POLYPHONIC PITCH TRACKING USING JOINT BAYESIAN ESTIMATION OF MULTIPLE FRAME PARAMETERS

Paul J. Walmsley, Simon J. Godsill and Peter J. W. Rayner

Signal Processing Group
Cambridge University Engineering Department
Trumpington St., Cambridge, CB2 1PZ, UK.
{pjw42,sjg,pjwr}@eng.cam.ac.uk
http://www-sigproc.eng.cam.ac.uk/

ABSTRACT

We present a novel approach to pitch estimation and note detection in polyphonic audio signals. We pose the problem in a Bayesian probabilistic framework, which allows us to incorporate prior knowledge about the nature of musical data into the model. We exploit the high correlation between model parameters in adjacent frames of data by explicitly modelling the frequency variation over time using latent variables. Parameters are estimated jointly across a number of adjacent frames to increase the robustness of the estimation against transient events. Individual frames of data are modelled as the sum of harmonic sinusoids. Parameter estimation is performed using Markov chain Monte Carlo (MCMC) methods.

1. INTRODUCTION

Pitch estimation of polyphonic musical signals has received little attention compared to that of monophonic signals. Most monophonic techniques are not suited to polyphonic data, for instance cepstral and pitch-synchronous methods. There have been some very diverse approaches for polyphonic signals, however. Macleod [1] interpolates the discrete Fourier spectrum to obtain high resolution frequency estimates. Klapuri [2] detects prime number harmonics to resolve harmonically related notes in chords. Rossi *et al.* [3] search for spectral patterns from a database, exploiting the inharmonicity in piano strings. Martin's blackboard system [4] successively abstracts STFT frequency tracks into higher level constructs of partials, notes and chords.

We present a model-based approach to polyphonic pitch estimation where the estimation of the fundamental frequency[1] and other parameters is performed along with model order detection, *i.e.*, determination of the number of concurrently sounding notes, and the number of harmonics in each. We adopt a harmonic signal model, which is a natural choice given the nature of the sound production mechanisms of many musical instruments (*e.g.*, source-filter, resonant cavities, *etc.*). Our method has the advantage that detailed *a priori* knowledge of the characteristics of individual instrument is not required (unlike [2, 3]), and is applicable to a wide range of instruments.

Most pitch estimation and sinusoidal analysis methods consider data on a frame-by-frame basis, and then infer frequency tracks as a later step. Here we exploit the fact that in musical signals, the frequencies of interest are those with a significant duration (*i.e.*, longer than a single frame), and model this explicitly. This has the chief advantage that the incidence of spurious frequency estimates arising from transient events is greatly reduced.

2. HARMONIC MODEL

Data is segmented into frames $\mathbf{d}_i$ of length N, chosen to make the frame duration around 20ms, during which time we assume the data is stationary. The model is constructed as the sum of an unknown number of concurrently sounding *notes*, where the parameters of note q in frame i are: fundamental frequency ω_i^q, number of harmonics H_i^q and harmonic amplitudes $\mathbf{b}_i^q$.

A maximum limit of Q notes is imposed, but each note can be switched into or out of the model via a binary indicator variable γ_i^q, which is estimated along with the other parameters, hence the model order selection is implicitly carried out within the estimation process. Each note is expressed as a General Linear Model (GLM) [5] in terms of a harmonic basis matrix $\mathbf{G}_i^q$, the amplitudes $\mathbf{b}_i^q$, and an error term $\mathbf{e}_i$ which is assumed Gaussian, independent and identically distributed with variance $\sigma_{e_i}^2$,

$$\mathbf{d}_i = \sum_{q=1}^{Q} \gamma_i^q \mathbf{G}_i^q \mathbf{b}_i^q + \mathbf{e}_i \tag{1}$$

$$\mathbf{G}_i^q = [\, \mathbf{s}(\omega_i^q) \ \ldots \ \mathbf{s}(H_i^q \omega_i^q) \, \mathbf{c}(\omega_i^q) \ \ldots \ \mathbf{c}(H_i^q \omega_i^q)] \tag{2}$$

$$\mathbf{s}(\omega) = [\, \sin(\omega t_1) \ \sin(\omega t_2) \ldots \sin(\omega t_N) \,]^{\mathrm{t}} \tag{3}$$

$$\mathbf{c}(\omega) = [\, \cos(\omega t_1) \ \cos(\omega t_2) \ldots \cos(\omega t_N) \,]^{\mathrm{t}}. \tag{4}$$

Denoting the note parameters by $\Theta_i^q = \{\gamma_i^q, \omega_i^q, H_i^q, \mathbf{b}_i^q\}$, the likelihood for frame i is given by

$$p(\mathbf{d}_i | \{\Theta_i^q; q = 1 \ldots Q\}, \sigma_{e_i}^2) = \frac{1}{(2\pi\sigma_{e_i}^2)^{\frac{N}{2}}} \exp\left[-\frac{\|\mathbf{e}_i\|^2}{2\sigma_{e_i}^2} \right] \tag{5}$$

A least-squares or maximum likelihood method would seek to maximise eq. (5), but here we pose the model in a Bayesian framework which enables us to impart prior information into the model via *a priori* probability densities on the parameters, and which also provides a basis for probabilistic model selection.

