Sparse Linear Regression With Structured Priors and Application to Denoising of Musical Audio

Cédric Févotte, Bruno Torrésani, Laurent Daudet, and Simon J. Godsill

***Abstract*—We describe in this paper an audio denoising technique based on sparse linear regression with structured priors. The noisy signal is decomposed as a linear combination of atoms belonging to two modified discrete cosine transform (MDCT) bases, plus a residual part containing the noise. One MDCT basis has a long time resolution, and thus high frequency resolution, and is aimed at modeling tonal parts of the signal, while the other MDCT basis has short time resolution and is aimed at modeling transient parts (such as attacks of notes). The problem is formulated within a Bayesian setting. Conditional upon an indicator variable which is either 0 or 1, one expansion coefficient is set to zero or given a hierarchical prior. Structured priors are employed for the indicator variables; using two types of Markov chains, persistency along the time axis is favored for expansion coefficients of the tonal layer, while persistency along the frequency axis is favored for the expansion coefficients of the transient layer. Inference about the denoised signal and model parameters is performed using a Gibbs sampler, a standard Markov chain Monte Carlo (MCMC) sampling technique. We present results for denoising of a short glockenspiel excerpt and a long polyphonic music excerpt. Our approach is compared with unstructured sparse regression and with structured sparse regression in a single resolution MDCT basis (no transient layer). The results show that better denoising is obtained, both from signal-to-noise ratio measurements and from subjective criteria, when both a transient and tonal layer are used, in conjunction with our proposed structured prior framework.**

***Index Terms*—Bayesian variable selection, denoising, Markov chain Monte Carlo (MCMC) methods, nonlinear signal approximation, sparse component analysis, sparse regression, sparse representations.**

I. Introduction

MOST commonly used representations of audio signals, for example for coding or denoising purposes, make use of local Fourier bases. Among these, lapped orthogonal transforms [1] such as the modified discrete cosine transform (MDCT) are a popular choice since they provide an orthonormal decomposition without blocking effects, and have fast implementations based on the fast Fourier transform (FFT). Atoms corresponding to the MDCT transform of a signal of length $N = l_{\text{frame}} \times n_{\text{frame}}$ and a frame length l_{frame} are defined as

$$\begin{aligned}\Phi_{(q,n)}(t) &= w\left(t-(n-1)l_{\text{frame}}\right) \\ &\quad \times \cos\left[\frac{\pi}{l_{\text{frame}}}\left(t-(n-1)l_{\text{frame}}+\frac{l_{\text{frame}}+1}{2}\right)\left(q-\frac{1}{2}\right)\right]\end{aligned} \tag{1}$$

with $q = 1,\ldots,l_{\text{frame}}$ being a frequency index and $n = 1,\ldots,n_{\text{frame}}$ being a frame index. w is a window of length $2l_{\text{frame}}$ that meets symmetry and energy-preservation constraints. Decomposing a signal $x \in \mathbb{R}^{N\times 1}$ onto the dictionary $\{\Phi_{(q,n)}\}$ is simply done with dot products: $x = \sum_k \tilde{s}_k \Phi_k$ with $\tilde{s}_k = \langle \Phi_k, x\rangle$.[1] In other words, as with any orthonormal transform, the *synthesis* coefficients $\{\tilde{s}_k\}$ are identical to the *analysis* coefficients.

A main reason for the success of such expansions is the fact that they are *sparse*: the signal is characterized by a small number of coefficients, the remaining ones being either equal to zero or at least numerically negligible. However, for most audio signals, using the MDCT with a constant frame size l_{frame} does not provide approximations that are sufficiently sparse, i.e., where most of the coefficients are small and can be neglected. Typically, one would use a frame size of 23 ms (1024 coefficients at 44.1-kHz sampling rate), which is adequate for the tonal part of the signal. However, there might also be a number of so-called "transient" components, e.g., at attacks of percussive notes, that evolve on much smaller time scales, typically a few milliseconds. For audio coding purposes, this leads to an overly large number of coefficients to encode, and current state-of-the-art transform coders such as MPEG 2 Advanced Audio Coder (AAC [2]) switch to a shorter frame size at transients. Similarly, for denoising purposes, with a model of the kind

$$x = \sum_{k=1}^{N} \tilde{s}_k \Phi_k + e \tag{2}$$

using a single frame size l_{frame} results in a large number of small coefficients at the attacks, that are thresholded to zero together with the noise term e. This leads to a loss of percussive

[1]Here, k replaces (q,n) and $\Phi_k = [\Phi_k(1),\ldots,\Phi_k(N)]^T$.

Manuscript received June 30, 2006; revised August 7, 2007. Part of this work was done while C. Févotte was a Research Associate with Cambridge University Engineering Department, Cambridge, U.K., and also while visiting the Laboratoire d'Analyse, Topologie et Probabilités, Université de Provence, Marseille, France. This work was supported in part by the European Commission funded Research Training Network HASSIP under Grant HPRN-CT-2002-00285. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Te-Won Lee.

C. Févotte is with CNRS-GET/Télécom Paris (ENST), 75014 Paris, France (e-mail: fevotte@tsi.enst.fr).

B. Torrésani is with the Laboratoire d'Analyse, Topologie et Probabilités, Université de Provence, 3453 Marseille cedex 13, France (e-mail: torresan@cmi.univ-mrs.fr).

L. Daudet is with the Lutheries Acoustique Musique/Institut Jean Le Rond d'Alembert, Université Pierre et Marie Curie-Paris 6, 75015 Paris, France (e-mail: daudet@lam.jussieu.fr).

S. J. Godsill is with the Signal Processing Group, Cambridge University Engineering Department, Cambridge CB2 1PZ, U.K. (e-mail: sjg@eng.cam.ac.uk).

