PRIOR STRUCTURES FOR TIME-FREQUENCY ENERGY DISTRIBUTIONS

Ali Taylan Cemgil, Paul Peeling, Onur Dikmen, Simon Godsill

Signal Processing and Communications Laboratory, University of Cambridge
Department of Engineering, Trumpington Street, Cambridge, CB2 1PZ, UK
{atc27,php23,od225,sjg}@eng.cam.ac.uk

ABSTRACT

We introduce a framework for probabilistic modelling of time-frequency energy distributions based on correlated Gamma and inverse Gamma random variables. One advantage of the approach is that the resulting class of models are conjugate which makes inference easier. Moreover, both positivity and additivity follow naturally in this framework. We illustrate how generic models (applicable to a broad class of signals) and more specialised models can be designed to model harmonicity, spectral continuity and/or changepoints. We show simulation results that illustrate the potential of the approach on a large spectrum of audio processing applications such as denoising, source separation and transcription.

1. INTRODUCTION

Time-Frequency energy distributions are of central importance in analysis of nonstationary processes, and in particular for audio and acoustical signal analysis. The Gabor transform and STFT (short time Fourier transform) or the MDCT (modified discrete cosine transform) are undoubtly among the most widely used approaches [1, 2, 3]. In all these representations, the signal $\mathbf{y} = (y_1, \dots, y_n, \dots y_N)$ is represented by a linear combination $y_n = \sum_{\nu,\tau} \phi_n^{(\nu,\tau)} s_{\nu,\tau}$ where the set of basis functions ϕ are windowed sinusoidals indexed by the frequency ν and time shift τ. The very popular and visually appealing spectrogram representation is obtained by viewing the log-magnitude of the expansion coefficients $\frac{1}{2}\log s_{\nu,\tau}^2$ as a function of frequencies ν and time indices τ. However, in many applications, the spectrogram is typically viewed as a preprocessing or a feature extraction step.

In recent years, there has been a growing interest in modelling time-frequency energy distributions. One implicit modelling approach has focused on non-negativity of the spectrogram $S = \{s_{\nu,\tau}^2\}$ and enforcing a factorisation as $S = WR$ where both W and R are matrices with positive entries [4, 5]. These have a rough interpretation as a codebook of templates and R is the matrix of activations, somewhat analogous to a musical score. The primary advantage of these methods is computational attractiveness due to fast converging iterative matrix factorization techniques. However, lacking an explicit signal model, it is hard to incorporate prior knowledge and one may have to resort to heuristics, since the construction of the dictionary is entirely data driven. Moreover, this representation is physically unrealistic, since the energy is a quadratic quantity; in general for two sources s_1 and s_2 we have $(s_1 + s_2)^2 \neq {s_1}^2 + {s_2}^2$.

At the other extreme of the spectrum are the dynamical system models which explicitly model the time evolution of the phases, amplitudes and discontinuities [6, 7, 8]. While these class of models are quite powerful and close to reality from a generative perspective, the computational requirements have somewhat limited their use in data intensive applications.

An alternative and often effective approach is to model sources directly in a transform domain. In audio processing, the energy content of a signal is typically time-varying hence it is natural to model audio with a process with a time varying power spectral density on a time frequency plane using switches [9, 2, 10], a histogram [11] or source filter models in cepstral domain [12].

In this paper, we follow a transform domain modelling approach and focus on the following hierarchical source model

$$p(\mathbf{s}|\mathbf{v})p(\mathbf{v}) = \left(\prod_{\nu,\tau} p(s_{\nu,\tau}|v_{\nu,\tau}) \right) p(\mathbf{v})$$

where $\mathbf{s} = s_{1:W,1:T}$ are the collection of transform coefficients $s_{\nu,\tau}$ and $\mathbf{v} = v_{1:W,1:T}$ are the associated variances . To lighten the notation, we will denote each time-frequency atom by $k \equiv (\nu, \tau)$ and write $1 : K \equiv (1 : W, 1 : T)$. In particular, we will assume that the expansion coefficients s_k are conditionally Gaussian and the variances v_k are nonnegative random variables assumed to be distributed by an inverse-Gamma distribution[1]:

$$s_k \sim \mathcal{N}(s_k; 0, v_k I_d) \qquad v_k \sim \mathcal{IG}(v_k, a(\mathbf{v}_{-k}), b(\mathbf{v}_{-k}))$$

