Bayesian Methods to Perform High-Quality Audio Source Separation: Technical Milestone Report

Max Fryer, supervised by Prof. S. Godsill

January 19, 2023

0.1 Summary

When listening to classical music, skilled musicians can identify specific instruments from the fusion of sounds, analyse its key characteristics (including key signature, time signature etc.), and even transcribe it (write to sheet music). There has been extensive research on the processing of monophonic (single instrument) lines, but much less on polyphonic (multi instrument) recordings. Of these techniques fewer still involve using prior knowledge to apply Bayesian approaches, even while musicians (particularly in western music) are largely constrained by a set of strict musical rules and the physics of their instruments. In this project we tackle a small part of the polyphonic music modelling task, separating polyphonic recordings into separate monophonic recording using modern Bayesian methods, to allow for more extensive use of other techniques.

Inspired by Gabor-atoms we construct a model for our sound in a defined time-frequency grid. To ensure smooth, natural sounding transitions between weighted sinusoids, windowing functions are used at a length-scale that ensures flexibility enough to capture the rich complexity of musical tones while able to smooth out noise.
We begin by applying a maximum likelihood estimation algorithm to explore the effectiveness of our music model, before constructing a more complex polyphonic Bayesian model that more naturally allows the inclusion of meaningful, extensive priors, a technique devised originally by Davy, Godsill and Idier in their 2006 paper *"Bayesian analysis of polyphonic western tonal music"*.
Not only is a novel model suggested but a method for estimating parameters using Markov Chain Monte Carlo (MCMC) techniques especially suited to this problem.

0.2 Linear Gaussian Model & ML Estimation

We assume to begin with that the data $\mathbf{x}$ is generated from a parametric function $\mathbf{g}$ with some additive Gaussian noise term $\mathbf{e}$. Here our noise is assumed drawn from an independent, identically distributed source (i.i.d),

$$x_n = g_n(\theta, e_n)$$

Acoustic instruments produce complex tones that contain a root frequency and several harmonics (shown in Fig. 1). Many acoustic oscillators, such as a bowed violin produce overtones that are almost perfectly periodic (imperfection owing to the non-linear response of strings to stimulation).[1] For this reason we use a sinusoidal model for our signal,

$$x_n = \sum_{i=1}^{M} a_i sin(\omega_i n) + b_i cos(\omega_i n)$$

We choose an eleventh order model based on Fig. 1 and because harmonics of degree > 11 will exceed the Nyquist frequency (Sampling frequency unless otherwise stated is 44.1KHZ).
Our linear model expression is therefore,

$$\mathbf{x} = \mathbf{G}\theta + \mathbf{e}$$

with basis functions and weights given by,

$$\mathbf{G} = [\mathbf{c}(\omega_1) \quad \mathbf{s}(\omega_1) \quad \cdots \quad \mathbf{c}(\omega_{11}) \quad \mathbf{s}(\omega_{11})],$$

$$\theta = \begin{bmatrix} a_1 & b_1 & a_2 & b_2 & \cdots & a_{11} & b_{11} \end{bmatrix}^T$$

Beginning simply with a ML estimator we estimate the parameters from the data using no information *a priori* by maximising the likelihood,

$$\theta^{ML} = \underset{\theta}{\text{argmax}}[p(\mathbf{x}|\theta)]$$

We first experiment with a simple recording of a single piano note of known frequency (c_4, 261.6 hz), a snapshot of whose frequency spectrum can be seen in Fig. 1.

1. Take samples of length 500 samples ($\sim$ 10ms) around each 1000^{th} sample ($\sim$ 20ms).

2. Calculate maximum likelihood weights up to and including 11 degrees of harmonics. $\theta^{ML} = (G^T G)^{-1} G^T y$.

3. We then compute all samples using linear interpolation (or triangular weighting functions) before resynthisizing using our model.

