Efficient Bayesian inference for harmonic models via adaptive posterior factorization

Emmanuel Vincent, Mark D. Plumbley

I R I S A

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION INTERNE

N° 1841

EFFICIENT BAYESIAN INFERENCE FOR HARMONIC MODELS VIA ADAPTIVE POSTERIOR FACTORIZATION

EMMANUEL VINCENT AND MARK D. PLUMBLEY

ISSN 1166-8687

IRISA

CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Efficient Bayesian inference for harmonic models via adaptive posterior factorization

Emmanuel Vincent * and Mark D. Plumbley **

Systèmes cognitifs
Projet METISS

Publication interne n° 1841 — Avril 2007 — 20 pages

Abstract: Harmonic sinusoidal models are an essential tool for audio signal analysis. Bayesian harmonic models are particularly interesting, since they allow the joint exploitation of various priors on the model parameters. However existing inference methods often rely on specific prior distributions and remain rather slow for realistic data. In this article, we investigate a generic inference method based on approximate factorization of the joint posterior into a product of independent distributions on small subsets of parameters. We discuss the conditions under which this factorization holds true and propose two criteria to choose these subsets adaptively. We evaluate the resulting performance experimentally for the task of musical score transcription using different levels of factorization.

Key-words: Bayesian inference, harmonic model, adaptive factorization, posterior dependence

(Résumé : tsvp)

E. Vincent was partly funded by EPSRC grant GR/S75802/01.

* METISS group, IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France – emmanuel.vincent@irisa.fr
** Centre for Digital Music, Department of Electronic Engineering, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom – mark.plumbley@elec.qmul.ac.uk

Centre National de la Recherche Scientifique (UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique et en Automatique – unité de recherche de Rennes

Inférence bayésienne efficace pour les modèles harmoniques par factorisation adaptative de la distribution *a posteriori*

Résumé : Les modèles sinusoïdaux harmoniques constituent un outil essentiel pour l'analyse des signaux audio. Les modèles harmoniques bayésiens sont particulièrement intéressants, car ils permettent d'exploiter conjointement les différents *a priori* disponibles sur les paramètres des modèles. Cependant, les méthodes d'inférence existantes reposent souvent sur des distributions *a priori* spécifiques et restent relativement lentes sur des données réalistes. Dans cet article, nous étudions une méthode d'inférence générique basée sur la factorisation approchée de la distribution *a posteriori* en un produit de distributions indépendantes sur des petits sous-ensembles de paramètres. Nous discutons les conditions sous lesquelles cette factorisation est vraie et nous proposons deux critères pour choisir ces sous-ensembles de façon adaptative. Nous évaluons la performance de cette méthode expérimentalement pour la tâche de transcription de la partition musicale pour divers niveaux de factorisation.

Mots clés : Inférence bayésienne, modèle harmonique, factorisation adaptative, dépendance *a posteriori*

1 Introduction

Music and speech involve different types of sounds, including periodic, transient and noisy sounds. Short-term stationary periodic sounds composed of sinusoidal partials at harmonic or near-harmonic frequencies are perceptually essential, since they contain most of the energy of musical notes and vowels. Harmonicity means that at each instant the frequencies of the partials are multiples of a single frequency called the fundamental frequency. Estimating the periodic sounds underlying a given signal, *i.e.* estimating their fundamental frequencies and the amplitudes and phases of their partials, is required or useful for many applications, such as speech prosody analysis [1], musical score transcription and instrument recognition [2] and low bit-rate compression [3]. This problem is particularly difficult for polyphonic signals, *i.e.* signals containing several concurrent periodic sounds, since different periodic sounds may exhibit partials overlapping at the same frequencies.

Existing methods for polyphonic fundamental frequency estimation are often based on one of two approaches [2]: either validation of fundamental frequency candidates given by the peaks of a short-term auto-correlation function [4, 5, 6] or inference of the hidden states of a probabilistic model of the signal short-term power spectrum based on learned template spectra [7, 8, 9]. These approaches have achieved limited performance on complex polyphonic signals so far [2, 6]. Moreover neither approach provides estimates for the amplitudes and phases of the partials, which are needed for musical instrument recognition or low bit-rate compression.

A promising way to address these issues is to rely on a probabilistic model of the signal waveform incorporating various prior knowledge. Two families of such models have been proposed in the literature for music signals. One family introduced in [10, 11] models each musical note signal in state-space form by a discrete fundamental frequency and a fixed number of damped oscillators at harmonic frequencies with independent transition noises. Decoding is achieved either via linear Kalman filtering or variational approximation [12], depending whether the damping factors are fixed or subject to additional transition noises. These inference methods restrict the prior distribution of the transition noises to be Gaussian or from a class of conjugate priors [13] respectively. Another family of models described in [14, 15, 16] represents musical note signals by continuous fundamental frequency, amplitude and phase parameters, inferred using Markov Chain Monte Carlo (MCMC) methods [13]. These methods are theoretically applicable to all prior distributions, but tend to be rather slow in practice. Thus the chosen priors are partly motivated by computational issues [16]. In particular, the amplitudes of the partials are modeled by independent uniform priors or by conjugate zero-mean Gaussian priors.

For both families of models, the above priors exhibit some differences with the empirical parameter distributions. In particular, they do not penalize partials with zero amplitude. This can lead to missing estimated notes for signals composed of several notes at simple rational fundamental frequency ratios [14, 16], or to erroneous fundamental frequency estimates, typically equal to a multiple or a submultiple of the true fundamental frequencies [16]. To help solving these limitations, we recently proposed a harmonic model including probabilistic priors motivated by empirical parameter distributions and used a variant of the diagonal Laplace method for fast parameter inference [3].