[1] For the purposes of this paper, we define pitch as the fundamental frequency of a harmonic signal.

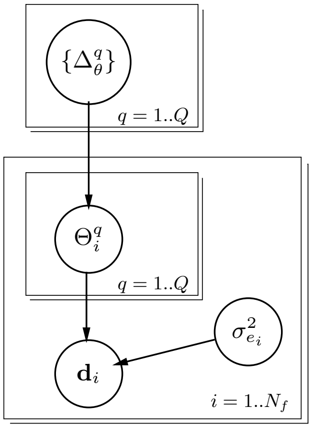

Figure 1: Graphical model showing how the parameters of multiple frames can be tied together through the hyperparameters $\{\Delta_\Theta^q\}$.

2.1. Graphical Model

Graphical models are a useful tool for investigating different model structures. They reflect the independence structure between model parameters and provide a simple means for evaluating the full conditional densities, which will be used in the numerical methods adopted for the parameter estimation. Figure 1 shows a graphical model representing eq. (1) for a set of N_f frames of data (termed a *block*). The *a priori* densities of the parameters of each note are dependent upon the *hyperparameters* $\{\Delta_\Theta^q\}$ which represent any prior knowledge we may have about the note parameters. Hyperparameters represent the constraints upon the underlying variation of the note parameters over the course of the block, and hence the prior distributions for the parameters are conditional upon the block hyperparameters, and provide a good local model fit.

2.2. Choice of priors

Figure 2 shows the detail of the graphical model for the parameters of note q in frame i. The block hyperparameters are $\{\nu^q, \sigma^2_{\omega^q}\}$ which denote respectively the pitch over the block and a measure of its spread, and Γ^q which represents whether the note is active (switched on) in this block. The arrows pointing to the note parameters signify dependence, such that the prior for ω_i^q is dependent on ν^q and $\sigma^2_{\omega^q}$, for example. The prior distributions used are:

$$p(\mathbf{b}_i^q) = \mathbb{I}_{[-\frac{B}{2},\frac{B}{2}]}(\mathbf{b}_i^q)/B^{2H_{\max}} \tag{6}$$

$$p(H_i^q = h) = \mathbb{I}_{[1,H_{\max}]}\, B^{-2(h-1)} \tag{7}$$

$$p(\gamma_i^q|\Gamma^q) = \begin{cases} 1-\alpha_\gamma, & \text{if } \gamma_i^q = \Gamma^q \\ \alpha_\gamma, & \text{otherwise} \end{cases} \tag{8}$$

$$p(\omega_i^q|\nu^q, \sigma^2_{\omega^q}) = \mathrm{LN}(\nu^q, \sigma^2_{\omega^q}) \tag{9}$$

$$p(\sigma^2_{e_i}) = \mathcal{IG}(\alpha_\sigma, \beta_\sigma) \tag{10}$$

The prior for the harmonic amplitudes $\mathbf{b}_i^q$ is a uniform distribution over the maximum dimension $2H_{\max}$, where $\mathbb{I}_Y(y)$ is the indicator function (unity if $y \in Y$, zero otherwise) and $\pm B/2$ is the maximum allowable range of any element of $\mathbf{b}_i^q$. The parameter H_i^q has a range of $1 \ldots H_{\max}$ and acts as a selector into $\mathbf{b}_i^q$, such that only the first H_i^q in-phase and quadrature (*i.e.*, sine and cosine) components are used. The prior is chosen to penalise higher model orders as the improvement in model fit achieved by increasing H_i^q must outweigh the cost of increasing the model dimensions. The prior for the switch variable γ_i^q has a Bernoulli form and encourages γ_i^q to follow the trend of Γ^q. The prior for the frequency is a lognormal density (LN) which has the useful property that the width of the distribution is proportional to frequency for a given $\sigma^2_{\omega^q}$. The error variance $\sigma^2_{e_i}$ is given a diffuse Inverse Gamma (IG) prior.