where $\mathbf{v}_{-k}$ is the collection of all $\mathbf{v}$ excluding v_k. I_d is an identity matrix that is taken 1×1 for MDCT (real coefficients) and 2×2 for a STFT representation (real and complex coefficients). The inverse-Gamma distribution is the conjugate prior for the variance v of a Gaussian distribution[2]. This fact is the consequence of a simple algebraic observation: when the prior $p(v)$ is inverse-Gamma, the posterior distribution $p(v|s)$ can be represented as an inverse-Gamma distribution since the logarithm of a Gaussian is a polynomial in v^{-1} and $\log v^{-1}$.

Such a model is useful as a building block for many different audio processing applications such as denoising, transcription and source separation. For example, in single channel source separation, we write the observed signal y_k as a superposition of J source models $y_k = \sum_{j=1}^{J} s_{k,j}$. If all the latent variances $\mathbf{v}_j$ for $j = 1 \dots J$ would have been known, by straightforward application of the Bayes theorem, the source coefficients can be reconstructed in closed form by

$$\langle s_{k,j} \rangle \quad = \quad \kappa_{k,j} y_k \qquad \left\langle s_{k,j}^2 \right\rangle - \langle s_{k,j} \rangle^2 = v_{k,j}(1 - \kappa_{k,j})$$

where $\kappa_{k,j} = v_{k,j}/(\sum_{j'} v_{k,j'})$ and $\langle f(s) \rangle$ denotes the expectation of the function $f(s)$ under the posterior distribution $p(s|y,v)$.

[1] $\mathcal{IG}(v; a, z) \equiv \exp((a+1)\log v^{-1} - z^{-1}v^{-1} + a \log z^{-1} - \log\Gamma(a))$ with $\Gamma(a)$ being the Gamma (generalized factorial) function.

[2] $\mathcal{N}(s; \mu, v) \equiv \exp\left(-(s-\mu)^2 v^{-1}/2 + \log v^{-1}/2 - \log(2\pi)/2\right)$

Note that $\kappa_{k,j}$ are nonnegative such that $\sum_j \kappa_{k,j} = 1$ for all k. Intuitively, this means that each reconstructed source $s_{k,j}$ gets a fraction $\kappa_{k,j}$ of the observation y_k. We name κ as *responsibilities* (also know as *Wiener filter factors*). In reality, of course, the variances $v_{k,j}$ are unknown but we can postulate realistic prior structures by considering physical properties such as harmonicity, damping e.t.c. Another appealing property of this approach is that if we integrate out the unknown expansion coefficients $\mathbf{s}$ analytically, we obtain a model that is *additive* in variances as

$$\int d\mathbf{s} p(\mathbf{y}|\mathbf{s}_{1:J}) p(\mathbf{s}_{1:J}|\mathbf{v}_{1:J}) p(\mathbf{v}_{1:J}) = \mathcal{N}\left(\mathbf{y}; 0, \sum_{j=1}^{J} \mathbf{v}_j\right) p(\mathbf{v}_{1:J})$$

Denoising is a special case of the above model: we just assign one source (e.g. $j = J$) to be the "noise". For example, when the noise is additive and the stationary Gaussian we have $v_k = v$ for all k where v is the noise variance to be estimated. Similarly, transcription can be formulated by postulating additional indicator variables $\mathbf{r} \equiv r_{1:T}$ upper layers in the hierarchy $p(v_{1:W,1:T}|r_{1:T})\ p(r_{1:T})$ where each r_τ is a discrete variable that selects a prior structure. By computing the marginal MAP configuration $\arg\max_{\mathbf{r}} \int d\mathbf{v}\ p(\mathbf{x}|\mathbf{v})\ p(\mathbf{v}|\mathbf{r}) p(\mathbf{r})$, we can find the most likely score, e.t.c.

In all these applications, the main modelling issue is finding appropriate prior structures on variances and this is the focus of this paper.

