MULTIDIMENSIONAL OPTIMISATION OF HARMONIC SIGNALS

Paul J. Walmsley, Simon J. Godsill, Peter J. W. Rayner*
Signal Processing Group
Department of Engineering, University of Cambridge,
Cambridge CB2 1PZ, UK
pjw42@eng.cam.ac.uk, sjg@eng.cam.ac.uk, pjwr@eng.cam.ac.uk

ABSTRACT

Harmonic models are a common class of sinusoidal models which are of great interest in speech and musical analysis. In this paper we present a method for estimating the parameters of an unknown number of musical notes, each with an unknown number of harmonics. We pose the estimation task in a Bayesian framework which allows for the specification of (possibly subjective) *a priori* knowledge of the model parameters. We use indicator variables to represent implicitly the model order and employ a Metropolis-Hastings algorithm to produce approximate maximum *a posteriori* parameter estimates. A novel choice of transition kernels is presented to explore the parameter space, exploiting the structure of the posterior distribution.

1 INTRODUCTION

Sinusoidal models are popular in analysis of musical and speech signals due to considerations of the physical basis and periodic nature of voiced speech and of many musical instruments [5, 9, 11]. The signal is modelled as a series of frames, with the parameters regarded constant over the duration of each frame. We impose harmonic constraints upon the model, such that all frequencies are integer multiples of a fundamental, we hope to produce a good model fit whilst reducing the number of parameters to estimate [4, 6]. We describe each set of related harmonics as a *note*. Each note has a fundamental frequency (or pitch) ω_q with H_q harmonics, and is generated by a harmonic basis matrix $\mathbf{G}_q$,

$$\begin{aligned}\mathbf{G}_q &= [\mathbf{s}(\omega_q)\ldots\mathbf{s}(H_q\omega_q)\,\mathbf{c}(\omega_q)\ldots\mathbf{c}(H_q\omega_q)]\\ \mathbf{s}(\omega) &= [\sin(\omega t_1)\,\sin(\omega t_2)\ldots\sin(\omega t_N)]^{\mathrm{T}}\\ \mathbf{c}(\omega) &= [\cos(\omega t_1)\,\cos(\omega t_2)\ldots\cos(\omega t_N)]^{\mathrm{T}}\end{aligned}$$

and a vector of amplitudes $\mathbf{b}_q$. A frame of the signal, $\mathbf{d}$ (of length N), is modelled as the sum of up to Q notes, each of which is represented as a general linear model [2, 10], with each note turned on or off by a binary indicator variable γ_q. This approach means that the model order selection task is implicit and is performed jointly with the parameter estimation, as in reversible jump methods (*e.g.*, see [1]), but the effective size of the parameter space is fixed (by the choice of the maximum number of notes Q) such that changes in the number of notes and harmonics present do not involve a change of dimensions. The error term is assumed Gaussian, $\mathbf{e}\sim N(\mathbf{0},\sigma^2\mathbf{I}_N)$, and the model becomes

$$\mathbf{d}=\sum_{q=1}^{Q}\gamma_q\mathbf{G}_q\mathbf{b}_q+\mathbf{e}.$$

Rewriting the set of parameters of each note as $\Phi_q=\{\omega_q,H_q,\gamma_q,\mathbf{b}_q\}$ and the whole parameter set as $\Phi=\{\Phi_1,\Phi_2,\ldots,\Phi_Q,\sigma\}$, the likelihood is given by

$$p(\mathbf{d}|\Phi)=(2\pi\sigma^2)^{-\frac{N}{2}}\exp\left[\frac{-\left\|\mathbf{d}-\sum_{q=1}^{Q}\gamma_q\mathbf{G}_q\mathbf{b}_q\right\|^2}{2\sigma^2}\right].$$

2 BAYESIAN FORMULATION

For the structure of the *a priori* distributions of each note we write

$$\begin{aligned}p(\Phi_q) &= p(\omega_q,H_q,\gamma_q,\mathbf{b}_q)\\ &= p(\omega_q|\gamma_q,H_q)\,p(\gamma_q)\,p(\mathbf{b}_q|H_q)\,p(H_q).\end{aligned}$$

The prior for the amplitude parameters is chosen to be constant for values in the region $|\mathbf{b}_q[i]|<B/2\ (\forall i=1\ldots2H_q)$, and zero outside it, in order not to influence the inference procedure. The density in this region is $p_B(\mathbf{b}_q|H_q)=1/B^{2H_q}$.