Digital Object Identifier 10.1109/TASL.2007.909290

strength, typical of many denoising algorithms. Again, adaptive switching of the frame size is possible but does not reflect the additive nature of sounds; it is indeed quite common to have steady tones together with percussive transient signals, and it is not desirable that the analysis of one component introduces a bias in the analysis of the second. Therefore, one needs overcompleteness, with a basis of atoms $\{\Phi_{1,k}\}$ having long frame length l_{frame1} together with a basis of atoms $\{\Phi_{2,k}\}$ having short frame length l_{frame2}. Our signal model now becomes

$$x = \sum_{k=1}^{K} \tilde{s}_k \Psi_k + e \tag{3}$$

where $K = 2T$ and the atoms Ψ_k belong to the dictionary

$$\mathcal{D} = \{\Phi_{1,k}, k = 1, \ldots, N\} \cup \{\Phi_{2,k}, k = 1, \ldots, N\}. \tag{4}$$

Because of the nonuniqueness of the expansion (3), finding the expansion coefficients $\tilde{s}_k$ above involves more than just computing scalar products, and an additional selection criterion has to be introduced. We choose to emphasize sparsity: among all possible decompositions one has to find one that is (nearly) optimally sparse, according to some prespecified sparsity criteria. This problem is often referred to as sparse linear regression. Numerous practical methods have been developed for finding sparse approximations in overcomplete dictionaries, with different computational complexities. Seminal contributions include Matching Pursuit [3], Basis Pursuit [4], the FOCUSS algorithm and its regularized version [5], [6], as well as Figueiredo's algorithm [7]. Computational complexity of these algorithms can be reduced when the dictionary is the union of two orthonormal bases, as described in [8] and [9]. However, none of these methods considers dependencies between significant coefficients, and this often results in a number of isolated large coefficients that are nearly equally well represented in both bases. In the reconstructed signal, after thresholding small coefficients, these isolated components give rise to so-called "musical noise." Clearly, in such situations, one would like to favor clusters of coefficients rather than isolated coefficients, along spectral lines for the tonal part (the amplitude-varying harmonics), or across adjacent frequency bins at a given time frame for the transient part (attacks). This strategy will penalize those isolated coefficients that have no physical meaning. We will term such additional constraints *structure*, and when interpreted in a probabilistic setting they will be used to define a *structured prior distribution* over the basis coefficients.

In [10] and [11], structural information in a similar tones + transients + noise model was enforced through the use of hidden Markov chains for time persistency in the tonal MDCT layer, and hidden Markov trees for the transient part in the transient discrete wavelets layer. However, in this case, the estimation of the two layers is sequential (first the tonal part is estimated and subtracted, then the transient part is estimated); this sometimes leads to a biased estimate in the relative importance of these layers. References [12] and [13] study the use of structured priors (both vertical, horizontal, and spatial in the time–frequency plane) in an overcomplete Gabor regression framework, operating however in one single time–frequency resolution.

The goal of this paper is to present a framework for simultaneous estimation of both layers, while imposing structural constraints on the set of selected coefficients with the help of Markov chains: "horizontal structures" for the tonal layers and "vertical structures" for the transients layer. Unlike prior works implementing horizontal and vertical time–frequency structures, our approach allows one to avoid the sequential approach used in [10], [11], and [14] and estimate tonal and transient layers simultaneously. Within our Bayesian setting, inference of the targeted expansion coefficients is done through Markov chain Monte Carlo (MCMC) inference, using similar inference methodology as in [9], [12], [13], and [15]. Though these computational methods are more demanding than their expectation-maximization (EM)-like counterparts, they offer increased robustness (reduced problems of convergence to local minima) and a complete Monte Carlo description of the posterior density of the parameters. Preliminary results can be found in [16]; here, we propose significant improvements to the signal model (particular care is brought to modeling of the initial probabilities of the Markov chains, and the use of frequency profiles is investigated), we include additional technical details (including efficient sampling schemes for the Markov chain parameters) and present detailed results.

Although we focus here on the single application of music denoising, which allows a rather straightforward quantitative evaluation, we shall emphasize that what we describe in this paper is a semantic object-based representation of musical audio, where the sound objects are transient and tonal components. It provides a *mid-level representation* of sound from which many audio signal processing tasks could benefit, such as very low bit-rate audio coding, automatic music transcription, and more general processing tasks such as source separation, interpolation of missing data, and removal of impulse noise.

This paper is organized as follows. In Section II, we detail the signal model and develop the explicit form of the structured priors. The estimation technique is presented in Section III, where a Gibbs sampler-based MCMC scheme is described. Section IV presents denoising results over a short Glockenspiel excerpt and over a longer polyphonic music excerpt. We compare the benefits of our approach with respect to overcomplete unstructured sparse regression on the one hand, and with sparse regression in a single long time resolution MDCT basis with horizontal structures (no transient layer) on the other hand. Finally, Section V is devoted to conclusions and perspectives.

II. Signal Model

We here formalize the concepts introduced above and specify more precisely the functional model and the priors on all the parameters of our model.

A. Functional Model

Starting from a couple of MDCT bases $\{\Phi_{1,k}, k=1,\dots N\}$ and $\{\Phi_{2,k}, k=1,\dots N\}$ of $\mathbb{R}^N$ [see (1)], and using the dictionary defined in (4), we rewrite the model (3) as

$$x = \sum_{k=1}^{N} \tilde{s}_{1,k}\Phi_{1,k} + \sum_{k=1}^{N} \tilde{s}_{2,k}\Phi_{2,k} + e. \tag{5}$$

Note that this three-layer model is similar to the sines + transients + noise models used in many low bit-rate parametric audio coders [17]. In essence, in the generative model described by (5), e is an error term, i.e, the noise term, that will be modeled as Gaussian white noise with variance σ^2.[2] A central ingredient of the model to be presented is the fact that the two vectors $\tilde{s}_1 = [\tilde{s}_{1,1},\dots,\tilde{s}_{1,N}]^T$ and $\tilde{s}_2 = [\tilde{s}_{2,1},\dots,\tilde{s}_{2,N}]^T$ (which generate respectively the tonal and transients layers) are *sparse*, i.e., most coefficients $\tilde{s}_{i,k}$ vanish, while the noise term is *dense* and does not admit any sparse expansion with respect to the dictionary. The signal model will also assume some *structure* in the coefficient domain that will be expressed in terms of suitable prior distributions, as we describe below.