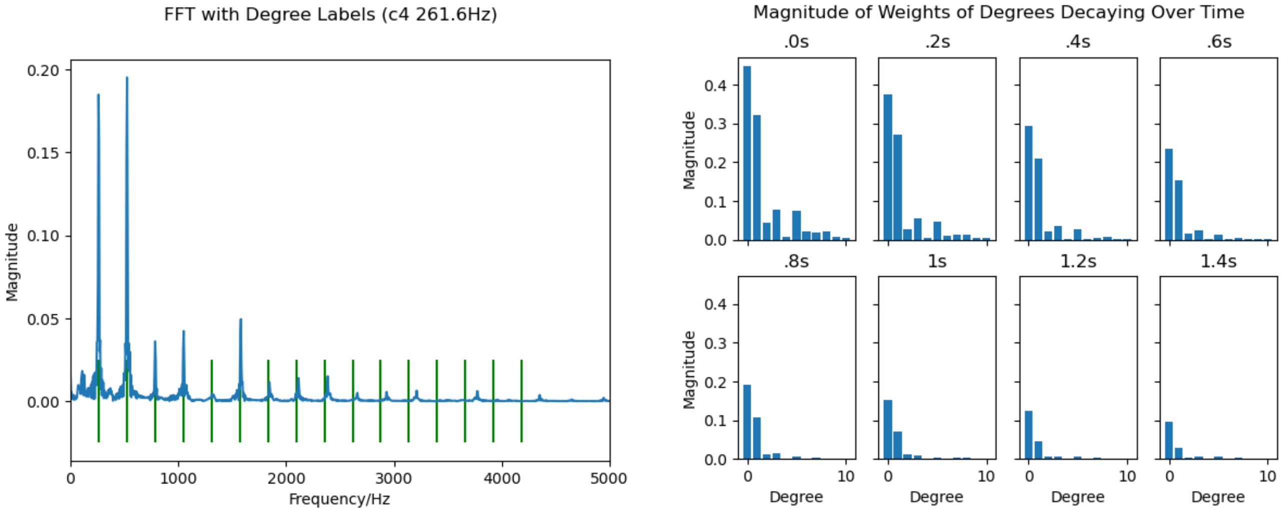

Figure 1: Snapshot of frequency domain of piano note (left), showing importance of harmonics (shown on axis), ML estimation of weights over time (right)

[1] Tones that are integer multiples of the root frequencies are termed “partial harmonics” or “harmonics” for short (though very rarely are they perfect harmonics)

0.3 Polyphonic Bayesian Harmonic Model

We next introduce a varient of the more flexible mathematical model of harmonic music originally developed by Godsill, Davy & Idier.[2] Our parameters are not truly random variables, we have prior knowledge that we ideally want to express probabilisticly before the data are observed.

Equipped with both the data and some knowledge about our parameters *a priori* we can state Bayes' Theorem of estimating posterior distribution,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

The Bayesian approach has a second advantage over ML estimation by returning a fully interpretable probability distribution (ML gives 'point estimate'), in theory everything we could want to know about the distribution.

As well as an more flexible probabilistic framework we implement a more robust mathematical model of harmonic music inspired by the Gabor atom-based approach. Here we model recording samples as a sum of K notes (fundamental frequencies), each with a number of harmonics, M_k all of which exist within I parameterized Hanning windows $\Phi[t]$ shown in Fig. 2,

$$y[t] = \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=0}^{I} \Phi[t - i\Delta_t] \Bigg\{ a_{k,m,i} cos(\frac{\omega_{k,m}}{\omega_s} t) + b_{k,m,i} sin(\frac{\omega_{k,m}}{\omega_s} t) \Bigg\} + \nu[t]$$