2. GAMMA MODELS FOR VARIANCES

One possible approach for defining a prior distribution over variances is to define a Gaussian process $\{l_\tau\}_{\tau=1,2,\dots}$, e.g. a random walk, in the log domain as

$$l_\tau \quad \sim \quad \mathcal{N}(l_\tau; l_{\tau-1}, q^{-1}) \qquad\qquad v_\tau = \exp(l_\tau)$$

It is easy to see that v_τ will be strictly positive, the distribution of $p(v_\tau|l_{\tau+1}, l_{\tau-1})$ will be log-normal and v_τ and $v_{\tau-1}$ will be marginally positively correlated. However, the joint distribution will have non-convex terms such as $\log^2 v_\tau$ that will render inference harder and it will be necessary to resort to generic Monte Carlo integration techniques. While this is in principle not an obstacle, it is desirable to construct a model that retains some form of conjugacy for fast inference since in audio applications that we are interested in, K will be very large and $p(\mathbf{v})$ will be typically embedded into a hierarchical model.

It is possible to define a Markov chain on inverse-Gamma random variables in a straightforward way by $v_\tau|v_{\tau-1} \sim \mathcal{IG}\ (v_\tau; a, v_{\tau-1}/a)$. The full conditional distribution $p(v_\tau|v_{\tau-1}, v_{\tau+1})$ is conjugate, i.e. it is also inverse-Gamma. However, by this construction it is not possible to attain positive correlation between v_τ and $v_{\tau-1}$. The basic idea is to introduce latent auxiliary variables z_τ between v_τ and $v_{\tau-1}$ such that when z_τ are integrated out we restore positive correlation between v_τ and $v_{\tau-1}$ while retaining conjugacy [13]. We define an *Inverse-Gamma Markov* chain (IGMC) for $k = 1 \dots K$ as follows

$$v_\tau|z_\tau \sim \mathcal{IG}(v_\tau; a, z_\tau/a) \qquad z_{\tau+1}|v_\tau \sim \mathcal{IG}(z_{\tau+1}; a_z, v_\tau/a_z)$$

Here, z_τ are auxiliary variables that ensure the full conditionals $p(v_\tau|z_\tau, z_{\tau+1})$ and $p(z_\tau|v_\tau, v_{\tau-1})$ are inverse-Gamma. An equivalent "mixed" construction is obtained by letting $\lambda_\tau = 1/v_\tau$

$$\lambda_\tau|z_\tau \sim \mathcal{G}(\lambda_\tau; a, z_\tau/a) \qquad z_{\tau+1}|\lambda_\tau \sim \mathcal{IG}(z_{\tau+1}; a_z, 1/(a_z\lambda_\tau))$$

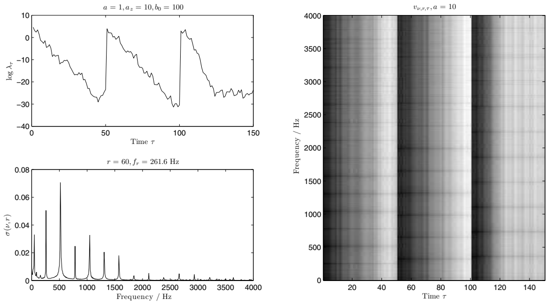

Figure 1: A typical trajectory λ_τ sampled from the changepoint process (1) (upper-left). Typical spectral templates from pitched musical instruments (lower-left). A time frequency energy distribution sampled from Eq.(2) scaled by λ_τ (right).

where $\mathcal{G}$ is a *Gamma* distribution[3]. By integrating out over the auxiliary variable z_τ we obtain the effective transition kernel of the Markov chain, $p(v_\tau|v_{\tau-1})$, which has positive correlation for various shape parameters a_z and a. The absolute value of a_z and a control the strength of the correlation and the ratio a_z/a controls the skewness. For $a_z/a < 1$ ($a_z/a > 1$), the probability mass is shifted towards the interval $v_\tau < v_{\tau-1}$ ($v_\tau > v_{\tau-1}$) hence, typical trajectories from a IGMC will exhibit a systematic negative (positive) drift. The chain structure can be generalised in a straightforward manner to 2-D random fields with arbitrary connection topology [13].

Positive correlations between variances can be used for modelling harmonic continuity observed for many acoustical sources. Using this strategy, we can design *generic* prior structures that are potentially useful for a broad class of audio signals. For example, we can tie the energies in a time-frequency plane across time (to model harmonic continuity) or across frequency (to model impulsive sources) or both.

