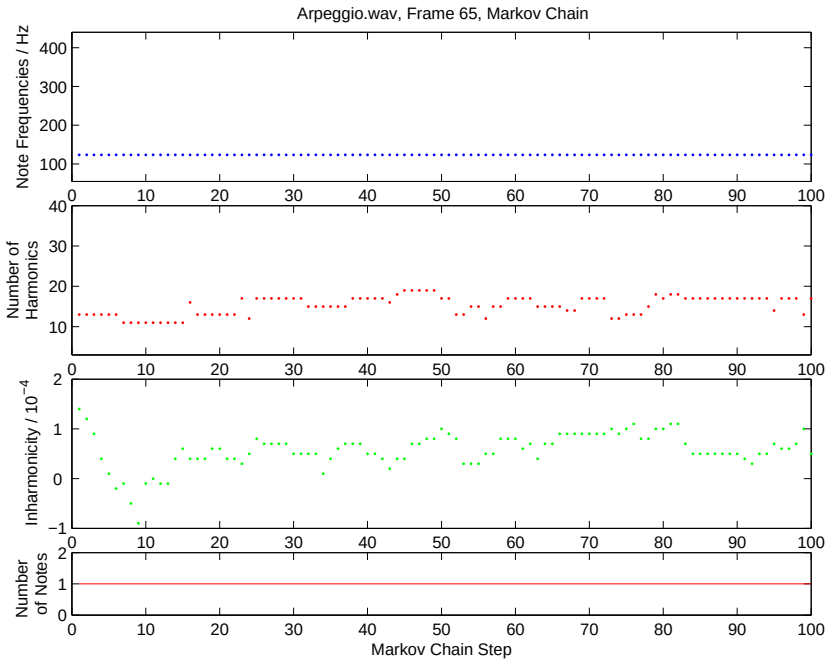
$\{\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t\}$ moves:

- We would like to sample from $P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1}) q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$
- Very low acceptance rates
- Collapse $P(\boldsymbol{\theta}_{t-1} | \mathbf{Y}_{1:t-1})$ into a univariate histogram.
- Approximate with a product of marginal distributions.

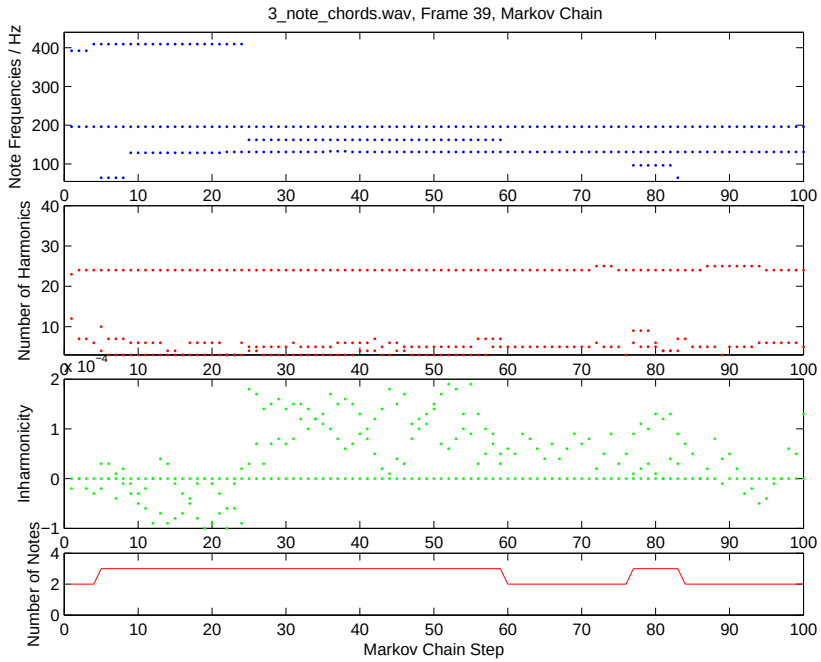
Reversible jump framework to estimate number of notes.