In this article, we propose a more accurate fast inference method for probabilistic harmonic models, based on approximate factorization of the joint posterior into a product of independent distributions on subsets of parameters. This method is designed for models of the form described in [14, 15, 16, 3], involving explicit frequency, amplitude and phase parameters. It is generic, in that it can be applied to a wide range of priors, and adaptive, since the level of factorization depends on the observed signal and the hypothesized notes. This constitutes a crucial difference compared to variational approximation methods, where the terms of the factorization are fixed *a priori* and their parameters can only be computed for certain classes of priors. We complete our preliminary work [17] by discussing the extension of this method to nongaussian residuals and alternative model structures, investigating a new criterion for the choice of the parameter subsets and providing a detailed experimental evaluation.

The structure of the rest of the article is as follows. In section 2, we present a possible Bayesian network structure for harmonic models and make some mild assumptions about the parameter priors. Then, we describe the proposed inference method in section 3 and extend it to alternative model structures. In section 4, we evaluate its performance for the task of musical score transcription on short time frames. We conclude in section 5 and suggest some perspectives for future research.

2 Assumptions about the model

The harmonic models in [14, 15, 16, 3] are variations of the same concept. They all represent the observed music signal as a sum of note signals, each composed of several sinusoidal partials parametrized by a sequence of random variables spanning successive time frames. However, the chosen variables and their conditional dependency structure are slightly different for each model. For the sake of clarity, we first discuss our approach for the model structure in [3], which involves fewer variables.

2.1 Bayesian network structure

On each time frame, the model described in [3] exhibits the four-layer Bayesian network structure shown in Figure 1. Each layer models the observed signal frame $x(t)$ of length T at a different abstraction level.

The bottom layer represents the underlying musical score. In western music, the normalized fundamental frequency f_p of each note may vary across frames but remains close to a discrete pitch of the form

$$\mu_p = \frac{440}{F_s} 2^{\frac{p-69}{12}} \tag{1}$$

where F_s is the sampling frequency in Hz and p an integer value on the MIDI semitone scale. Assuming no unison, *i.e.* several notes corresponding to the same discrete pitch cannot be present at the same time, each point p on the MIDI scale is associated with a binary activity state S_p determining whether a note with that discrete pitch is active or not.

The signal $s_p(t)$ corresponding to each active note is then defined in the middle layers for $0 \leq t \leq T-1$ by

$$s_p(t) = w(t) \sum_{m=1}^{M_p} a_{pm} \cos(2\pi m f_p t + \phi_{pm}) \tag{2}$$

where $w(t)$ is the framing window and f_p, a_{pm} and ϕ_{pm} are respectively its normalized fundamental frequency and the amplitude and the phase of its m-th partial. The amplitudes of the partials are assumed to depend on an amplitude scale factor r_p accounting for the total power of note p. The number of partials M_p is constrained as a function of the note pitch p to

$$M_p = \min\left(\frac{1}{2\mu_p}, M_{\max}\right) \tag{3}$$

so that the partials fill the whole observed frequency range up to a maximal number of partials $M_{\max}$. Finally, the observed signal is modeled in the top layer as

$$x(t) = \sum_{p \text{ s.t.} S_p=1} s_p(t) + e(t) \tag{4}$$

where $e(t)$ is the residual.

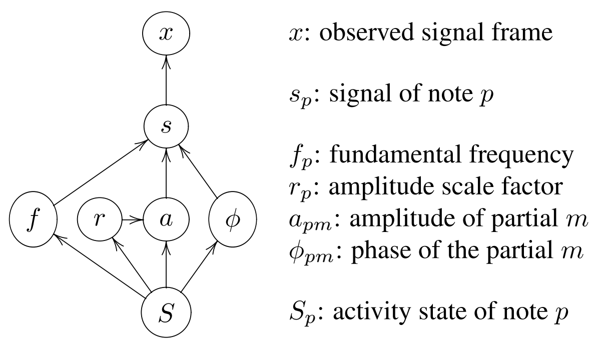

Figure 1: Bayesian network structure of the harmonic model in [3]. Circles denote vector random variables (some of variable size) and arrows conditional dependencies.

2.2 Assumptions about the parameter priors

The inference method proposed below is valid given some mild assumptions about the parameter priors. Classically, we assume that the fundamental frequencies f_p of different notes p and the phases ϕ_{pm} of different partials (p, m) are independent *a priori* and that the amplitudes a_{pm} of different partials are independent *a priori* given the amplitude scale factors

r_p. We also assume that the prior distribution of each fundamental frequency f_p enforces proximity to the underlying discrete pitch μ_p. Finally, we make the hypothesis that the residual $e(t)$ has a continuous distribution and that its values at distinct frequencies are almost independent *a priori*, *i.e.* its prior distribution $P(e) = P(x|f, a, \phi)$ can be approximately factored as

$$P(e) \approx \prod_{\nu=0}^{T-1} P(E_\nu) \tag{5}$$

where E_ν are the discrete Fourier transform coefficients of $e(t)$.

Note that the ubiquitous time-domain Gaussian i.i.d. prior satisfies this hypothesis, since it is equivalent to a Gaussian i.i.d. prior on the Fourier coefficients. A more general prior of particular interest in the following is the frequency-weighted Gaussian distribution [3]

$$P(e) = (2\pi\sigma^2)^{-T/2} \prod_{\nu=0}^{T-1} \exp\left(-\frac{\gamma_\nu |E_\nu|^2}{2\sigma^2}\right) \tag{6}$$

where γ_ν are constant positive weights. This prior can be rewritten as $P(e) = (2\pi\sigma^2)^{-T/2} \exp(-\|e\|_\gamma^2/(2\sigma^2))$ where $\|e\|_\gamma^2 = \sum_{\nu=0}^{T-1} \gamma_\nu |E_\nu|^2$ is the squared weighted Euclidean norm of the Fourier coefficients.