In the following, we will use the matrix notation $\Phi_i = [\Phi_{i,1},\dots,\Phi_{i,N}] \in \mathbb{R}^{N\times N}$, $i=1,2$. Φ_1 is a MDCT basis with long time resolution l_{frame1} and thus high-frequency resolution (aiming at representing tonals), $\Phi_2 \in \mathbb{R}^{N\times N}$ is an MDCT basis with short time resolution l_{frame2} (aiming at representing transients). The index $k=1,\dots,N$ will sometimes be more conveniently replaced by (q,n) with $q=1,\dots,l_{\text{frame}i}$ being a frequency index (where $l_{\text{frame}i}$ is either l_{frame1} or l_{frame2}) and $n=1,\dots,n_{\text{frame}i}$ being a frame index, with $l_{\text{frame}i}\times n_{\text{frame}i} = N$ and such that $k=(n-1)l_{\text{frame}i}+q$.

B. Coefficients Priors

Sparsity is explicitly modeled in the coefficients through introduction of indicator random variables $\gamma_{i,k}\in\{0,1\}$ attached to all coefficients, and use of the following hierarchical prior for $\tilde{s}_{i,k}$, $i=1,2$, $k=1,\dots,N$

$$p(\tilde{s}_{i,k}|\gamma_{i,k},v_{i,k}) = (1-\gamma_{i,k})\delta_0(\tilde{s}_{i,k}) + \gamma_{i,k}\mathcal{N}(\tilde{s}_{i,k}|0,v_{i,k}) \tag{6}$$

$$p(v_{i,k}|\alpha_i,f_{i,k}) = \mathcal{IG}(v_{i,k}|\alpha_i,f_{i,k}) \tag{7}$$

where $\mathcal{N}(u|\mu,v)$ and $\mathcal{IG}(u|\alpha,\beta)$ are the normal and inverted-Gamma distributions defined in Appendix I, and $\delta_0(u)$ is the Dirac delta function. As can be seen from the above, when $\gamma_{i,k}=0$, $\tilde{s}_{i,k}$ is set to zero and sparsity is precisely enforced for that coefficient; when $\gamma_{i,k}=1$, $\tilde{s}_{i,k}$ has a normal distribution conditional upon $v_{i,k}$, which is itself given a conjugate inverted-Gamma prior.[3]

[2]Colored or non-Gaussian noise can routinely be incorporated into the same framework, but they will affect the computational efficiency of the coefficient sampling steps, as discussed in Section V.

[3]If a parameter θ is observed through data x via the likelihood $p(x|\theta)$, the prior $p(\theta)$ is said to be conjugate when $p(\theta)$ and $p(\theta|x)\propto p(x|\theta)p(\theta)$ belong to the same family of distributions. Here, $v_{i,k}$ is observed through $\tilde{s}_{i,k}$ (when $\gamma_{i,k}=1$) via $p(\tilde{s}_{i,k}|\gamma_{i,k}=1,v_{i,k})$, its prior is $\mathcal{IG}(v_{i,k}|\alpha_i,f_{i,k})$ and its posterior $p(v_{i,k}|\gamma_{i,k}=1,\tilde{s}_{i,k},\alpha_i,f_{i,k})$, given in (19), is also inverted-Gamma. Conjugate priors belonging to families of distributions easy to sample from or whose moments are analytically available are often used in Bayesian estimation, because they allow to keep the inference tractable [18].

This sparsity-enforcing coefficient implies that the marginal distribution of any given coefficient is a mixture of a Dirac point mass at zero and a Student-t distribution.[4]

$f_{i,k}$ is a parametric frequency profile whose expression is given by

$$f_{i,k} = f_{\lambda_i,\nu_i,\eta_i}(q) = \frac{\lambda_i}{\left(1+\left(\frac{q-1}{\eta_i}\right)^{\nu_i}\right)} \quad \text{where } k=(q,n). \tag{8}$$

This frequency profile aims at modeling the expected energy distribution of audio signals, which is typically decreasing with frequency. Here, we chose a frequency shaping based on the frequency response of a Butterworth low-pass filter, where λ_i acts as a gain or scale parameter, η_i acts as a cutoff frequency, and ν_i acts as the filter order. However, any other profile can be chosen (and readily fits in the proposed framework). Apart from the frequency profiles, the model defined by (6) and (7) is similar to the sparse prior used in [9].

C. Indicator Variable Priors

The sequences of coefficients $\tilde{s}_1$ and $\tilde{s}_2$ are each modeled as independent conditionally upon γ_1 and γ_2. The *structural* properties of the prior are obtained through dependent prior distributions over the binary indicator variables $\{\gamma_{i,k}\}$. As discussed above, the dependency is either across time for the first (tonal) basis or across frequencies for the second (transient) basis. Below, we use the conventional representation for the time–frequency plane, and refer to the time axis as the *horizontal axis* and the frequency axis as the *vertical axis*.

1) *"Horizontal" Markov model for tonals*: In order to model persistency in time of time–frequency coefficients corresponding to tonal parts, we give a horizontal prior structure to the indicator variables in the first basis. For a fixed frequency index q, the sequence $\{\gamma_{1,(q,n)}\}_{n=1,\dots,n_{\text{frame1}}}$ is modeled by a two-state first-order Markov chain with transition probabilities $P_{1,00}$ and $P_{1,11}$, assumed equal for all frequency indices $q=1,\dots,l_{\text{frame1}}$. The initial distribution $\pi_1 = P(\gamma_{1,(q,1)}=1)$ of each chain is taken to be its stationary distribution (see remark below), namely

$$\pi_1 = \frac{1-P_{1,00}}{2-P_{1,11}-P_{1,00}} \text{ and } (1-\pi_1) = \frac{1-P_{1,11}}{2-P_{1,11}-P_{1,00}} \tag{9}$$

[4]Indeed, $v_{i,k}$ can be "integrated out" of (6) as follows:

$$\begin{aligned} &p(\tilde{s}_{i,k}|\gamma_{i,k}=1,\alpha_i,f_{i,k}) \\ &\quad = \int_{v_{i,k}} p(\tilde{s}_{i,k},v_{i,k}|\gamma_{i,k}=1,\alpha_i,f_{i,k})dv_{i,k} \\ &\quad = \int_{v_{i,k}} \mathcal{N}(\tilde{s}_{i,k}|0,v_{i,k})\mathcal{IG}(v_{i,k}|\alpha_i,f_{i,k})dv_{i,k} \\ &\quad = t(\tilde{s}_{i,k}|2\alpha_i,\sqrt{f_{i,k}/\alpha_i}). \end{aligned}$$

where $t(u|\alpha,\lambda)$ is the Student t density, defined in Appendix I. The hierarchical formulation (6), (7) is preferred because the auxiliary variable $v_{i,k}$ allows to update $\tilde{s}_{i,k}$ easily, by alternatively updating $\tilde{s}_{i,k}$ conditionally upon $v_{i,k}$ and $v_{i,k}$ upon $\tilde{s}_{i,k}$, as shown in Section III. Updating $\tilde{s}_{i,k}$ directly from its Student-t prior formulation would require more elaborate strategies.