MCMC output

Markov chain states:

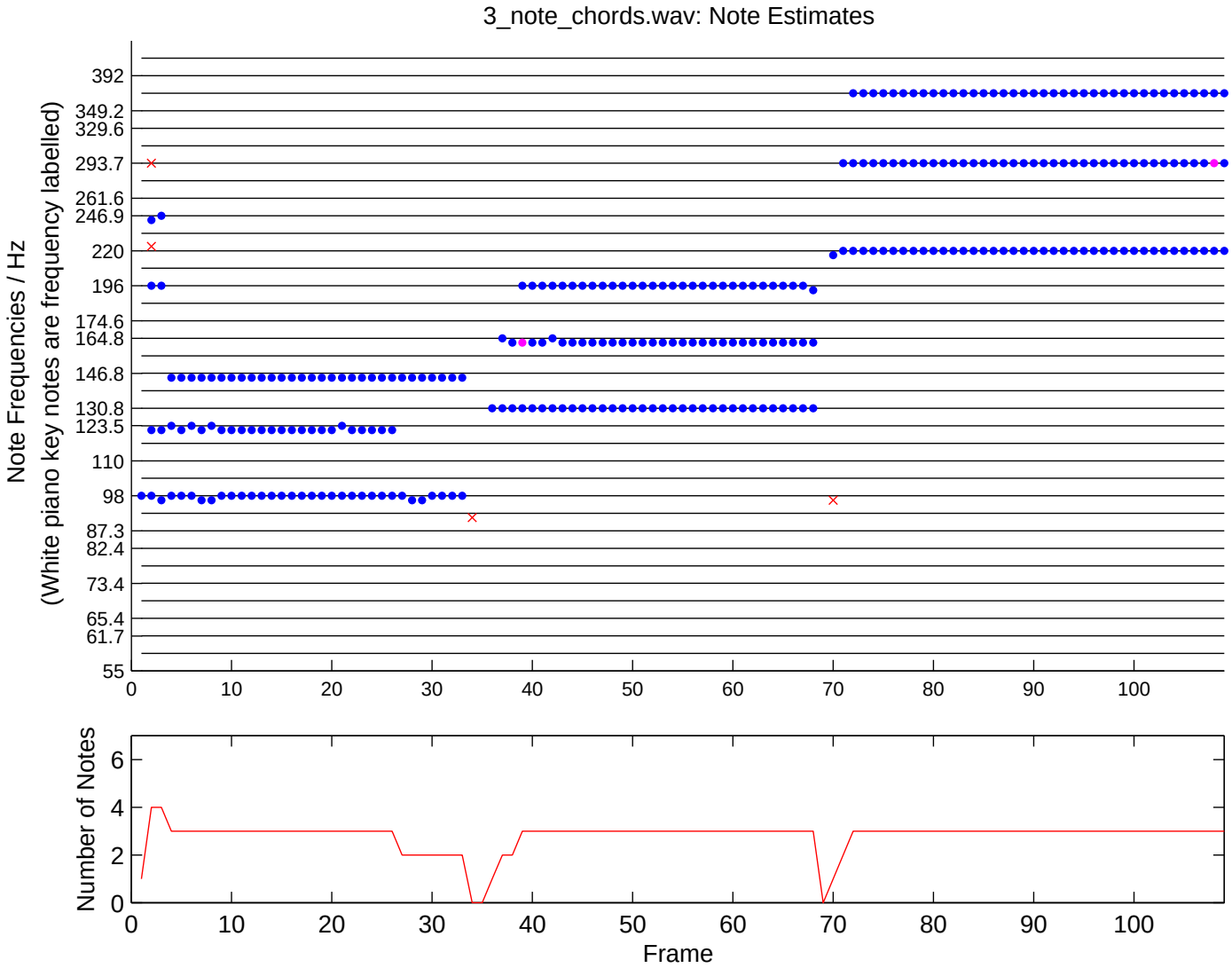


(a) Single Note

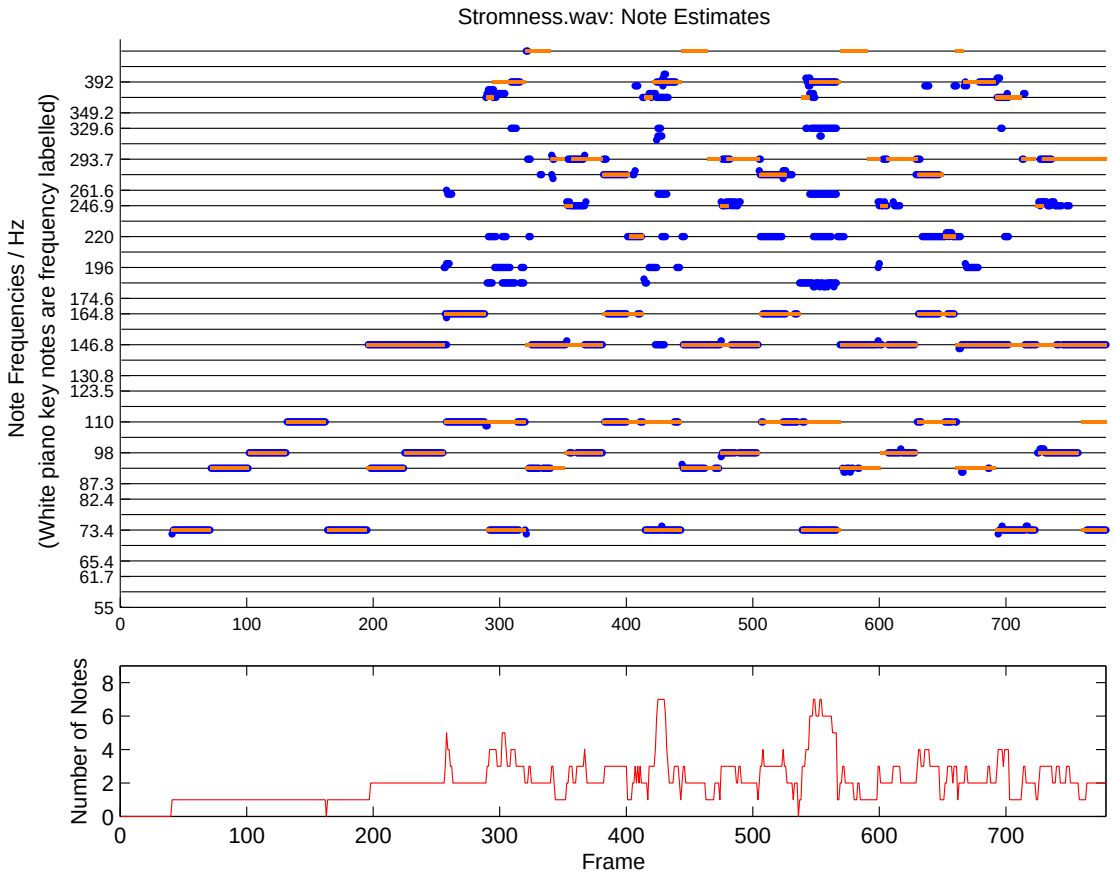


(b) Three Notes

Results I


3_note_chords.wav: Note Estimates
Note Frequencies / Hz
(White piano key notes are frequency labelled)
392
349.2
329.6
293.7
261.6
246.9
220
196
174.6
164.8
146.8
130.8
123.5
110
98
87.3
82.4
73.4
65.4
61.7
55
0
10
20
30
40
50
60
70
80
90
100
Number of Notes
6
4
2
0
Frame

Results II



Estimates in blue. Ground truth in orange.

Conclusions and Possible Extensions

- Poisson point process assumption leads to a simple likelihood model for frame spectra.
- Simple models for note evolution
- Sequential MCMC-particles algorithm allows inference for multiple notes including number of notes, fundamental frequencies, numbers of partials, and inharmonicities.
- Many extensions possible:
  - Peak amplitudes
  - Phase
  - More complex dynamic models
- Promising results from very simple models.

References

A.T. Cemgil, S.J. Godsill, P. Peeling, and N. Whiteley,

"Bayesian statistical methods for audio and music processing,"
in *Handbook of Applied Bayesian Analysis*, A. O'Hagan and M. West, Eds. Oxford University Press, 2010, (To Appear).

P. Peeling, C. Li, and S. Godsill,

"Poisson point process modeling for polyphonic music transcription,"
Journal of the Acoustical Society of America Express Letters, vol. 121, no. 4, pp. EL168–EL175, 2007.

S.K. Pang, S.J. Godsill, J. Li, and F. Septier,

"Sequential inference for dynamically evolving groups of objects,"
in *Inference and Learning in Dynamic Models (To Appear)*, Barber, Cemgil, and Chiappa, Eds. CUP, 2010.

S. Godsill and M. Davy,

"Bayesian computational models for inharmonicity in musical instruments,"
in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, 2005, pp. 283–286.

P.J. Green,

"Reversible jump markov chain monte carlo computation and bayesian model determination,"
Biometrika, vol. 82, no. 4, pp. 711, 1995.