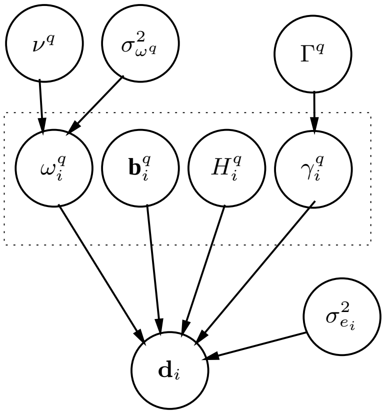

Figure 2: Structure of a harmonic model showing the dependencies of the parameters of one note for a single frame of data.

The hyperparameters are also given prior distributions:

$$p(\Gamma^q) = \begin{cases} \alpha_\Gamma^{N_f}, & \text{if } \Gamma^q = 1 \\ 1-\alpha_\Gamma^{N_f}, & \text{otherwise} \end{cases} \tag{11}$$

$$p(\nu^q) \propto \begin{cases} (2g\,\nu^q_{\text{prev}})^{-1}, & \text{if } |\nu^q - \nu^q_{\text{prev}}| < g\,\nu^q_{\text{prev}} \\ \beta, & \text{otherwise} \end{cases} \tag{12}$$

The latent switch variable Γ^q is given a Bernoulli prior which reflects the prior probability that a note should be switched on, where $0 < \alpha_\Gamma < 1$. The latent frequency variable ν^q is given a prior which has high probability within a small fraction g of the frequency in the previous block ν^q_{prev}, and a very low probability β outside that range. This prior creates a dependence between successive blocks, to encourage continuous frequency tracks and a value of $g = 0.05$ is used to make the width of $p(\nu^q)$ plus or minus one semitone. The value of the hyperparameter $\sigma^2_{\omega^q}$ should be specified *a priori* as it provides a constraint on the deviation from ν^q of the frequency of a single note over the block.

3. PARAMETER ESTIMATION

The joint posterior distribution for all parameters is obtained via Bayes' theorem as

$$p(\{\Theta_i^q\},\{\Delta_\theta^q\},\{\sigma_{e_i}^2\}|\{\mathbf{d}_i\}) \propto p(\{\Theta_i^q\},\{\Delta_\theta^q\},\{\sigma_{e_i}^2\}) \times \prod_{i=1}^{N_f} p(\mathbf{d}_i|\{\Theta_i^q\},\sigma_{e_i}^2). \qquad (13)$$

Our aim is to produce optimal parameter and hyperparameter estimates by locating the regions of highest probability of eq. (13), but as it is intractable to optimise this expression analytically we employ Markov chain Monte Carlo (MCMC) methods to simulate a Markov chain with eq. (13) as its stationary distribution [6]. A stream of dependent samples from the posterior are generated and the final state of the Markov chain is used for the parameter estimates.[2]

The successive states of the Markov chain are generated by the Metropolis-Hastings (M-H) algorithm [6]. A transition kernel $T(\Theta^k,\Theta^*)$ generates a proposal state Θ^* from the current state Θ^k which is then accepted with probability $\min(1, Q(\Theta^k,\Theta^*))$ where $Q(\cdot,\cdot)$ is the M-H acceptance function:

$$Q(\Theta^k,\Theta^*) = \frac{p(\Theta^*|\mathbf{d})}{p(\Theta^k|\mathbf{d})}\frac{T(\Theta^*,\Theta^k)}{T(\Theta^k,\Theta^*)}. \qquad (14)$$

$T(\Theta^k,\Theta^*)$ here is also used to represent the transition probability of proposing the move from Θ^k to Θ^*. If the proposal only affects the subset $\theta \subset \Theta$, then $Q(\cdot)$ simplifies to a function of the full conditional for θ,

$$Q(\theta^k,\theta^*) = \frac{p(\theta^*|\mathbf{d},\Theta^k_{-\{\theta\}})}{p(\theta^k|\mathbf{d},\Theta^k_{-\{\theta\}})}\frac{T(\theta^*,\theta^k)}{T(\theta^k,\theta^*)}. \qquad (15)$$

where $\Theta_{-\{\theta\}}$ denotes all parameters except those in θ. The benefit of sampling from the full conditionals is that the distributions are simpler and cheaper to calculate than the joint posterior. The simulation method consists of iteratively sampling for each note, and so great efficiency savings can be made by defining a *residual* $\mathbf{r}_i^q$ and expressing the error as a GLM in terms of the residual and the note parameters:

$$\mathbf{r}_i^q = \mathbf{d}_i^q - \sum_{q'\neq q}\gamma_i^{q'}\mathbf{G}_i^{q'}\mathbf{b}_i^{q'} \qquad (16)$$

$$\mathbf{e}_i^q = \mathbf{r}_i^q - \gamma_i^q\mathbf{G}_i^q\mathbf{b}_i^q. \qquad (17)$$