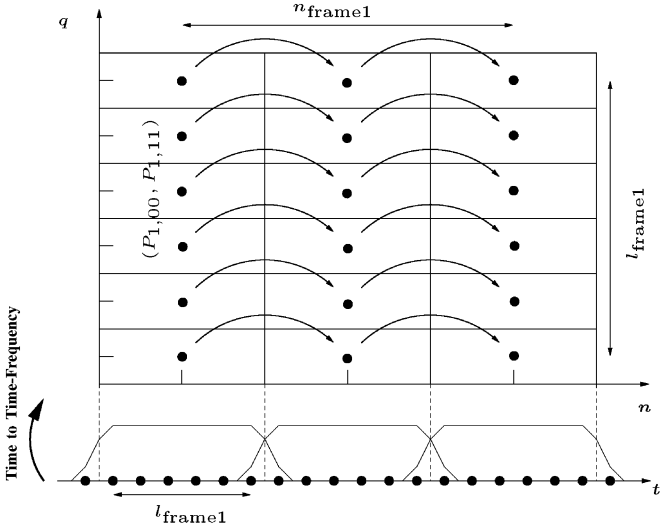

Fig. 1. This figure illustrates the tonal model. Each square of the time–frequency tiling corresponds to a MDCT atom. To each atom corresponds an indicator variable $\gamma_{1,k}$ which controls whether this atom is selected ($\gamma_{1,k} = 1$) or not ($\gamma_{1,k} = 0$) in the signal expansion described in (5). The set of indicator variables γ_1 is modeled as "horizontal and parallel" Markov chains of order 1, with common transition probabilities $P_{1,00}$ and $P_{1,11}$, and with initial probability taken as its equilibrium value.

The tonal model is illustrated in Fig. 1. The transition probabilities are estimated and given Beta priors $\mathcal{B}(P_{1,00}|\alpha_{P_{1,00}}, \beta_{P_{1,00}})$ and $\mathcal{B}(P_{1,11}|\alpha_{P_{1,11}}, \beta_{P_{1,11}})$.

2) *"Vertical" Markov model for transients*: we favor vertical structures for the transients. For a fixed frame index n, the sequence $\{\gamma_{2,(q,n)}\}_{q=1,..,l_{\text{frame2}}}$ is modeled by a two-state first-order Markov chain with transition probabilities $P_{2,00}$ and $P_{2,11}$, assumed equal for all frames. The transition probabilities are estimated and given Beta priors $\mathcal{B}(P_{2,00}|\alpha_{P_{2,00}}, \beta_{P_{2,00}})$ and $\mathcal{B}(P_{2,11}|\alpha_{P_{2,11}}, \beta_{P_{2,11}})$. The initial distribution $\pi_2 = P(\gamma_{2,(1,n)} = 1)$ is learned and given a Beta prior $\mathcal{B}(\pi_2|\alpha_{\pi_2}, \beta_{\pi_2})$. The transients model is illustrated in Fig. 2.

Remark 1: The stationarity of the horizontal Markov chain is important, as it implies that the distribution of the corresponding indicator variables is shift invariant. Therefore, the tonal layer possesses some built-in weak form of stationarity property, as follows. Denote by $\mathbb{E}_{\tilde{s}_1}$ the expectation taken with respect to the random coefficients (integration over $p(\tilde{s}_1|v_1, \gamma_1, \tilde{s}_2, \sigma^2, x)$), and by $\mathbb{E}_{\gamma_1}$ the expectation with respect to the indicator random variables of the tonal layer (integration over $p(\gamma_1|v_1, \sigma^2, \tilde{s}_2, \sigma^2, x)$). Denoting by $s_1 = \sum_{q,n} \tilde{s}_{1,(q,n)} \Phi_{1,(q,n)}$ the tonal signal, one readily shows that

$$\begin{aligned}\mathbb{E}_{\tilde{s}_1}\{s_1(t + ml_{\text{frame1}})s_1(t' + ml_{\text{frame1}})\} &= \sum_{q,n} \gamma_{1,(q,n)} v_{1,(q,n)} \Phi_{1,(q,n-m)}(t)\Phi_{1,(q,n-m)}(t') \\ &= \sum_{q,n} \gamma_{1,(q,n+m)} v_{1,(q,n+m)} \Phi_{1,(q,n)}(t)\Phi_{1,(q,n)}(t'). \quad (10)\end{aligned}$$

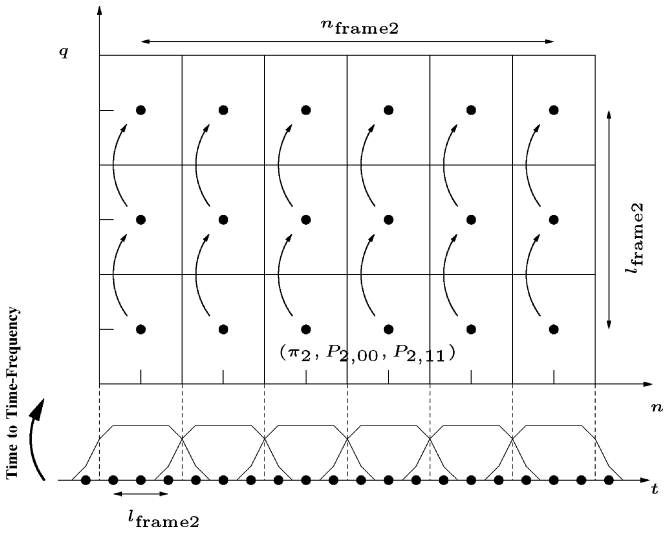

Fig. 2. This figure illustrates the transients model. A shorter time resolution than the one used for tonals is used in order to capture short sound components. The set of indicator variables γ_2 is modeled as "vertical and parallel" Markov chains of order 1, with common transition probabilities $P_{2,00}$ and $P_{2,11}$ and initial probability π_2.