The drift property can be exploited to model damping effects and onsets. This path of modelling can lead to *specific* prior structures. We first extend the chain to a changepoint model as follows: We introduce latent discrete variables o and let

$$\lambda_\tau|z_\tau, o_\tau \sim \left\{ \begin{array}{ll} \mathcal{G}(\lambda_\tau; a, z_\tau/a) & o_\tau \neq \text{onset} \\ \mathcal{G}(\lambda_\tau; a, b_0/a) & o_\tau = \text{onset} \end{array} \right. \tag{1}$$

when an onset occurs, the chain is reinitialized from a prior. Given λ_τ as a decaying positive process with occasional reinitializations, we can interpret it as an average energy. Conditioned on λ, we can define a model for time varying spectral energy for each time-frequency bin

$$v_{\nu,\tau}|r_\tau, \lambda_\tau \quad \sim \quad \mathcal{IG}(v_{\nu,\tau}; a/2, 2/(\lambda_\tau \sigma(\nu; r_\tau)a)) \tag{2}$$

Here, $\sigma(\nu; r)$ is a positive spectral template function which represents the expected distribution of energy among frequency bins ν and r is an index variable. For transcription, we can choose r to correspond to individual pitch or chord labels. An application of this model to score following is reported elsewhere [14]. A spectral template for a pitched instrument and a typical sample generated from the model is shown in Figure 1.

[3] $\mathcal{G}(\lambda; a, b) \equiv \exp((a-1)\log\lambda - b^{-1}\lambda + a\log b^{-1} - \log\Gamma(a))$

3. INFERENCE

Exact inference in Gamma chains is difficult since the marginals have complicated closed form expressions. Fortunately, various powerful numerical integration methods can be employed, notably based on sampling (Monte Carlo-stochastic) or analytic approximation (Variational-deterministic). Here, we focus on sampling based methods, in particular sequential Monte Carlo [15] and the Gibbs sampler (e.g., see [16]). Algorithmically similar variational methods to Gibbs sampling, notably variational Bayes [17], can be derived easily by exploiting conjugacy [13].

The Gibbs sampler [16] is a particular Monte Carlo method that relies on generating samples $\{\boldsymbol{\xi}^{(t)}\}_{t=1,2,\ldots}$ via simulation of a Markov chain with the desired target density $p(\boldsymbol{\xi})$. The algorithm proceeds by sampling each random variable $i \in \mathcal{V}$ from the full conditional distribution $p(\xi_i|\xi_{-i}^{(t-1)})$ where $-i \equiv \mathcal{V} \setminus i$. For a Gamma chain, the expressions are particularly simple; due to the local connectivity structure, for each i, the conditional distribution depends only on immediate neighbours of i. Moreover, since the model is conjugate, this expression is readily available in closed form as $\xi_i^{(t)} \sim p(\xi_i|\xi_i^{(t-1)}) = \mathcal{IG}(\xi_i; \alpha_i, \beta_i)$ where α_i and β_i are functions of $\xi_i^{(t-1)}$.

One problem with the Gibbs sampler is convergence speed. In practice, the chain takes a prohibitively long time to converge and occasionally becomes trapped in local maxima. To speed up convergence, special strategies, such as tempering (gradually changing the target density) or blocking (grouping random variables) need to be employed [16]. Additionally, for time series models the algorithm is inherently a batch processing method. An alternative set of methods, known as sequential Monte Carlo (SMC) [15] are based on sequential importance sampling (also known as particle filtering) have proved to be quite powerful, especially for time series models. SMC methods have the advantage that they are inherently online, simple to implement and quite flexible. In many applications, by merely increasing the amount of computation, the estimation results tend to improve.

Sequential Monte Carlo methods [15] approximate a target density (often the posterior of a time series model) $p(\mathbf{x}_{0:K}|y_{0:K})$ of a hidden Markov process $\mathbf{x}_{0:K}$ as a set of N particle trajectories $\{\mathbf{x}_{0:K}^{(i)}, i = 1, \ldots, N\}$ which are drawn from a importance function π, which allows the normalised importance weights $\tilde{w}_k^{(i)} = w_k^{*(i)} / \sum_{i=1}^{N} w_k^{*(i)}$, where

$$w_k^{*(i)} = \frac{p(y_{0:k}|\mathbf{x}_{0:k})p(\mathbf{x}_{0:k})}{\pi(\mathbf{x}_{0:k}|y_{0:k})} = \frac{p(y_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})}{\pi(\mathbf{x}_k|\mathbf{x}_{0:k-1}, y_{0:k})} w_{k-1}^{*(i)}$$