3 Bayesian inference via adaptive posterior factorization

Harmonic models are typically employed to solve the score transcription task, which consists of estimating the Maximum *A Posteriori* (MAP) vector of activity states $\hat{S} = \arg\max P(S|x)$. The posterior probability of S equals

$$P(S|x) = \int P(S, f, r, a, \phi|x)\, df\, dr\, da\, d\phi \tag{7}$$

where f, r, a, ϕ denote the vectors of parameters f_p, r_p, a_{pm}, ϕ_{pm}, and the joint posterior is given by Bayes law

$$P(S, f, r, a, \phi|x) \propto P(e)P(a|r, S)P(\phi|S)P(r|S)P(f|S)P(S). \tag{8}$$

The computation of the integral in (7) is known as the Bayesian marginalization problem [12].

A number of sampling techniques are available to compute such integrals [12]. However they appear unsatisfactory in this context. Numerical integration on a uniform grid is intractable, since the number of parameters is typically of the order of one hundred. Integration via importance sampling [18] is slow, since the variance of the importance weights, which is proportional to that of the estimate, increases sharply with the number of parameters [12]. Sampling of the joint posterior via reversible jump MCMC [13] is also slow [16].

Fast inference can be achieved instead by estimating the MAP parameter values $(\hat{f}, \hat{r}, \hat{a}, \hat{\phi})$ $= \arg\max P(S, f, r, a, \phi | x)$ using any standard nonlinear optimization algorithm and approximating the joint posterior around these values by a simpler distribution which can be integrated analytically or by tabulation. Relevant techniques include the diagonal Laplace approximation [19], which factors the posterior into a product of parameter-wise univariate Gaussian distributions, and its variant proposed in [3] with a specific nongaussian distribution for the phase parameters. The full Laplace approximation [19] performs poorly due to unbounded integration over the phase parameters [17].

The proposed inference method can be seen as a compromise between sampling-based and full factorization-based techniques, since it relies on partial factorization of the posterior and sampling over subsets of parameters. Various levels of factorization can be achieved depending on the MAP parameter values.

3.1 Conditional posterior factorization over the partials

Let us assume initially that the residual follows the frequency-weighted Gaussian prior (6) and that the harmonic partials of the hypothesized fundamental frequencies have "different enough" frequencies. This is true for a single hypothesized note, but generally not for several notes. Mathematically, this translates into the fact that the windowed complex sinusoidal signals

$$z_{pm}(t) = w(t) e^{2i\pi m f_p t} \tag{9}$$

corresponding to different partials are mutually orthogonal

$$\langle z_{pm}, z_{p'm'} \rangle_\gamma = 0 \quad \forall (p, m) \neq (p', m') \tag{10}$$

according to the dot product $\langle ., . \rangle_\gamma$ consistent with the weighted Euclidean norm $\|.\|_\gamma$. This dot product is defined for two signals $z(t)$ and $z'(t)$ by

$$\langle z, z' \rangle_\gamma = \sum_{\nu=0}^{T-1} \gamma_\nu Z_\nu \bar{Z}'_\nu \tag{11}$$

where Z_ν and Z'_ν are the discrete Fourier transform coefficients of $z(t)$ and $z'(t)$ and $\bar{Z}'_\nu$ is the complex conjugate of Z'_ν. The orthogonality property (10) formalizes the fact that partials with "different enough" frequencies have almost disjoint frequency supports and can be assumed to hold true for all possible frequency weights γ_ν. When the frequencies of the partials are not too close to Nyquist, the negative frequency sinusoidal signals $\bar{z}_{pm}(t) = w(t) e^{-2i\pi m f_p t}$ are also orthogonal to their positive counterparts: $\langle z_{pm}, \bar{z}_{p'm'} \rangle_\gamma = 0$ for all (p, m) and (p', m'). The observed signal $x(t)$ can then be decomposed into a sum of sinusoidal signals at the frequencies of the hypothesized partials by orthogonal projection onto the two-dimensional subspaces spanned by $(z_{pm}, \bar{z}_{pm})$

$$x(t) = \frac{1}{2} \sum_{p,m} \tilde{a}_{pm} (e^{i\tilde{\phi}_{pm}} z_{pm}(t) + e^{-i\tilde{\phi}_{pm}} \bar{z}_{pm}(t)) + \tilde{e}(t). \tag{12}$$

The projection coefficients given by

$$\tilde{a}_{pm}e^{i\tilde{\phi}_{pm}} = 2\frac{\langle x, z_{pm}\rangle_\gamma}{\|z_{pm}\|_\gamma^2} \tag{13}$$

represent the amplitude and phase values of each partial minimizing the norm of the residual $e(t)$. Given hypothesized values a_{pm} and ϕ_{pm}, the residual can be decomposed as a sum of mutually orthogonal terms

$$\begin{aligned} e(t) = \frac{1}{2}\sum_{p,m} &\left(\tilde{a}_{pm}e^{i\tilde{\phi}_{pm}} - a_{pm}e^{i\phi_{pm}}\right) z_{pm}(t) \\ &+ \left(\tilde{a}_{pm}e^{-i\tilde{\phi}_{pm}} - a_{pm}e^{-i\phi_{pm}}\right) \bar{z}_{pm}(t) + \tilde{e}(t). \end{aligned} \tag{14}$$

The squared norm of the residual then equals by analytical computation

$$\|e\|_\gamma^2 = \sum_{p,m} D_{pm} + D_0 \tag{15}$$

with $D_0 = \|\tilde{e}\|_\gamma^2$ and

$$D_{pm} = \frac{1}{2}\|z_{pm}\|_\gamma^2 \left((a_{pm} - \tilde{a}_{pm})^2 + 4\tilde{a}_{pm}a_{pm}\sin^2\frac{\phi_{pm} - \tilde{\phi}_{pm}}{2}\right). \tag{16}$$

Using (8) and the relationship between $P(e)$ and $\|e\|_\gamma^2$, this decomposition leads to the exact factorization of the joint posterior into a product of partial-wise bivariate conditional distributions over amplitude and phase parameters

$$\begin{aligned} P(S, f, r, a, \phi|x) \propto\ & P_0(x, f)P(r|S)P(f|S)P(S) \\ & \times \prod_{p,m} P_{pm}(a_{pm}, \phi_{pm}; x, f_p)P(a_{pm}|r_p)P(\phi_{pm}) \end{aligned} \tag{17}$$

where $P_0(x, f) = (2\pi\sigma^2)^{-T/2}\, e^{-D_0/(2\sigma^2)}$ is a constant and $P_{pm}(a_{pm}, \phi_{pm}; x, f_p) = \exp(-D_{pm}/(2\sigma^2))$ a bivariate parametric distribution that can be quickly computed, since it depends on three hyper-parameters only: $\|z_{pm}\|_\gamma^2$, $\tilde{a}_{pm}$ and $\tilde{\phi}_{pm}$. The top part of Figure 2 illustrates the validity of this factorization.