The choice of the other priors is left general, but we can use a Bernoulli distribution for γ_q and a uniform distribution for H_q in the absence of other prior information. The frequency prior $p(\omega_q|\gamma_q,H_q)$ can be chosen to favour the continuity of frequency tracks between frames of data. A Jeffreys' prior is used for the scale parameter, $p(\sigma)=R_\sigma/\sigma$, chosen for its uninformative nature [2] and ease of marginalisation.

From Bayes theorem we obtain the posterior distribution

$$p(\Phi|\mathbf{d})=\frac{p(\mathbf{d}|\Phi)p(\Phi)}{p(\mathbf{d})}$$

*Work supported by the Engineering and Physical Sciences Research Council.

where $\boldsymbol{p}(\mathbf{d})$ is treated as a constant to give

$$p(\Phi|\mathbf{d}) \propto \prod_{q=1}^{Q} \left(\frac{P_q}{B^{-2H_q}}\right) \frac{R_\sigma (2\pi\sigma^2)^{-\frac{N}{2}}}{\sigma} \times \exp\left[\frac{-\left\|\mathbf{d} - \sum_{q=1}^{Q} \gamma_q \mathbf{G}_q \mathbf{b}_q\right\|^2}{2\sigma^2}\right] \quad (1)$$

$$P_q = p(\omega_q|\gamma_q, H_q)\, p(\gamma_q)\, p(H_q).$$

This expression is difficult to maximise analytically, and so a stochastic optimisation technique is employed. We apply a local Metropolis-Hastings algorithm [3] to sample for the parameters of one note whilst keeping the others constant (see [4] for a deterministic approach.) For this we sample from the full conditional posterior distributions for each note. Defining the q-th note residual as

$$\mathbf{r}_q = \mathbf{d} - \sum_{\substack{i=1 \\ i\neq q}}^{Q} \gamma_i \mathbf{G}_i \mathbf{b}_i,$$

the conditional posterior can be written as

$$p(\Phi_q, \sigma|\Phi_{-(q)}, \mathbf{d}) \propto \frac{P_q}{B^{-2H_q}} \frac{R_\sigma (2\pi\sigma^2)^{-\frac{N}{2}}}{\sigma} \times \exp\left[\frac{-\left\|\mathbf{r}_q - \gamma_q \mathbf{G}_q \mathbf{b}_q\right\|^2}{2\sigma^2}\right]. \quad (2)$$

It is possible to marginalise the amplitude parameters $\mathbf{b}_q$, which is desirable due to the high posterior correlation between ω_q and $\mathbf{b}_q$; another benefit is that the size of the parameter space to be optimised is reduced significantly. We can also marginalise σ, in a similar manner to [10], to yield[1]

$$p(\omega_q, H_q, \gamma_q|\ \Phi_{-(\mathbf{q})}, \mathbf{d}) \propto \frac{P_q}{B^{-M_q}} \frac{R_\sigma}{\left|\mathbf{G}_q^{\mathrm{T}}\mathbf{G}_q\right|^{\frac{\gamma_q}{2}}} \times \frac{\Gamma(\varepsilon_q)}{\pi^{\varepsilon_q}\left[\left\|\mathbf{r}_q\right\|^2 - \gamma_q \left\|\mathbf{f}_q\right\|^2\right]^{\varepsilon_q}} \quad (3)$$

$$\varepsilon_q = \frac{N - M_q}{2} \qquad M_q = 2\gamma_q H_q$$

$$\mathbf{f}_q = \mathbf{G}_q \hat{\mathbf{b}}_q \qquad \hat{\mathbf{b}}_q = \mathbf{G}_q (\mathbf{G}_q^{\mathrm{T}}\mathbf{G}_q)^{-1} \mathbf{G}_q^{\mathrm{T}} \mathbf{r}_q.$$