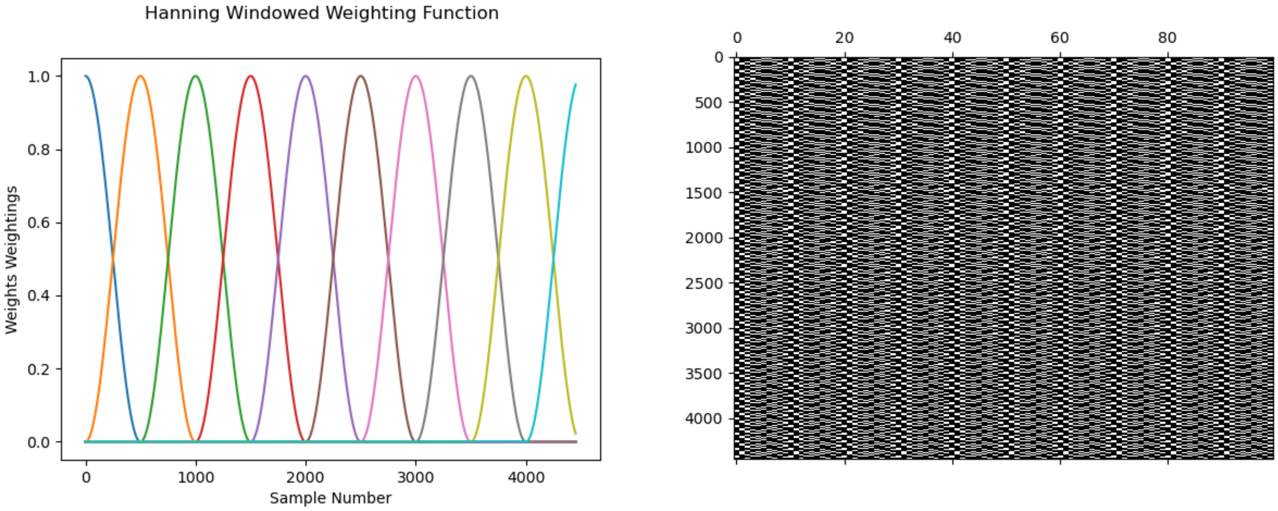

Figure 2: Hanning windows of length 1000 samples and 50% overlap (left), `spy(D)` showing cells whose value>0.5, visualization of 'G of G' matrix(right)

[2] Davy, Godsill, Idier (2006), *Bayesian analysis of polyphonic western tonal music*, The Journal of the Acoustical Society of America, 119, 2498

where parameterized envelope,$\Phi[t]$, is defined as a superposition of overlapping Hanning windowing,

$$\sum_{i=0}^{I} 0.5 - 0.5cos\left(\frac{2\pi n}{(i+1)M-1}\right),$$

$$0 \leq n \leq (i+1)M - 1.$$

Using previous notation, this can be thought of as a 'G of G matrices', which will now refer to as $\mathbf{D}$, multiplied by the Hanning window functions,

$$\mathbf{D} = \mathbf{G}_t^* = \Phi[t] \times [\mathbf{G}_1 \quad \mathbf{G}_2 \cdots \quad \mathbf{G}_K]$$

The unknown parameters in this model are the amplitudes, β, with length $2R(I+1)$, where $R = \sum_{k=1}^{K} M_k$ is the total number of partials. we can now compactly write our model

$$\mathbf{y} = \mathbf{D}\beta + \nu$$

Including the variance of the noise, σ_v^2 and the frequencies and number of partials the Polyphonic Bayesian Harmonic Model has total unknown parameters $R(2I+3)+2$.

0.4 Probabilistic Framework

Having explored our model we now fix it to a Bayesian probabilistic framework to allow useful inference from data.
Assuming again that noise is i.i.d we can define the likelihood function of the model parameters,

$$p(\mathbf{y}|\beta, \sigma_v^2, \omega, M, K) = (2\pi\sigma_v^2)^{-N/2} \exp\Bigg[$$

$$-\frac{1}{2\sigma_v^2}\|y - D\beta\|^2\Bigg]$$

[Remark: Currently in Progress] If we were to implement ML estimation now we would find a tendency towards solutions with too many partials and notes. To penalize over-fitting we incorporate priors to the parameters $\beta, \sigma_v^2 \&$ ω.

A sensible objective would be to use maximum *a posteriori* (MAP) to combine likelihood and priors into a posterior distribution

$$p(\beta, \sigma_v^2, \omega, M, K|\mathbf{y}) \propto$$

$$p(\mathbf{y}|\beta, \sigma_v^2, \omega, M, K)p(\beta, \sigma_v^2, \omega, M, K)$$

. The estimate of model parameters is given by,

$$(\hat{M}, \hat{K}) = \underset{(M,K)}{\text{argmax}} p(M, K|y)$$

, where,

$$p(M, K|y) = \int p(\beta, \sigma_v^2, \omega, M, K|\mathbf{y}).$$

However, as is pointed out in the Davy, Godsill and Idier (see above) this yields an over-simplified view-point since the ordering of individual notes in β is non unique (*label switching problem*).