Denoting by $\hat{a}$ and $\hat{\phi}$ the MAP amplitude and phase vectors given f and r and by $\hat{a}_{\underline{pm}}$ and $\hat{\phi}_{\underline{pm}}$ these vectors reduced by one coefficient corresponding to partial (p, m), the above expression can be equivalently rewritten as

$$P(S, f, r, a, \phi|x) = P(S, f, r, \hat{a}, \hat{\phi}|x) \prod_{p,m} \frac{P(a_{pm}, \phi_{pm}|S, f, r_p, \hat{a}_{\underline{pm}}, \hat{\phi}_{\underline{pm}}, x)}{P(\hat{a}_{pm}, \hat{\phi}_{pm}|S, f, r_p, \hat{a}_{\underline{pm}}, \hat{\phi}_{\underline{pm}}, x)}. \tag{18}$$

This equation admits the following interpretation: the first term is the joint posterior value for the MAP amplitude and phase parameters and each quotient term describes the relative drop of this value with different parameters as proportional to the posterior distribution of the parameters of each partial with other parameters being fixed. This equation remains approximately valid in the more general case where the residual follows a nongaussian prior satisfying (5), although quick computation of the quotient terms is not possible anymore. Indeed, when the amplitude and phase parameters are close to their MAP values, the Fourier coefficients of the signal associated with each partial (p, m) are near zero except for a few bins ν whose frequencies are close to mf_p. Thus, if the partials have "different enough" frequencies, each Fourier coefficient E_ν of the residual depends mostly on the parameters of the partial with closest frequency. The probability of the residual $P(e)$ can then be approximately factored into a product of binwise terms, each involving the parameters of at most one partial, which leads to (18) after simple analytical computation.

3.2 Conditional posterior factorization over subsets of partials

In the general case where several partials may have close frequencies, the terms of (18) can still be computed but this equation may not hold true, as shown in the middle part of Figure 2. It is however possible to group partials into subsets such that partials from different subsets have frequencies as different as possible. This can be mathematically formalized by grouping partials (p, m) and (p', m') if and only if

$$|mf_p - m'f_{p'}| \leq f_{\max} \tag{19}$$

where $f_{\max}$ is a manual frequency threshold. Similar arguments as above lead to the approximate factorization of the posterior into a product of multivariate conditional distributions over subsets of amplitude and phase parameters $a_g = \{a_{pm}, (p, m) \in g\}$ and $\phi_g = \{\phi_{pm}, (p, m) \in g\}$, whose terms can be quickly computed by orthogonal projection in the particular case where the residual follows a frequency-weighted Gaussian prior

$$P(S, f, r, a, \phi|x) \approx P(S, f, r, \hat{a}, \hat{\phi}|x) \prod_g \frac{P(a_g, \phi_g|S, f, r, \hat{a}_{\underline{g}}, \hat{\phi}_{\underline{g}}, x)}{P(\hat{a}_g, \hat{\phi}_g|S, f, r, \hat{a}_{\underline{g}}, \hat{\phi}_{\underline{g}}, x)}. \tag{20}$$

A higher threshold $f_{\max}$ increases the accuracy of this equation, but also leads to larger subsets. In practice, it is often possible to obtain a factored expression of similar accuracy with smaller subsets. Indeed there exist some conditions where partials at close frequencies may still be associated with different subsets. An example of such a condition is given in the bottom part of Figure 2 and discussed in [17]. Denoting by the vectors y and y' two disjoint subsets of variables and by $\hat{y}$ and $\hat{y}'$ their MAP values given the rest of the variables y'', we assess the accuracy of the approximation of the joint posterior distribution $P(y, y'|y'')$ by the factored distribution $P(y|\hat{y}', y'')P(y'|\hat{y}, y'')$ using the Kullback-Leibler divergence [12]

$$\mathcal{D}(y, y') = \int P(y, y'|y'') \log_2 \frac{P(y, y'|y'')}{P(y|\hat{y}', y'')P(y'|\hat{y}, y'')}\, dy\, dy'. \tag{21}$$