If the horizontal Markov chain is in its stationary regime, $\mathbb{E}_{\gamma_1}\{\gamma_{1,(q,n)}\}$ is independent of n, and further assuming that the variances $v_{1,(q,n)}$ are also independent of n, one is led to

$$\begin{aligned}\mathbb{E}_{\gamma_1}\mathbb{E}_{\tilde{s}_1}\{s_1(t + ml_{\text{frame1}})s_1(t' + ml_{\text{frame1}})\} &= \mathbb{E}_{\gamma_1}\mathbb{E}_{\tilde{s}_1}\{s_1(t)s_1(t')\}. \quad (11)\end{aligned}$$

In other words, the (doubly averaged) second-order moment of the tonal layer is invariant under time shifts that are multiple of the horizontal time resolution. Notice that this calculation does not assume $\mathbb{E}_{\gamma_1}\{\gamma_{1,(q,n)}\}$ to be independent of the frequency index. This assumption will be made in the class of models considered here, but can be relaxed easily.

Remark 2: In contrast with modeling of the tonal part, we do not see any good reason for assuming (frequency) stationarity of the transient indicator variables, i.e., the vertical Markov chain needs not be at equilibrium (for example, the "vertical wavelet chains" considered in [10] do not admit an equilibrium distribution). Moreover, significant physical information regarding the nature of transients is likely to be contained in such a lack of frequency translation invariance: very "percussive" transients have a much more important high-frequency content than smoother ones. This may be described by the behavior of the indicator variables $\gamma_{2,k}$ as well as the frequency profiles $f_{2,k}$.

D. Residual Model

The variance σ^2 of the residual signal e, assumed independent and identically distributed (i.i.d) zero-mean Gaussian, is given an inverted-Gamma (conjugate) prior $p(\sigma^2|\alpha_\sigma, \beta_\sigma) = \mathcal{IG}(\sigma^2|\alpha_\sigma, \beta_\sigma)$.

E. Frequency Profile Parameters Priors

In the following, only the scale parameter λ_i will be estimated, while the filter cutoff η_i and order ν_i are fixed in advance.

The cutoff frequency η_i is set to $l_{\text{frame}i}/3$ while several values of ν_i are considered in the results section. λ_i is given for each basis a Gamma (conjugate) prior $p(\lambda_i|\alpha_{\lambda_i},\beta_{\lambda_i}) = \mathcal{G}(\lambda_i|\alpha_{\lambda_i},\beta_{\lambda_i})$. The value of the degrees of freedom α_i in (7) was found to have little influence over the results and we set it to 1 in practice.

III. MCMC Inference

It is proposed to sample from the posterior distribution of the parameters $\boldsymbol{\theta} = \{\tilde{s}_i, v_i, \alpha_i, \lambda_i, P_{i,00}, P_{i,00}\}_{i=1,2} \cup \{\sigma^2, \pi_2\}$, using a Gibbs sampler. The Gibbs sampler is a standard MCMC technique which simply requires to sample, iteratively with replacement, from the distribution of each parameter conditioned upon the others [19]. Point estimates or more generally complete posterior density estimates can then be computed from the samples obtained from the posterior distribution $p(\boldsymbol{\theta}|x)$. Since most of the parameters in our model are chosen to have conjugate priors, derivations for the Gibbs sampling steps are rather straightforward, and have thus been skipped. Derivations that required particular care can be found in Appendix II. Note that except in a few cases where Metropolis–Hastings (M–H) steps are needed, all conditional posterior distributions can be easily sampled.

A. Alternate Sampling of $(\gamma_1, \tilde{s}_1)$ and $(\gamma_2, \tilde{s}_2)$

One approach is to sample $\gamma = [\gamma_1^T \gamma_2^T]^T$ and $\tilde{s} = [\tilde{s}_1^T \tilde{s}_2^T]^T$ successively. Denoting $\Phi = [\Phi_1\ \Phi_2]$, this strategy is akin to standard Bayesian variable selection [20] and requires the storage and inversion of matrices of the form $(\Phi^T\Phi + \sigma^2 \text{diag}([v_1^T\ v_2^T]^T)^{-1})$ at each iteration of the sampler, which might not be feasible when N is large. The structure of our dictionary Φ allows for efficient alternate block sampling of $(\gamma_1, \tilde{s}_1)$ and $(\gamma_2, \tilde{s}_2)$, in the fashion of [8] and [9]. Indeed, because the Euclidean norm is invariant under rotation, the likelihood of the observation x can be written

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp -\frac{1}{2\sigma^2}\|x - \Phi_1\tilde{s}_1 - \Phi_2\tilde{s}_2\|_2^2 \\ &= (2\pi\sigma^2)^{-N/2} \exp -\frac{1}{2\sigma^2}\left\|\underbrace{\Phi_2^T(x - \Phi_1\tilde{s}_1)}_{\tilde{x}_{2|1}} - \tilde{s}_2\right\|_2^2 \\ &= (2\pi\sigma^2)^{-N/2} \exp -\frac{1}{2\sigma^2}\left\|\underbrace{\Phi_1^T(x - \Phi_2\tilde{s}_2)}_{\tilde{x}_{1|2}} - \tilde{s}_1\right\|_2^2. \end{aligned}$$

This means that conditionally upon $\tilde{s}_2$ (resp. $\tilde{s}_1$) and the other parameters, inferring $\tilde{s}_1$ (resp. $\tilde{s}_2$) is a simple filtering problem with data $\tilde{x}_{1|2}$ (resp. $\tilde{x}_{2|1}$), variable $\tilde{s}_1$ (resp. $\tilde{s}_2$) modeled as i.i.d conditionally upon γ_1 (resp. γ_2), and i.i.d noise, and thus does not require any matrix inversion. In the following, we will write, for $i = 1, 2$

$$\tilde{x}_{i|-i} = \tilde{s}_i + \tilde{e}_i \tag{12}$$

where $\tilde{x}_{i|-i}$ is either $\tilde{x}_{1|2}$ or $\tilde{x}_{2|1}$ and $\tilde{e}_i = \Phi_i^T e$. $\tilde{e}_i$ is Gaussian i.i.d with variance σ^2.