With suitable values for model order parameters, the weight parameters, β, for a particular model could then be estimated using minimum mean squared error (MMSE),

$$\hat{\beta} = \int \beta p(\beta, \sigma_v^2, \omega|y, \hat{M}, \hat{K}) d\beta d\sigma_v^2 d\omega$$

.

Because parameter estimation requires calculating intractable integrals, we resort instead to numerical techniques, in particular Markov Chain Monte Carlo (MCMC). This allows us to perform Monte Carlo estimates of the unknown parameters using random samples of $(\tilde{\beta}^{(l)}, \tilde{\sigma_v^2}^{(l)}, \tilde{\omega}^{(l)}, \tilde{M}^{(l)}, \tilde{K}^{(l)})$ according to the joint distribution $p(\beta, \sigma_v^2, \omega, M, K|\mathbf{y})$.

0.5 Priors

The priors can be simplified into a product of more elementary priors, so that each term is more pleasant to understand and model,

$$p(\beta, \sigma_v^2, \omega, M, K) = p(\beta|\sigma_v^2, \omega, M, K)p(\omega|M, K)$$
$$p(M|K)p(K)p(\sigma_v^2)$$

(i) $p(\beta|\sigma_v^2, \omega, M, K)$ The prior is selected to be a zero-mean Gaussian with parameterised covariance matrix $(\sigma_v^2/\zeta^2)\mathbf{I}$, where ζ^2 can be thought of as a signal-to-noise ratio. We treat ζ^2 as another parameter to be estimated and give it an inverted gamma prior as shown in Fig. 3,

$$p(\zeta^2) = IG(\alpha_\zeta, \beta_\zeta) \propto \frac{e^{\beta_\zeta/\zeta^2}}{\zeta^{2(\alpha_\zeta+1)}}$$

(ii) $p(M|K)$ The prior for frequencies given the the root frequency and overtones. A critical component of accurate estimation is the number of partials for each note. Too few and not enough information is captured to accurately represent the signal and resynthisis may be difficult. Too many and we may include noise or even worse estimate the fundamental frequency as being a number of partials below the true root frequency. The solution suggested in *"Bayesian analysis of polyphonic western tonal music"*, implemented here is to select a Poisson distribution for each M_k,

$$p(M_k = m|\Lambda_k) = e^{-\Lambda_k}\frac{\Lambda_k^m}{m!}, \quad \text{for } k = 1, \dots, K$$

where the prior Λ_k is gamma distributed. This is equivalent to the single prior,

$$p(M_k = m) \propto \frac{\Gamma(m + \alpha_\Lambda}{\Gamma(\alpha_\Lambda m!}(\beta_\Lambda + 1)^{-m}$$

where $\Lambda(\dot{})$ is the Gamma function.

(iii) $p(\sigma_v^2)$ The noise prior is an inverted Gamma distribution

$$p(\sigma_v^2) \propto (\sigma_v^2)^{-\psi_0/2-1}e^{-2/\mu_0\sigma_v^2},$$

where ψ_0 and μ_0 are small.