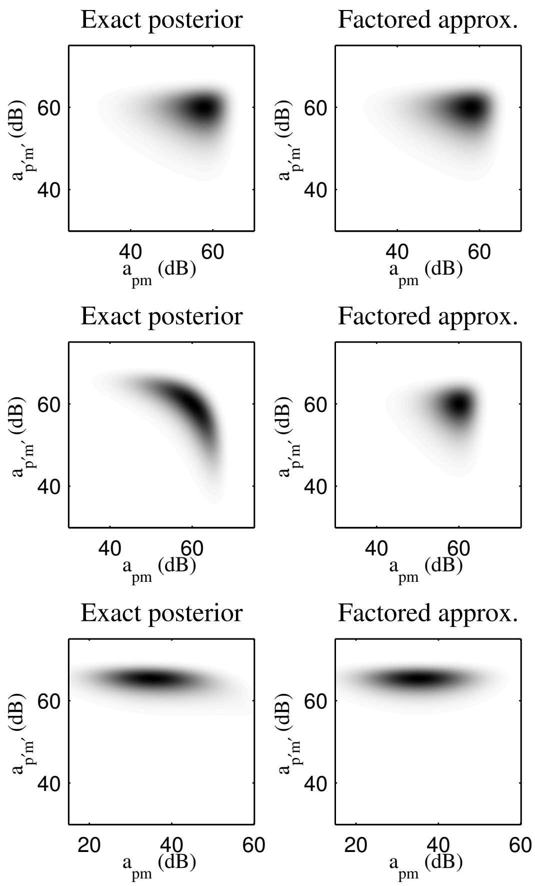

Figure 2: Shape of the joint posterior for a signal containing two partials (p, m) and (p', m') with 60 dB amplitudes, using the priors defined in [3]. Dark areas denote high probability. Top: partials at different frequencies with mean prior amplitudes of 50 dB and 60 dB. Middle: partials at the same frequency with identical mean prior amplitudes of 60 dB. Bottom: partials at the same frequency with mean prior amplitudes of 40 dB and 60 dB. The posterior dependence between a_{pm} and $a_{p'm'}$ equals 0, 0.99 and 0.093 bits respectively.

This quantity is always positive and equal to zero only when the approximation is exact. It can be seen as a measure of the local posterior dependence between y and y' expressed in bits. Indeed, it is analogous to mutual information [12], except that the marginal distribution of each variable is replaced here by its posterior distribution given the MAP value of the other. This suggests that partials (p,m) and (p',m') should belong to the same subset if and only if

$$\mathcal{D}(\{a_{pm},\phi_{pm}\},\{a'_{pm},\phi'_{pm}\}) \geq c_{\min} \tag{22}$$

where $c_{\min}$ is a manual threshold.

Figure 3 shows that the posterior dependence between the parameters of two partials tends to decrease as a function of their frequency difference. However, this decrease is not monotonic: the posterior dependence is typically smaller for frequency differences corresponding to certain zeroes of the discrete Fourier transform of the framing window $w(t)$. Also, for a given frequency difference, posterior dependence values differing by up to three orders of magnitude can be observed. Figure 4 depicts the posterior dependence between the parameters of two notes with a fundamental frequency ratio of 1.5, considering parameters one by one. The third and sixth partials of the lower note have the same frequency as the second and fourth partial of the higher note. The posterior dependence between the parameters of these pairs of partials equals between $10^{-2.4}$ and $10^{1.3}$ bits, while it is smaller than 10^{-11} bits between other pairs.

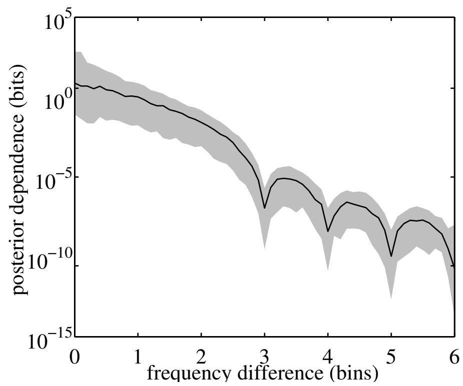

Figure 3: Posterior dependence between the parameters of two partials as a function of their frequency difference, expressed in number of bins of the discrete Fourier transform. The black curve and the gray area denote respectively the median and the two-tailed 95th percentile of the values computed for all the data of Section 4.

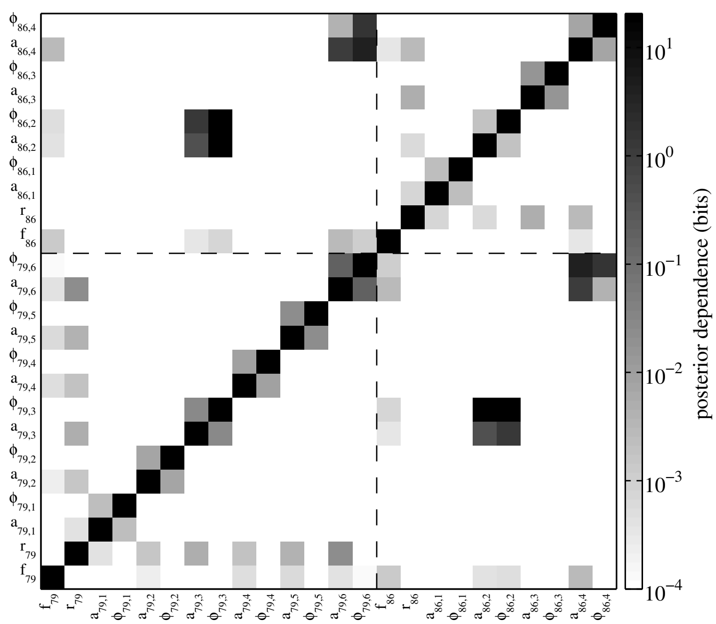

Figure 4: Posterior dependence between the parameters of two notes with pitches $p = 79$ and 86 for a signal consisting of these notes, using the priors defined in [3]. Upper partials are not shown for legibility. Black squares on the diagonal denote infinite posterior dependence values and white squares values below 10^{-4} bits.

3.3 An exploitable posterior factorization

The conditional factorization (20) can be exploited for numerical integration of the posterior, either by sampling on a uniform grid or by importance sampling. Indeed integration over amplitude and phase parameters can be achieved by multiplying lower dimension integrals over the parameters of each subset of partials. Using sampling on a uniform grid and denoting by N the number of grid points for each scalar variable, P the number of hypothesized notes, $M = \sum_p M_p$ their total number of partials and G the size of the largest subset of partials, this results in a maximal complexity of $\mathcal{O}(\frac{M}{G} N^{2P+2G})$. This is smaller than the complexity of $\mathcal{O}(N^{2P+2M})$ associated with straightforward integration of the joint posterior, but still intractable.