Here, $\hat{\mathbf{b}}_q$ is the familiar least-squares expression for the harmonic amplitudes, and $\mathbf{f}_q$ is the projection of the residual $\mathbf{r}_q$ onto the model given by the parameters $\{\hat{\mathbf{b}}_q, \omega_q, H_q\}$. The dominant term of (3) is the expression in the denominator which is an energy-fitting function — as the energy in the projection approaches that of the residual, the term tends to zero. Most of the other terms are model order dependent, and penalise overfitting. The use of a conditional distribution based on the residual $\mathbf{r}_q$ rather than a joint distribution based on the signal $\mathbf{d}$ has the advantage that execution time scales linearly with the number of notes Q rather than as Q^3, and the $\mathbf{G}_q^{\mathrm{T}}\mathbf{G}_q$ matrices are less prone to ill-conditioning.[2]

[1] This is approximate, due to the limits of integration on $\mathbf{b}_q$ and σ.

[2] Since the joint $\mathbf{G}$ matrix would be composed of the catenation of each $\mathbf{G}_q$, and if any harmonic of one note is close to a harmonic of another note, then $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ would be ill-conditioned, producing nonsensically large $\hat{\mathbf{b}}$ estimates.

3 CHOICE OF TRANSITION KERNELS

A novel choice of transition kernels is presented here for the generation of samples of ω_q, H_q and γ_q from their conditional distributions. The kernels generate trial parameter values from a proposal density, which are then accepted (by the M-H acceptance function) on the basis of how they affect the conditional posterior.

The form of (3) is a likely to be a complex multi-modal distribution with sharp peaks. In order to ensure that the parameter space is explored efficiently, we combine a number of transition kernels.

The first exploits the nature of a harmonic series and proposes a new value which is some multiple of the current value. The factor is chosen randomly from a set of values such that the proposal distribution is reversible; the set $\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{2}, 2, 3\}$ works well in practice.

The second transition kernel for ω_q is an independence sampler whose proposal distribution is independent of the current state [12]. This approach has been recently used for sinusoidal models with a proposal density based on the periodogram [1], to coerce the ω_q samples into high probability regions. We adopt a similar technique, but instead use a harmonic transform (HT) of the residuals in the proposal density. We define a P-th order HT, $\mathcal{H}_{k,P}(\mathbf{x})$ as

$$X_p[k] = \sum_{n=0}^{N_f - 1} x[n]\, e^{-j2\pi pkn/N_f}$$

$$\mathcal{H}_{k,P}(\mathbf{x}) = \sum_{p=1}^{P} X_p^*[k]\, X_p[k]$$

and the proposal distribution as

$$q(\omega) = c \sum_{k=1}^{K} \delta(\omega - k\Delta\omega)\, \mathcal{H}_{k,P}(\mathbf{r}_q)$$

where $c = 1/\sum_{k=1}^{K} \mathcal{H}_{k,P}(\mathbf{x})$, $K = \lfloor N_f/P \rfloor$ and $\Delta\omega$ is the frequency bin spacing, since it can be shown, for $\omega_q = k\Delta\omega$, that

$$\left\|\mathbf{f}_q\right\|^2 \approx \frac{2}{N} \left\|\mathbf{G}^{\mathrm{T}}\mathbf{r}_q\right\|^2 \approx \frac{2}{N} \mathcal{H}_{k,H_q}(\mathbf{r}_q)$$

which has its modes in the same locations as (3). The HT can be calculated efficiently from the FFT since $X_1[k]$ is the DFT of $\mathbf{x}$ and $X_p[k] = X_1[pk]$.