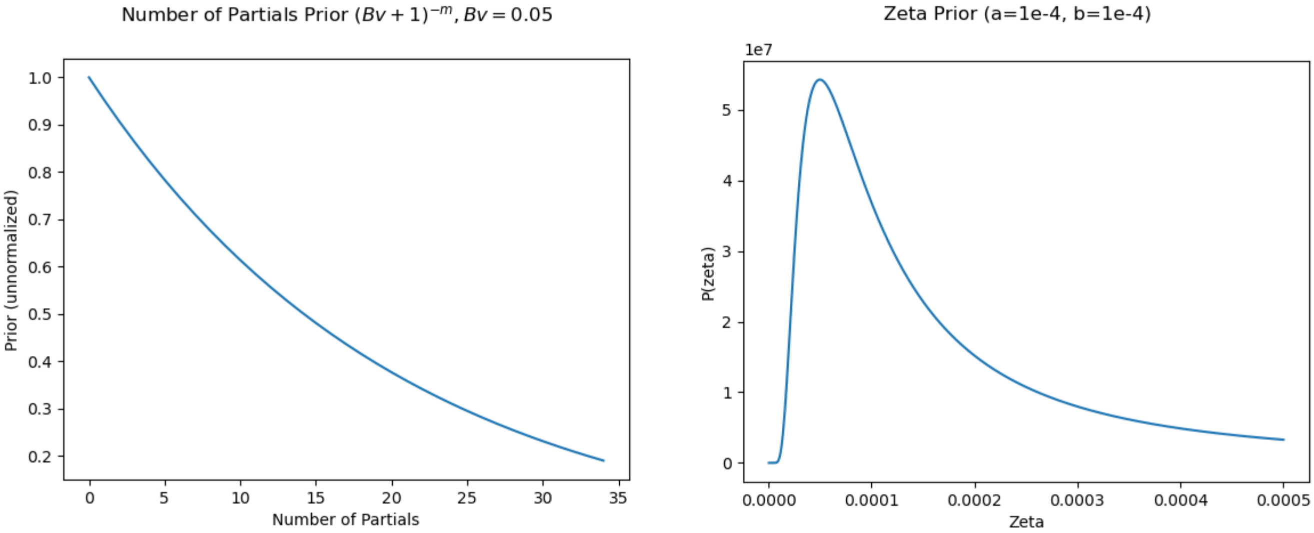

Figure 3: Prior on number of priors M_k for each Note, K (left), prior on zeta, which in turn defines diagonal covariance prior on weights vector, β (right)

0.6 Future Work

We now briefly outline the various directions of research which we hope to take over the course of Lent and Easter terms, in order of decreasing priority/interest. Naturally less technical in nature owing to it being a range of ideas yet to be implemented in detail.

0.6.1 Algorithm for parameter estimation (MCMC)

Monte Carlo techniques estimate intractable integrals of the following kind,

$$I(\theta) = E_{\theta|x}[f(\theta)] = \int_{\Theta} f(\theta)\pi(\theta|x)d\theta$$

via stochastic simulation, that is,

$$\hat{I}_N(\theta) = \frac{1}{N}\sum_{n=1}^{N} f(\theta^{(n)}) \longrightarrow E_{\theta|x}[f(\theta)]$$

.

With a means to obtain unbiased, consistent estimators of expectations with respect to distributions (even intractable ones), we require a method to obtain i.i.d samples from the target distribution $\pi(\theta|x)$. One such method studied in 4M24 is to employ Markov chains whose transition kernel returns an invariant distribution that is our desired distribution. We state the resulting MCMC method (full derivation can be found in lecture notes[3]), where $q(x,y)$ is our proposal density and the acceptance probability is given by

$$\alpha(x,y) = min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right)$$

```
for $j = 1 \Rightarrow N$ do
    Simulate y from $q(x^{(j)}, \cdot)$
    Simulate u from $U(0,1)$
    if u $\leq \alpha(x^{(j)}, y)$ then
        Set $x^{(j+1)} = y$
    else
        Set $x^{(j+1)} = x^{(j)}$
    end if
end for
Return $[x^{(1)}, x^{(2)}, \cdots, x^{(N)}]$
```

However, for large dimensional problems such as these the space of acceptable samples collapses. Methods developed for high-dimensional MCMC include Gaussian Random Walk Metropolis Hastings (GRW-MH) and pre-conditioned Crank Nicolson (pCN).

0.6.2 Include more priors

As it stands we are not using a prior on the distribution of frequencies given notes and partials, $p(\omega|M,K)$. The prior structure recommended in Davy, Godsill and Idier is,

$$p(\omega|M,K) = \prod_{k=1}^{K}\left[p(\omega_{k,1}|M_k)\prod_{m=1}^{M_k} p(\delta_{k,m})\right]$$

.

0.6.3 Piano harmonic model

A more accurate piano harmonic model is proposed by Fletcher and Rossing[4] where partial frequencies have frequencies

$$\omega_{k,m} = m\omega_{k,1}\sqrt{\frac{1+m^2B}{1+B}}$$

[3] M. Girolami, 4M24 Computational Statistics & Machine Learning, (Lec.5)

[4] N. Fletcher and T. Rossing, *The Physics of Musical Instruments*, 2^{nd} edition. (Springer, Berlin, 1998)