In order to get faster integration, it is necessary to remove additional parameter dependencies. An approximate solution is to replace the free fundamental frequency and amplitude scale parameters by their MAP values within the conditional distributions of amplitude and phase parameters. The remaining joint posterior distribution of fundamental frequency and amplitude scale parameters can be similarly factored over each parameter. This gives

$$P(S,f,r,a,\phi|x) \approx P(S,\hat{f},\hat{r},\hat{a},\hat{\phi}|x) \prod_p \frac{P(f_p|S,\hat{f}_{\underline{p}},\hat{a},\hat{\phi},x)}{P(\hat{f}_p|S,\hat{f}_{\underline{p}},\hat{a},\hat{\phi},x)} \times \prod_p \frac{P(r_p|S,\hat{a}_p)}{P(\hat{r}_p|S,\hat{a}_p)} \prod_g \frac{P(a_g,\phi_g|S,f,r,\hat{a}_{\underline{g}},\hat{\phi}_{\underline{g}},x)}{P(\hat{a}_g,\hat{\phi}_g|S,f,r,\hat{a}_{\underline{g}},\hat{\phi}_{\underline{g}},x)}. \quad (23)$$

This equation allows approximate numerical integration of the posterior with a maximal complexity of $\mathcal{O}(\frac{M}{G} N^{2G})$. Although it is not straightforward to justify mathematically, it appears experimentally valid when the prior distribution of the fundamental frequency parameters enforces proximity to the underlying discrete pitches, in the sense that the error introduced by factorization of the joint posterior over fundamental frequency and amplitude scale parameters is smaller than the one introduced by factorization over amplitude and phase parameters for typical values of $c_{\min}$. This is illustrated in Figure 4, where the posterior dependence between fundamental frequency and amplitude scale parameters and other parameters lies below $10^{-1.6}$ bits, which is smaller than the values of $c_{\min}$ chosen in Section 4.

3.4 Extension to alternative model structures

The proposed marginalization could be extended to alternative harmonic model structures, such as those described in [14, 15, 16]. Indeed, the approximate posterior independence property of partials at different frequencies remains valid.

Among all structural differences, these models consider the number of partials per note M_p as a random variable subject to a certain prior. With narrow fundamental frequency priors as in [14, 15], the proposed method can be directly applied to compute the integrals of the joint posterior for each value of M_p. Note that this results in little additional cost compared to fixed M_p. Indeed, when increasing or decreasing M_p by one, only one subset of partials needs to be updated, while the integral over the other subsets remains constant. With wider fundamental frequency priors as in [16], the posterior becomes multimodal with local maxima at all fundamental frequencies present in the signal and rational multiples of these. Amplitude and phase parameters then exhibit a strong dependence with fundamental frequency parameters. The proposed method can still be applied by splitting the fundamental frequency range into disjoint narrow bands, similar to the semitone bands considered above, and summing the integrals of the joint posterior within each band.

Another difference is that the models in [15, 16] involve additional parameters, namely one global inharmonicity parameter and one spectral shape parameter per note in [15] and one local inharmonicity parameter per partial in [16]. The proposed method can be directly applied in the second case by grouping local inharmonicity parameters with amplitude and phase parameters from the same partials, yielding a maximal complexity of $\mathcal{O}(\frac{M}{G}N^{3G})$. We believe that it could also be applied in the first case after additional factorization of the joint posterior over global inharmonicity and spectral shape parameters. Indeed these parameters are physically similar to fundamental frequency and amplitude scale parameters and should exhibit a similar level of posterior dependence with other parameters.

Finally, the models in [14, 15, 16] describe the residual by a Gaussian whose variance is considered as a random variable. Although this prior does not satisfy (5), the proposed method can still be applied after additional factorization of the posterior over this variance parameter. We believe that this factorization remains approximately accurate provided that the posterior distribution of the variance is unimodal and narrow.

4 Evaluation

The precision of the integral estimates obtained by the proposed marginalization method cannot be assessed for realistic signals, since ground truth integral values are not available. However, the aim of marginalization is often not to compute accurate estimates of the state posteriors $P(S|x)$, but rather to provide an accurate musical score by selecting the right MAP state $\hat{S}$. Therefore we evaluated the performance of the proposed method for the score transcription task.

4.1 Data and evaluation procedure

The parameter priors were chosen as in [3], without assuming knowledge of the true number of notes: the activity states S_p were modeled by Bernoulli priors, the fundamental frequencies f_p, the amplitude scale factors r_p and the amplitudes of the partials a_{pm} by log-Gaussian priors, the phases of the partials ϕ_{pm} by uniform priors and the residual $e(t)$ by a frequency-weighted Gaussian prior. Note that the prior over a_{pm} helps to avoid partials with zero amplitude. The means and variances of these priors were learned on a subset of the RWC Musical Instrument Database[1], while test signals were generated by selecting and mixing

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/

isolated note signals played by five different wind instruments from the University of Iowa Musical Instrument Samples database[2]. More precisely, the test set included 100 one-note signals spanning all discrete pitches from $p = 40$ to 87 and 100 two-note signals corresponding to all possible pitch intervals between 1 and 25 semitones with four different lower pitches $p = 40$, 47, 54 and 61. All signals were sampled at 22.05 kHz and framed with a Hanning window $w(t)$ of length $T = 1024$ (46 ms).