B. Update of $(\gamma_i, \tilde{s}_i)$

As pointed out in [20], an implementation of the Gibbs sampler consisting of sampling alternatively $\tilde{s}_{i,k}|\gamma_{i,k}$ and $\gamma_{i,k}|\tilde{s}_{i,k}$ cannot be used as it leads to a nonconvergent Markov chain (the Gibbs sampler gets stuck when it generates a value $\tilde{s}_{i,k} = 0$). Instead, we need to sample $(\tilde{s}_{i,k}, \gamma_{i,k})$ jointly, by:

1) sampling $\gamma_{i,k}^{(l)}$ from $p(\gamma_{i,k}|\gamma_{i,-k}, \theta_{\gamma_i}, v_i, \sigma^2, \tilde{x}_{i|-i,k})$;
2) sampling $\tilde{s}_{i,k}^{(l)}$ from $p(\tilde{s}_{i,k}|\gamma_{i,k}^{(l)}, v_i, \sigma^2, \tilde{x}_{i|-i,k})$

where $\gamma_{i,-k}$ denotes the set $\{\gamma_{i,1}, \ldots, \gamma_{i,k-1}, \gamma_{i,k+1}, \ldots, \gamma_{i,N}\}$ and where θ_{γ_i} is the set of probabilities in the Markov model for γ_i.

The computation of the first posterior distribution is akin to solving a hypothesis testing problem [21], with

$$(H_1) \iff \gamma_{i,k} = 1 \iff \tilde{x}_{i|-i,k} = \tilde{s}_{i,k} + \tilde{e}_{i|-i,k} \tag{13}$$
$$(H_0) \iff \gamma_{i,k} = 0 \iff \tilde{x}_{i|-i,k} = \tilde{e}_{i|-i,k}. \tag{14}$$

The ratio $\tau_{i,k} = p(\gamma_{i,k} = 1|\gamma_{i,-k}, \theta_{\gamma_i}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) / p(\gamma_{i,k} = 0|\gamma_{i,-k}, \theta_{\gamma_i}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k})$ is thus simply expressed as

$$\begin{aligned} \tau_{i,k} &= \frac{\mathcal{N}(\tilde{x}_{i|-i,k}|0, v_{i,k} + \sigma^2)}{\mathcal{N}(\tilde{x}_{i|-i,k}|0, \sigma^2)} \frac{p(\gamma_{i,k} = 1|\gamma_{i,-k}, \theta_{\gamma_i})}{p(\gamma_{i,k} = 0|\gamma_{i,-k}, \theta_{\gamma_i})} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2 + v_{i,k}}} \exp\left(\frac{\tilde{x}_{i|-i,k}^2 v_{i,k}}{2\sigma^2(\sigma^2 + v_{i,k})}\right) \\ &\quad \times \frac{p(\gamma_{i,k} = 1|\gamma_{i,-k}, \theta_{\gamma_i})}{p(\gamma_{i,k} = 0|\gamma_{i,-k}, \theta_{\gamma_i})}. \end{aligned} \tag{15}$$

Values of the ratio $\tau_{i,k}^{\text{prior}} = p(\gamma_{i,k} = 1|\gamma_{i,-k}, \theta_{\gamma_i}) / p(\gamma_{i,k} = 0|\gamma_{i,-k}, \theta_{\gamma_i})$ are given in Appendix II-A. $\gamma_{i,k}$ is thus drawn from the two states discrete distribution with probability masses

$$p(\gamma_{i,k} = 0|\gamma_{i,-k}, \theta_{\gamma_i}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) = 1/(1+\tau_{i,k}) \tag{16}$$
$$p(\gamma_{i,k} = 1|\gamma_{i,-k}, \theta_{\gamma_i}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) = \tau_{i,k}/(1+\tau_{i,k}). \tag{17}$$

When a value $\gamma_{i,k} = 0$ is drawn, $\tilde{s}_{i,k}$ is set to zero. Otherwise, when $\gamma_{i,k} = 1$, inferring $\tilde{s}_{i,k}$ conditionally upon $v_{i,k}$ simply amounts to inferring a Gaussian parameter embedded in Gaussian noise, i.e, Wiener filtering. The posterior distribution of $\tilde{s}_{i,k}$ is thus written as

$$\begin{aligned} p(\tilde{s}_{i,k}|\gamma_{i,k}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) &= (1 - \gamma_{i,k})\delta_0(\tilde{s}_{i,k}) \\ &\quad + \gamma_{i,k}\mathcal{N}\left(\tilde{s}_{i,k}|\mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}}\right) \end{aligned} \tag{18}$$

with $\sigma^2_{\tilde{s}_{i,k}} = (1/\sigma^2 + 1/v_{i,k})^{-1}$ and $\mu_{\tilde{s}_{i,k}} = (\sigma^2_{\tilde{s}_{i,k}}/\sigma^2)\tilde{x}_{i|-i,k}$.

C. Update of v_i

The conditional posterior distribution of $v_{i,k}$ is simply

$$\begin{aligned} p(v_{i,k}|\gamma_{i,k}, \tilde{s}_{i,k}, f_{i,k}) &= (1 - \gamma_{i,k})\mathcal{IG}(v_{i,k}|\alpha_i, f_{i,k}) \\ &\quad + \gamma_{i,k}\mathcal{IG}\left(v_{i,k}\left|\frac{1}{2} + \alpha_i, \frac{\tilde{s}_{i,k}^2}{2} + f_{i,k}\right.\right). \end{aligned} \tag{19}$$

When a value $\gamma_{i,k} = 0$ is generated, $\tilde{v}_{i,k}$ is simply sampled from its prior (no posterior information is available); otherwise, it is inferred from the available value of $\tilde{s}_{i,k}$. In the latter case, the

posterior distribution is easily calculated because of the use of a conjugate prior for $v_{i,k}$.