The third transition kernel is a perturbation step — a random-walk sampler with a small variance — which ensures that the Markov chain is ergodic [7, 8]. The combination of all three kernels ensure that the Markov chain finds the prominent modes in the distribution, and other harmonically related modes, whilst performing small steps to find local maxima.

The transition kernels for H_q can be much simpler than those for ω_q— good results were obtained from combining an independence sampler with a uniform probability distribution and a random-walk sampler. Sampling for γ_q is a matter of using the trial value $\gamma_q^* = 1 - \gamma_q^k$ since the M-H acceptance function then considers the posterior for both states.

It was also found that octave errors (*i.e.*, frequency estimates of half or double the true value) can be reduced by including a M-H step which proposes a joint move for ω_q and H_q. This step takes advantage of the non-uniqueness of the harmonic representation, *e.g.*, if the fundamental frequency is halved and the number of harmonics is doubled, then the new parameters will generate a similar or identical $\mathbf{f}_q$. If a move is made which reduces H_q and increases ω_q whilst keeping $\mathbf{f}_q$ almost constant, then this may yield a higher posterior probability due to the lower model order.[3]

4 OBTAINING MAP ESTIMATES

The Bayesian inference we wish to perform on the chain of samples is simply to find the MAP parameter estimates. The output of the Markov chain is a sequence of (dependent) samples drawn from the joint posterior. The numerical value of the joint posterior for any set of parameter values in the chain can be calculated from the marginalisation of the linear parameters and error standard deviation from the full joint posterior (1),

$$p(\{\omega_1, H_1, \gamma_1\}, \ldots, \{\omega_Q, H_Q, \gamma_Q\}|\mathbf{d}) = \int \cdots \int p(\Phi|\mathbf{d})\, d\mathbf{b}_1 \ldots d\mathbf{b}_Q\, d\sigma.$$

For most practical purposes, MAP parameter estimates can be obtained by histogramming the Markov chain output, although these are effectively marginal rather than joint parameter estimates.

5 RESULTS

We analysed an extract consisting of the superposition of two completely independent monophonic musical phrases, one vocal and the other a played by a saxophone. Both are reasonably *legato* (*i.e.*, have a smooth pitch variation) and neither have sharp initial transients. Figure 1 shows the pitch variation over time — the x axis shows the window number (window length of 1000 samples with 50% overlap, at a rate of 88.2 frames/sec) and the y axis shows the fundamental frequency in Hz. The top plot shows the pitch variation, obtained by monophonic analysis ($Q = 1$) of each part separately using our algorithm, and the lower plot shows the results of the algorithm applied to the sum of the signals.

A simple prior was imposed on ω_q, giving a higher probability to values close to the mean of the frequency in the previous 4 frames, which improved the continuity of the frequency tracks and eliminated a few octave errors. The algorithm performed well in this example, even over the closely-spaced region in frames 200–250. The loss of detection in frames 300–330 is due to the presence of an unvoiced sound in the vocal part.[4]

6 FUTURE WORK

Future work will focus on the generalisation of the model to better represent 'real-world' data, *e.g.*, AR modelling of residuals, variable waveform start and end-points, inharmonicity, and amplitude and frequency variations. Higher level modelling, particularly joint estimation over a number of frames, will also be required to produce better continuity of frequency tracks.

7 CONCLUSIONS

We have presented a method of estimating the parameters of the linear combination of an unknown number of harmonic signals. The use of indicator variables implicitly represents the model order, and maximum *a posteriori* estimation of the model parameters and indicator variables is performed with a Metropolis-Hastings algorithm.

A novel choice of transition kernels is proposed for the frequency and harmonic number parameters, which combine different proposal distributions to exploit the structure of the posterior distribution of a harmonic signal, and perform different types of parameter moves to explore the whole parameter space.

Other envisaged applications of this method include interpolation, coding, enhancement and restoration of musical material.