In order to avoid testing all possible vectors of activity states S, 6 candidate vectors (3 with one active note and 3 with two active notes) were pre-selected for each test signal as those minimizing the residual of the orthogonal projection of the observed magnitude spectrum onto the subspace spanned by the typical magnitude spectra of the active notes derived from the amplitude prior, as explained in [3]. The MAP parameters values $(\hat{f}, \hat{r}, \hat{a}, \hat{\phi})$ were computed for each candidate using the subspace trust region optimization algorithm implemented in Matlab's lsqnonlin function[3]. The factored expression (23) was then obtained by grouping the partials using either the frequency difference criterion (19) or the posterior dependence criterion (22). The latter was computed by numerical integration on a uniform grid with 11 points per variable (or about 1.5×10^4 samples per pair of partials) for all pairs of partials with frequency difference smaller than 2.5 bins. The thresholds $f_{\max}$ and $c_{\min}$ were varied between 0 and 2 bins and between 10^3 and $10^{-1.5}$ bits respectively, resulting in a variation of the maximal number of partials per subset from one to three. Each term of the factored posterior was subsequently integrated by sampling on a uniform grid with N points per variable, resulting in a total of $N_{\text{tot}} = N^{2P} + \sum_g N^{2|g|}$ samples per candidate where $|g|$ denotes the number of partials in subset g. The average value of N_{tot} over all test signals and all candidates was varied between 10^5 and 10^7. We also tried integration of these terms via importance sampling [18], but this did not significantly affect performance, despite an increased computation time. We also evaluated the variant of the diagonal Laplace method employed in [3] for comparison.

Each estimated note was considered to be correctly transcribed if it was actually present in the test signal. Performance was then classically assessed by the F-measure $F = 2RP/(R+P)$ in percent, where the recall R is the ratio of the total number of correctly transcribed notes divided by the true number of notes and the precision P is the proportion of correctly transcribed notes among the estimated notes [20, 6]. The computation time was measured for a Matlab implementation on a 1.2 GHz dual CPU computer.

4.2 Results

With one-note signals, the proposed method resulted in F=100% (R=100%, P=100%) for all settings of $f_{\max}$, $c_{\min}$ and N_{tot}. The method in [3] also gave perfect results with a faster average computation time of 1.1 s per candidate, mostly due to the optimization of the MAP parameters.

[2]http://theremin.music.uiowa.edu/MIS.html
[3]http://www.mathworks.com/access/helpdesk_r13/help/toolbox/optim/lsqnonlin.html

The results with two-note signals are depicted in Figure 5. The performance of the proposed method with a large number of integration samples $N_{\text{tot}} = 10^7$ increases from F=93.7% (R=89.5%, P=98.4%) to F=96.9% (R=94.5%, P=99.5%) for both grouping criteria when the average number of partials per subset increases. This difference is statistically significant, as confirmed by a McNemar's p value [21] of 5×10^{-4}. By comparison, the method in [3] achieved a performance of F=92.6% (R=88.0%, P=97.8%), which is not statistically different from that of the proposed method with a single partial per subset. The posterior dependence grouping criterion appears more robust towards a small number of integration samples. Indeed the performance curve with $N_{\text{tot}} = 10^5$ wanders less around the curve with $N_{\text{tot}} = 10^7$ for this criterion. The largest value of $c_{\min}$ yielding maximal performance is $c_{\min} \simeq 10^0$ bits, resulting in as little as 7% of partials found within subsets of two partials for two-note candidates and no partials found within subsets of three or more. The computation time with $N_{\text{tot}} = 10^5$ is then equal to 2.0 s per candidate on average. This can be split into about 1.1 s for the optimization of the MAP parameters, 0.2 s for the computation of the posterior dependence between the partials and 0.7 s for the numerical integration of the terms of the factored posterior. This is much faster than previously reported computation times for MCMC methods with similar models, *e.g.* 1080 s per note with $T = 6000$ using a 2.6 GHz dual CPU computer in [16], corresponding to about 800 s per test signal for two-note signals of length $T = 1024$ with our computer.

The remaining eleven errors made by the proposed method with the best setting are as follows. The upper note is missing from the transcription of three chords with fundamental frequency ratios of 2, two chords with fundamental frequency ratios of 3 and five chords from 1 to 5 semitones with lower pitch $p = 40$. In addition, the upper note is wrongly transcribed within a 6-semitone chord with lower pitch $p = 40$, corresponding to a fundamental frequency error ratio of 4. These errors correspond to situations well known to be difficult, where all the partials of one note overlap with the partials of the other or where the frequency resolution is too small to distinguish multiple notes at low fundamental frequencies. It is likely that these errors are due to the inherent uncertainties of the model rather than the chosen inference method. Note that correct transcription was nevertheless achieved for one chord with a fundamental frequency ratio of 2, two chords with fundamental frequency ratios of 3 and all (four) chords with fundamental frequency ratios of 4, while such situations typically result in transcription errors for other models [14, 16].

5 Conclusion

We proposed a method for the approximate factorization of the joint posterior of a harmonic model based on the use of a local posterior dependence criterion and exploited it for fast Bayesian inference. Although factorization based on this criterion is theoretically feasible for any Bayesian model, it does not necessarily provide small parameter subsets, which are needed for subsequent numerical integration. The key property of harmonic models demonstrated here is that the parameters of partials with different frequencies are approximately independent *a posteriori*. The proposed method is generic and adaptive, in the sense that it