D. Update of σ^2

The conditional posterior distribution of σ^2 is given by

$$p(\sigma^2|\tilde{s}_1,\tilde{s}_2,x)=\mathcal{IG}\left(\sigma^2\left|\frac{N}{2}+\alpha_\sigma,\frac{\|x-\Phi_1\tilde{s}_1-\Phi_2\tilde{s}_2\|_2^2}{2}+\beta_\sigma\right.\right). \quad (20)$$

E. Update of the Scale Parameters

The full posterior distribution of the scale parameters is

$$p(\lambda_i|v_i)=\mathcal{G}\left(\lambda_i\left|N\alpha_i+\alpha_{\lambda_i},\sum_k\frac{1}{1+\left(\frac{q-1}{\eta_i}\right)^{\nu_i}v_{i,k}}+\beta_{\lambda_i}\right.\right). \quad (21)$$

As noted in [9] and [12], because we are expecting sparse representations, most of the indicator variables $\gamma_{i,k}$ take the value 0, and thus most of the variances $v_{i,k}$ are sampled from their prior [see (19)]. Thus, the influence of the data in the full posterior distribution of λ_i becomes small, and the convergence of these parameters can be very slow. A faster scheme consists of making one draw from $p(\{v_{i,k}:\gamma_{i,k}=1\}|\{\tilde{s}_{i,k}:\gamma_{i,k}=1\},\lambda_i)$, then one draw from $p(\lambda_i|\{v_{i,k}:\gamma_{i,k}=1\})$ and finally one draw from $p(\{v_{i,k}:\gamma_{i,k}=0\}|\lambda_i)$.

Let us mention that the conditional posterior density of η_i and ν_i can also be written, yielding

$$p(\nu_i,\eta_i|v_i,\lambda_i)\propto\prod_{k=1}^{N}\left(1+\left(\frac{q-1}{\eta_i}\right)^{\nu_i}\right)^{-\alpha_i}\times\exp-\frac{\lambda_i}{\left(1+\left(\frac{q-1}{\eta_i}\right)^{\nu_i}\right)v_{i,k}}\quad p(\nu_i,\eta_i). \quad (22)$$

This posterior distribution is not easily sampled and, in an effort to estimate ν_i and possibly η_i as well, we resorted to Metropolis random walk strategies to address this task. We observed very slow converging chains for (λ_i,η_i,ν_i) and convergence did not seem to be obtained before several thousands of iterations. To complete this task more efficiently, a better sampling scheme is yet to be found. However, the results section will show that the exact value of ν_i (with η_i fixed to the reasonable value $l_{\text{frame}i}/3$) is not of the highest importance.

F. Update of the Markov Chains Transition and Initial Probabilities

The posterior distribution of probabilities $P_{2,00}$ and $P_{2,11}$ mostly involve counting the number of changes from $\gamma_{2,(q-1,n)}=i$ to $\gamma_{2,(q,n)}=j$, where $i,j\in\{0,1\}$ and the posterior distribution of π_2 involves counting the number of values of $\gamma_{2,(q,1)}$ equal to 1. These variables have Beta posterior distributions whose expressions are given in Appendix II-B2. Because we have assumed the initial probability of the chain to be equal to its equilibrium probability, the posterior distributions of $P_{1,00}$ and $P_{1,11}$ do not belong to a family of distributions easy to sample. Their expressions are given in Appendix II-B1 where we describe an exact M–H scheme as well as a deterministic scheme to update these variables.

IV. Results

A. Denoising of a Short Glockenspiel Excerpt

1) Experimental Setup: We present denoising results of a glockenspiel excerpt, sampled at 44.1 kHz with length $N=131072(\approx 3$ s). White Gaussian noise was added to the clean signal with input signal-to-noise ratios (SNRs) {0, 10, 20} (dB). We applied the following strategies to the noisy excerpt, with in every case $l_{\text{frame1}}=1024(\approx 23$ ms), $l_{\text{frame1}}=128(\approx 3$ ms):

1) the proposed dual-resolution approach, with $\nu_1=2$ and $\nu_2=1$;
2) the proposed dual-resolution approach, with $\nu_1=8$, $\nu_2=4$;
3) a single-resolution approach, in which no transient model is used. The signal x is solely decomposed as $\tilde{x}=\Phi_1\tilde{s}_1+e$, with horizontal structured priors used for γ_1. The MCMC inference strategy described in Section III still holds, with $i=1$ and $x_{i|-i}=\Phi_1^T x$;[5]
4) the dual-resolution approach of [9], in which independent Bernoulli (unstructured) priors are considered for γ_1 and γ_2 and flat frequency profiles are used.

Remark 3: In cases 1) and 2), ν_1 was chosen greater than ν_2 to model our belief that transients should have a slower decreasing frequency profile than tonals. The choice of frame lengths was motivated by our tests on real audio signals. Even though a short frame of approximately 3 ms does not make much sense from the point of view of acoustics (this is shorter than the duration of short attacks), this choice turned out to be better in practice, because the two frame lengths need to be sufficiently different to discriminate tonals and transients. For example, taking $l_{\text{frame2}}=256(\approx 6$ ms) generally results in worse separations, in the sense that transients start to sound significantly "tonal." This is why we have made such a choice, at the price of sometimes needing several consecutive "vertical lines" for describing a transient.

The Gibbs samplers of methods (1–4) were run for 1000 iterations, which, on Mac G4 clocked at 1.25 GHz with RAM 512 MB, takes 68 min for (1,2), 12 min for 3) and 63 min for 4). The hyperparameters of the priors for λ_1, λ_2, and σ^2 were chosen as to yield Jeffreys noninformative distributions. The hyperparameters $\alpha_{P_{i,00}}$ and $\beta_{P_{i,11}}$ were respectively fixed to 50 and 1, thus giving more weight to values ranging from 0.8 to 1. Finally, we set $\alpha_{\pi_2}=1$ and $\beta_{\pi_2}=5000$, yielding a prior density for π_2 favoring very low values of this parameter. If rather noninformative priors could be chosen for λ_i, σ^2, $\alpha_{P_{i,00}}$, and $\beta_{P_{i,11}}$ as enough data is available to estimate them, we found out in practice that, on the contrary, the prior parameters for π_2 had to be set to realistic values. Indeed, choosing a noninformative prior for this parameter could lead to unsatisfying results on some signals. The algorithm would find many spurious transients, yielding significance maps (see below) full of short vertical lines in the very low part of the frequency range. The values $\alpha_{\pi_2}=1$ and $\beta_{\pi_2}=5000$ yielded satisfactory results over a wide range of signals.