3.1. Choice of transition kernels

In this algorithm we apply global and local transition kernels: global kernels propose a state space move for hyperparameters and/or note parameters for a particular note q but across all frames $i = 1\ldots N_f$, whereas local kernels propose a move only in a single frame. The estimation algorithm is a two stage process, where the first stage is a stochastic scheme composed of global moves to steer the Markov chain into high probability regions, and the final stage is made up of local moves to obtain more accurate parameter estimates. Given good starting values (*e.g.,* initialised with the results of the previous block), convergence is rapid, and typically a total of 50–100 iterations in total is adequate.

The selection of transition kernels for harmonic models has been previously reported in [7]. Andrieu and Doucet [8] also describe a number of suitable kernels for estimation of multiple sinusoids. We employ a few different types of global transition kernels as described in the following subsection. In each instance, the M-H acceptance function is readily calculated from eq. (15). The local kernels are simply random perturbations about the current value and aren't described here in further detail.

3.1.1. Global kernels

• *Independence sampler.* The most important of the kernels in this algorithm is an independence sampling step that efficiently creates parameter proposals from a distribution which has its modes in similar locations to the posterior. A joint move is proposed for the following parameters of note q: $\{\nu^q, \{\omega_i^q, \mathbf{b}_i^q, H_i^q\}_i\}$.[3] Defining the order P harmonic transform [7] of a signal $\mathbf{x}$, $\mathcal{H}_P(\mathbf{x}, l)$ as

$$X_p[l] = \sum_{n=0}^{N_{\text{fft}}-1} x[n]\, e^{-\frac{j2\pi pln}{N_{\text{fft}}}} \qquad (18)$$

$$\mathcal{H}_P(\mathbf{x}, l) = \sum_{p=1}^{P} X_p^*[l]\, X_p[l], \qquad (19)$$

where l is the frequency bin number ($l = 1\ldots L$, $L = \lfloor N_{\text{fft}}/P \rfloor$), we define a proposal distribution $q(\omega^{q*}|H^{q*})$ to be

$$q(\omega^{q*}|H^*) \propto \frac{1}{N_f}\sum_{i=1}^{N_f}\mathcal{H}_{H^*}(\mathbf{r}_i^q, \left\lfloor \omega^{q^*}/\Delta\omega \right\rfloor). \qquad (20)$$

A value H^* is sampled from a discrete distribution $q(H^*)$ which has a peak roughly around $H^* = 4$, and using this value we calculate the harmonic transform of the residual $\mathbf{r}_i^q$ for all frames. The modes of this distribution are the fundamental frequencies corresponding to notes with significant energy in their first few harmonics, which is typical of musical notes. The proposal for the harmonic amplitudes is calculated for each $\{\omega_i^q, H_i^q\}_i$ pair from the least-squares projection: $\mathbf{b}_i^{q*} = (\mathbf{G}^{*\,t}\mathbf{G}^*)^{-1}\mathbf{G}^{*\,t}\mathbf{r}_i^q$.

• *Multiple step.* This is an efficient method of overcoming octave errors and some other problems which arise due to the non-uniqueness of the harmonic representation, in which the most compact representation is to be favoured. A joint move for $\{\nu^q, \{\omega_i^q, \mathbf{b}_i^q, H_i^q\}_i\}$ is proposed by sampling $u \sim \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{2}, 2, 3\}$ and setting

$$\nu^{q^*} = u\,\nu^{q^k} \quad \omega_i^{q^*} = \nu^{q^*} \quad H_i^{q^*} = \left\lceil H^{q^*}/u \right\rceil \qquad (21)$$

with $\mathbf{b}_i^{q*}$ chosen as in the independence step. This kernel traverses harmonically related modes of the posterior distribution.

• *Perturbation step*. Perturbations are applied to $\{H_i^q\}_i$ and to $\{\nu^q, \{\omega_i^q\}_i\}$ which constitute a random walk with a small variance about the current current parameter values.

• *Switch step.* A note is switched on or off across the entire block by a joint proposal $\{\Gamma^q, \{\gamma_i^q\}_i\}$, setting $\Gamma^{q^*} = 1 - \Gamma^{q^k}$ and $\gamma_i^{q^*} = \Gamma^{q^*}$. This can also be incorporated into the independence sampling step described above.

[2] Due to the very sharp multimodal posterior distribution, generally moves are only made to states with a higher posterior probability.

[3] Introducing the notation $\{\phi_i\}_i \equiv \{\phi_i; i = 1\ldots N_f\}$.