to be computed sequentially. The particle trajectories $\mathbf{x}_{0:k}^{(i)} \equiv \left(\mathbf{x}_{0:k-1}^{(i)}, \mathbf{x}_k^{(i)}\right)$ are constructed sequentially by sampling the importance function: $\mathbf{x}_k^{(i)} \sim \pi(\mathbf{x}_k|\mathbf{x}_{0:k-1}, y_{0:k})$. To keep the approximation accurate over time (by keeping the variance of the importance weights bounded), periodical resampling steps are applied. Here, particles are drawn according to a distribution based on the importance weights and the importance weights are set to $1/N$. The *bootstrap filter* [15] is the simplest SMC method, where the importance function is simply the prior, i.e. $\pi(\mathbf{x}_{0:K}|y_{0:K}) = p(\mathbf{x}_{0:K})$ and resampling is carried out at every iteration by copying each particle $N_k^{(i)}$ times according to a multinomial distribution with parameters $\tilde{w}_k^{(i)}$.

However, whilst the bootstrap filter works in many time series analysis problems, it is well known that when the latent state dimension high, it can quickly become ineffective. To render the SMC approach feasible many improvements are needed. One such improvement is "Rao-Blackwellisation", i.e. exploiting model structure for reducing the sampling dimension by analytically integrating out some of the variables, conditioned on some others. The model described in equations (1) and (2) is a such one; given λ, we can integrate out the variances $\mathbf{v}$ analytically so only indicators o, indices r and scale variables λ need to be sampled.

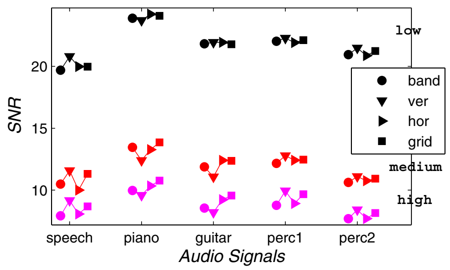

Figure 2: Signal-to-Noise ratio results for reconstructions obtained from the audio clips in `low, medium,high` noise conditions.

4. RESULTS

In this section we present results with *generic* models for denoising and single channel separation. We illustrate *specific* models, that aim to model details of harmonic structure and onsets for transcription and chord recognition. The audio extracts will be available online at `http://www-sigproc.eng.cam.ac.uk/~php23/publications/WASPAA`.

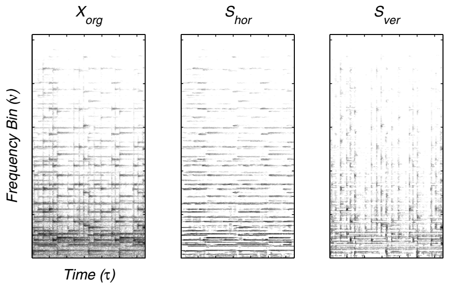

Figure 3: Single channel Source Separation example, left to right, log-MDCT coefficients of the original signal and reconstruction with horizontal and vertical IGMC models. The model seems to be able to separate transients and harmonic components.

In denoising simulations, 5 audio clips are used. Independent white Gaussian noise with variance $r \sim \mathcal{IG}(r; a_r, b_r)$ is added onto the MDCT source coefficients $(s_{(\nu,\tau)}^{true})$ and $x_{(\nu,\tau)}$ are obtained. The noise has the same characteristics in time domain, because MDCT is an orthonormal linear transform. Four IGMC topologies are used as priors: `vertical` (energies tied across frequency), `horizontal` (tied across time), `grid` (both) and `band` (an auxiliary variable for each frequency bin). The SNR results are presented in Figure 2.

We have used IGMCs for a single channel source separation as explained in Section 2. We used a vertical and a horizontal IGMC

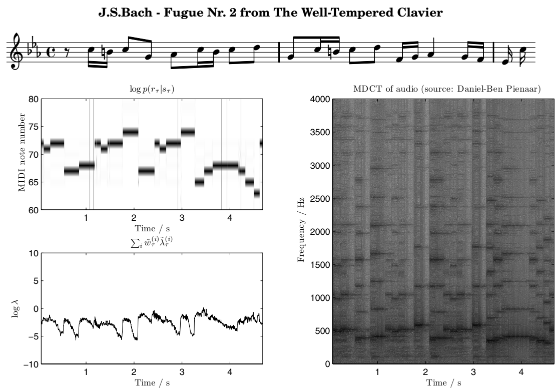

Figure 4: Monophonic piano transcription. The marginal filtering density over r_τ and the minimum mean-squared-error estimate of λ_τ are shown. The note pitch errors in the transcription can be eliminated by imposing a prior on how notes transition over time.