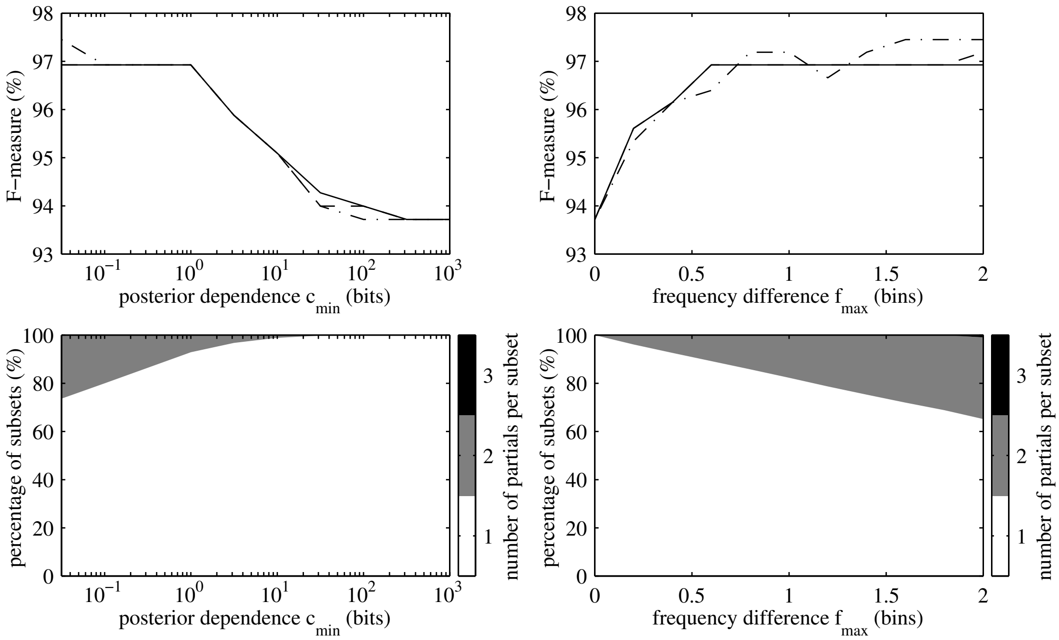

Figure 5: Score transcription results for two-note signals using posterior factorization based on posterior dependence (left) or frequency difference (right) and integration on a uniform grid. Top: F-measure with various grid sizes (plain: $N_{\text{tot}} = 10^7$, dashed: $N_{\text{tot}} = 10^6$, dash-dotted: $N_{\text{tot}} = 10^5$). Bottom: Percentage of partials from two-notes candidates within subsets of one, two or three partials. The percentage of partials within subsets of three equals 0.2% for $c_{\text{min}} = 10^{-1.5}$ bits, 1% for $f_{\text{max}} = 2$ bins and 0 in all other cases. All partials from one-note candidates are within subsets of one.

can be applied to a wide range of priors and that the level of factorization depends on the observed signal and the hypothesized notes. This is an important difference with variational approximation methods, which also rely on factorization of the posterior but are limited to certain classes of prior distributions and often assume a fixed factored expression. Another difference is that the proposed method relies on the computation of MAP parameter values instead of the iterative update of variational parameters, which is intrinsically faster.

To improve the accuracy of the factorization, it would be interesting to investigate transformations of the parameters resulting in a smaller posterior dependence. The minimization of the dependence between subsets of random variables described by a sequence of samples is known as the Independent Subspace Analysis (ISA) problem and can be solved in the case of linear transformations by Independent Component Analysis (ICA) followed by grouping of the transformed variables [22]. This approach could easily be combined with subsequent integration based on importance sampling, and this may also allow other Bayesian models, which do not readily satisfy the posterior independence property, to benefit from the proposed inteference method.

Acknowledgment

The first author wishes to thank Cédric Févotte and Simon J. Godsill for motivating discussions about the influence of prior distributions on the performance of harmonic models and the practical use of MCMC methods.

References

[1] M. Horne (Ed.), Prosody: theory and experiment, Kluwer Academic Publishers, Boston, MA, 2000.

[2] A. Klapuri, M. Davy, Signal processing methods for music transcription, Springer, New York, NY, 2006.

[3] E. Vincent, M. D. Plumbley, Low bitrate object coding of musical audio using Bayesian harmonic models, IEEE Trans. on Audio, Speech and Language Processing, To appear.

[4] D. P. W. Ellis, Prediction-driven computational auditory scene analysis, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA (1996).

[5] M. Wu, D. Wang, G. J. Brown, A multipitch tracking algorithm for noisy speech, IEEE Trans. on Speech and Audio Processing 11 (3) (2003) 229–241.

[6] M. P. Ryynänen, A. P. Klapuri, Polyphonic music transcription using note event modeling, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 319–322.

[7] C. Raphael, Automatic transcription of piano music, in: Proc. Int. Conf. on Music Information Retrieval (ISMIR), 2002, pp. 15–19.

[8] E. Vincent, Musical source separation using time-frequency source priors, IEEE Trans. on Audio, Speech and Language Processing 14 (1) (2006) 91–98.

[9] A. Cont, Realtime multiple pitch observation using sparse non-negative constraints, in: Proc. Int. Conf. on Music Information Retrieval (ISMIR), 2006, pp. 206–212.

[10] A. T. Cemgil, H. J. Kappen, D. Barber, A generative model for music transcription, IEEE Trans. on Audio, Speech and Language Processing 14 (2) (2006) 679–694.

[11] A. T. Cemgil, S. J. Godsill, Efficient variational inference for the dynamic harmonic model, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 271–274.

[12] D. J. C. McKay, Information theory, inference, and learning algorithms, Cambridge University Press, Cambridge, UK, 2003.

[13] G. Casella, C. P. Robert, Monte Carlo statistical methods, 2nd Edition, Springer, New York, NY, 2005.

[14] P. J. Walmsley, S. J. Godsill, P. J. W. Rayner, Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 1999, pp. 119–122.

[15] S. J. Godsill, M. Davy, Bayesian computational models for inharmonicity in musical instruments, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 283–286.

[16] M. Davy, S. J. Godsill, J. Idier, Bayesian analysis of western tonal music, Journal of the Acoustical Society of America 119 (4) (2006) 2498–2517.

[17] E. Vincent, M. D. Plumbley, Fast factorization-based inference for Bayesian harmonic models, in: Proc. IEEE Int. Conf. on Machine Learning for Signal Processing (MLSP), 2006, pp. 117–122.

[18] R. M. Neal, Probabilistic inference using Markov chain Monte Carlo methods, Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto (1993).

[19] D. M. Chickering, D. Heckerman, Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables, in: Proc. Conf. on Uncertainty in Artificial Intelligence (UAI), 1996, pp. 158–168.

[20] C. J. van Rijsbergen, Information retrieval, 2nd Edition, Butterworths, London, UK, 1979.

[21] D. J. Sheskin, Handbook of parametric and nonparametric statistical procedures, 2nd Edition, Chapman & Hall, Boca Raton, FL, 2000.

[22] J.-F. Cardoso, Multidimensional independent component analysis, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998, pp. IV–1941–1944.