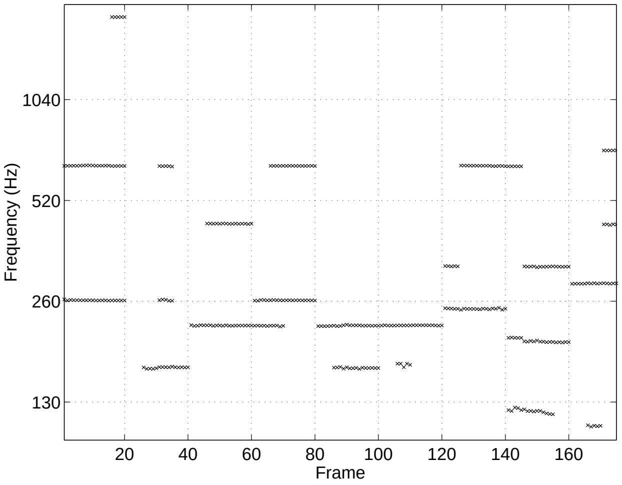

Figure 3: Log scale fundamental frequency tracks for a short extract of synthesised piano music. The parameters plotted are $\{\omega_i^q\}$, taken from the final state of the Markov chain. The frame length was 1000 samples ($F_s = 44.1$kHz), the block length was 5 frames, and a $Q = 5$ model was applied. Notes can clearly be distinguished in the 'piano roll' format. A threshold was applied so that only notes with an energy within 30dB of the signal energy are plotted.

4. RESULTS

A 'piano roll' plot of the fundamental frequencies ω_i^q from an extract of synthesised piano music are shown in figure 3. Smooth frequency tracks are obtained with very few outliers due to transients. There are some errors due to octave and perfect fifth intervals however, such as frames 15–20, which can be difficult to resolve when notes in the same chord are playing at the same time, since they will share a large number of harmonics. This is a common problem in harmonic models, since the representation can be ambiguous for chords, but major chords and octave intervals can often be detected successfully with this method, as can be seen in frames 45–60. This is possible since notes with fewer harmonics are favoured, which reduces the occurence of *harmonic roots* where the detected frequency is very low and the number of harmonics are high, and most harmonics have near zero amplitudes.

In addition to the fundamental frequency, we have estimates of the other model parameters, so the method can be applicable both to music transcription and signal separation. Both applications however require a higher level of modelling to determine the source of each frequency track. Musical instrument recognition relies on more than the steady-state harmonic amplitudes — the harmonic variation over time, and particularly the attack phase are very important characteristics [9]. A higher level model which accounts for the time and frequency domain variation of notes from different instruments and combines the currently disparate steps of signal modelling and musical context integration is clearly required, and is to be the subject of future work.

5. CONCLUSIONS

We have shown the effectiveness of a multiple frame approach to polyphonic pitch estimation which estimates harmonic model parameters jointly over a number of frames of data.

The algorithm is more robust to transient disturbances (*e.g.*, note attacks) and so fewer spurious frequency candidates are produced. The detection and estimation task is posed in a Bayesian framework, and the parameter full conditional densities have been obtained from graphical models to yield an efficient MCMC simulation scheme.

Work supported by the Engineering and Physical Sciences Research Council.

6. REFERENCES

[1] M. D. Macleod, "Fast nearly-ML estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 46, no. 1, pp. 141–148, January 1998.

[2] A. Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," in *Proc. EUSIPCO*, 1998.

[3] L. Rossi, G. Girolami, and M. Leca, "Identification of polyphonic piano signals," *Acustica*, vol. 83, pp. 1077–1084, 1997.

[4] K. D. Martin, "A blackboard system for automatic transcription of simple polyphonic music," Tech. Rep. No. 385, MIT Media Lab. perceptual computing section, July 1996. Available on-line[4].

[5] J. J. K. Ó Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer-Verlag, 1996.

[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., *Markov chain Monte Carlo in practice*, Chapman and Hall, 1996.

[7] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Multidimensional optimisation of harmonic signals," in *Proc. EUSIPCO*, 1998. Available on-line[5].

[8] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. on Signal Processing*. To appear.

[9] K. D. Martin, "Toward automatic sound source recognition: Identifying musical instruments," Presented at the NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy, 1-12 July 1998. Available on-line[6].

[4] ftp://sound.media.mit.edu/pub/Papers/kdm-TR385.ps.gz

[5] http://www-sigproc.eng.cam.ac.uk/~pjw42/ftp/eus98ps.zip

[6] ftp://sound.media.mit.edu/pub/Papers/kdm-comhear98.pdf