8 REFERENCES

[1] C. Andrieu and A. Doucet. Joint Bayesian detection and estimation of noisy sinusoids via reversible jump MCMC. Technical Report CUED/F-INFENG/TR.318, Cambridge University Engineering Department, 1998.

[2] G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, 1989.

[3] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[4] H.-T. Li and P. Djurić. An iterative procedure for joint Bayesian spectrum and parameter estimation of harmonic signals. In *IEEE International Symposium on Circuits and Systems*, volume 2, pages 513–516, Atlanta, May 1996. Proc. IEEE.

[5] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754, 1986.

[3] This is particularly true if a number of the harmonics are of very low amplitude. For example if all the odd harmonics are low, it suggests there is an octave error and the frequency should be doubled.

[4] The monophonic case may have detected reverberation from the previous voiced section.

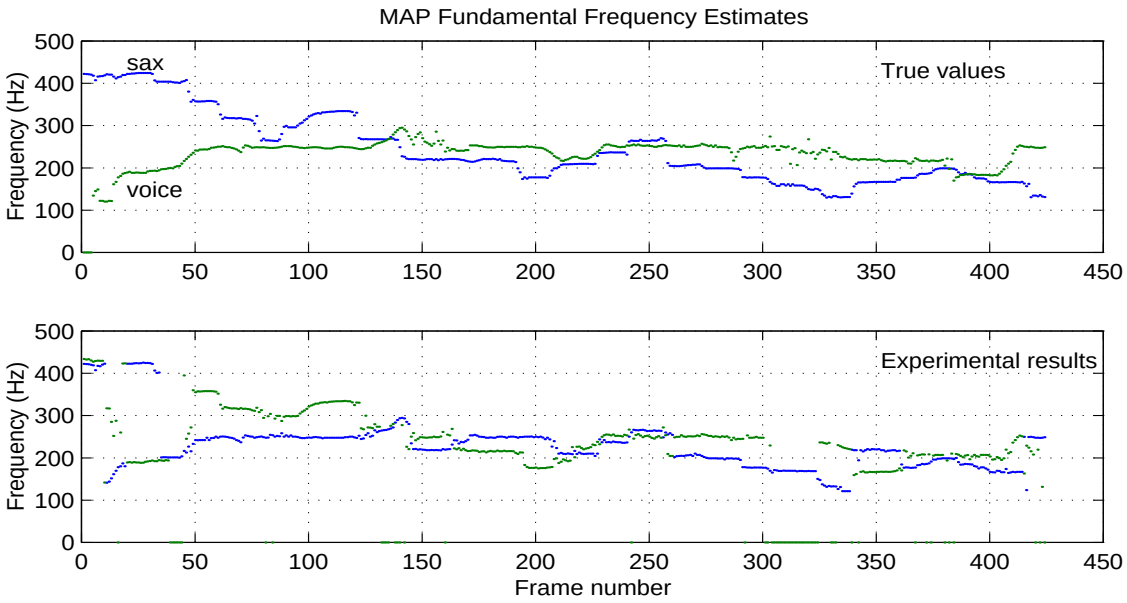

Figure 1: Frequency tracks over time for two superimposed monophonic musical extracts.

[6] R. J. McAulay and T. F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *Proc. ICASSP*, volume 1, pages 249–252, 1990.

[7] R. M. Neal. Probabilistic inference using MCMC methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Canada, September 1993. Available as Compressed Postscript[5].

[8] G. O. Roberts and J. S. Rosenthal. Markov chain Monte Carlo: some practical implications of theoretical results. Preprint, Feb 1997.

[9] X. Rodet. Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models. In *IEEE UK Symposium on applications of Time-Frequency and Time-Scale Methods*, pages 111–120, August 1997.

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

[11] X. Serra. *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*. PhD thesis, Stanford University, 1990.

[12] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1762, 1994. Available as Postscript[6].

[5] ftp://ftp.cs.utoronto.ca/pub/radford/review.ps.Z
[6] http://www.stat.umn.edu/PAPERS/tech-reports/tr560.ps