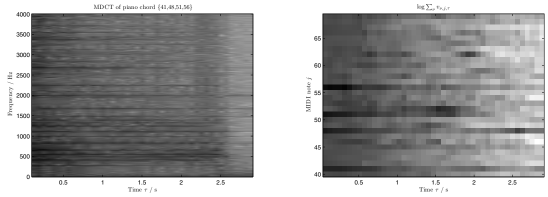

Figure 5: MDCT of a piano chord (left). Estimated $\lambda_\tau(r)$ (Midinotes $r = 40 \ldots 69$). The high energy onsets and subsequent decay of the four piano notes are clearly distinguishable (right).

to model the transients and harmonic components of an instrument sound separately. The results for a piano sound are shown in Figure 3.

If the variances are drawn according to the specific model (2) for all times τ, then we obtain an exact expression for the likelihood of the transform coefficients in frame τ as a Student-T distribution by integrating over v, see [14] for details. If we directly observe the transform coefficients, for instance for a monophonic music extract when we have only a single source $J = 1$, then the inference of the filtering density over the unknown variables $p(r_\tau, \lambda_\tau|\mathbf{s}_\tau)$ can be obtained by SMC. Figure 4 demonstrates the performance of the bootstrap filter with $N = 500$ particles.

These templates are also useful for guided source separation. Figure 5 shows the performance of Gibbs' sampler when jointly infering the energy distribution of 30 simultaneous sources at different note pitches, for a single piano chord with no changepoints. When changepoints are included, needed in a polyphonic transcription application for instance, Gibbs' sampler becomes trapped in local maxima, and more elaborate inference schemes, which will form the basis of future work, are necessary.

5. CONCLUSIONS

We introduce a framework for probabilistic modelling of time-frequency energy distributions based on correlated Gamma and inverse Gamma random variables. The approach is quite flexible in modelling a range of phenomena known to be present in general audio, such as harmonicity, spectral continuity, onsets e.t.c. Both positivity and additivity follow naturally in this framework and resulting models turn out to be conjugate, a technical condition which when satisfied renders inference easier. Both generic and more specific models can be designed. Using standard inference techniques such as sequential Monte Carlo, Gibbs sampling or variational Bayes, we show simulation results that illustrate the potential of the approach on denoising, source separation and transcription. Future work will include detailed testing the viability of this approach in terms of quality and computational cost in comparison to alternative approaches.

6. REFERENCES

[1] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[2] P. J. Wolfe, S. J. Godsill, and W. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation," *Journal of the Royal Statistical Society, Series B*, vol. 66, no. 3, pp. 575–589, August 2004.

[3] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 302–312, May 2004.

[4] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2003, pp. 177–180.

[5] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music using sparse coding," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 179–196, January 2006.

[6] Y. Qi, T. P. Minka, and R. W. Picard, "Bayesian spectrum estimation of unevenly sampled nonstationary data," MIT Media Lab, Tech. Rep. Vismod-TR-556, 2002.

[7] A. T. Cemgil and S. J. Godsill, "Efficient Variational Inference for the Dynamic Harmonic Model," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2005, pp. 271– 274.

[8] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible Bayesian approach," *IEEE transactions on Speech, Audio and Language Processing*, vol. 15, no. 4, pp. 1283–1295, May 2007.

[9] M. Reyes-Gomez, N. Jojic, and D. Ellis, "Deformable spectrograms," in *AI and Statistics Conference*, Barbados, 2005.

[10] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani, "Sparse regression with structured priors: Application to audio denoising," in *Proc. ICASSP*, Toulouse, France, May 2006.

[11] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2005, pp. 17–20.

[12] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Hawaii, USA, 2007.

[13] A. T. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," in *Submitted*, 2007.

[14] P. Peeling, A. T. Cemgil, and S. J. Godsill, "A probabilistic framework for matching music representations," in *Submitted*, 2007.

[15] A. Doucet, N. de Freitas, and N. J. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

[16] J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer, 2004.

[17] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.