We computed MMSE estimates of the parameters by averaging the last 300 sampled values. A source estimate was recon-

[5]This "tonal-only" model is very close to one of the models considered in [12], where a Wilson basis is employed instead of the MDCT.

TABLE I
OUTPUT SNRS (DB) OBTAINED WITH EACH OF THE METHODS FOR THREE DIFFERENT VALUES OF INPUT SNRS

Input SNR (dB)	0	10	20
1)	15.8	22.5	29.3
2)	16.0	22.5	29.2
3)	15.3	20.7	27.6
4)	13.7	20.9	28.1

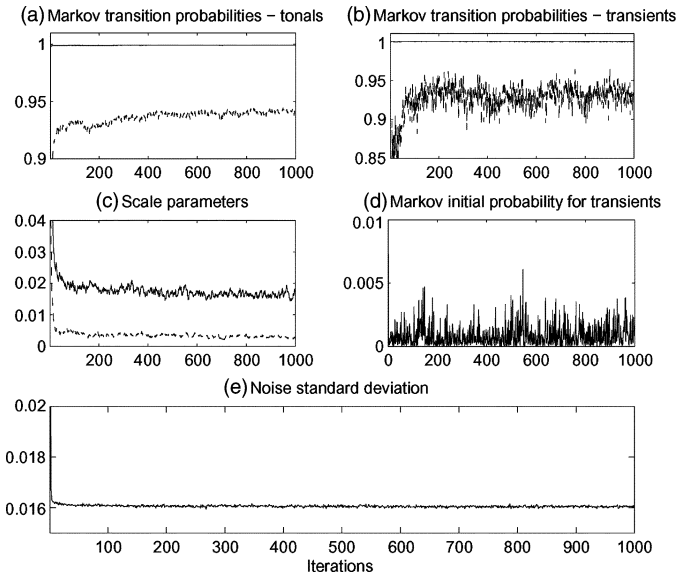

Fig. 3. Sampled values of (a): $P_{1,00}(-)$, $P_{1,11}(--)$, (b): $P_{2,00}(-)$, $P_{2,11}(--)$ (c): $\lambda_1(-)$, $\lambda_2(--)$, (d): π_2, (e): σ, in the 10-dB input SNR case and with approach 1). The original value of σ used in the simulation is 0.0158, its MMSE estimate is 0.0166 ± 0.0027.

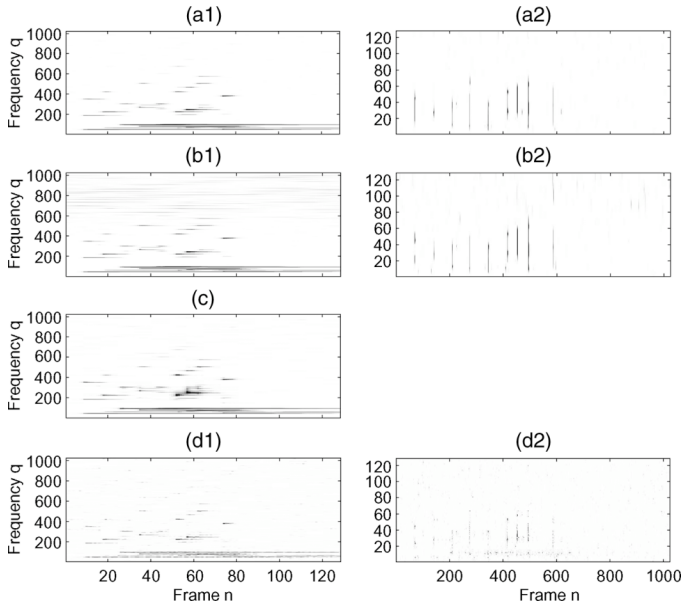

Fig. 4. Significance maps of the selected atoms in Φ_1 and Φ_2 for each method, in the 10-dB input SNR case. The maps are computed as the MMSE estimates of the indicator variables γ_1 and γ_2, so that values range continuously from 0 (white) to 1 (black). Significance maps from approach 1) to approach 4) are shown top to bottom.

structed by $s^{\mathrm{MMSE}} = \Phi_1 \tilde{s}_1^{\mathrm{MMSE}} + \Phi_2 \tilde{s}_2^{\mathrm{MMSE}}$. Table I shows the overall output SNR $20 \log_{10} \|s^{\mathrm{MMSE}} - s\| / \|s\|$ obtained with each method. Audio files can be found in [22]. Fig. 3 shows the values of $\{\lambda_1, \lambda_2, P_{1,00}, P_{1,11}, P_{2,00}, P_{2,11}, \sigma\}$ generated by the Gibbs sampler of approach 1), in the 10-dB input SNR case. Fig. 4 shows significance maps of the selected atoms in each basis with all methods, computed here as the MMSE estimates of γ_1 and γ_2 [only γ_1 in case 3)], in the 10-dB input SNR case.

2) Discussion: On the quality of denoising, we can draw two major conclusions from the latter results. One of them is that there is a gain at using structured priors. This is revealed on the one hand by the higher output SNRs obtained by approaches 1) and 2) as compared to approach 4), see Table I, and more convincingly on the other hand by the sound samples, which contains less artifacts in the first two cases. These artifacts originate from isolated atoms in the time–frequency plane, as illustrated on plots (d1) and (d2) of Fig. 4. Because of the structured priors employed in approaches 1) and 2), much of these isolated atoms have been removed, and the significance maps have been *regularized*, as can be seen on plots (a1), (a2), (b1), (b2) of Fig. 4.

Another conclusion is that there is a gain at modeling the transients as well as the tonals. This is revealed by the higher output SNRs obtained by approaches 1) and 2) as compared to approach 3) and also by the sound samples, which in the first two cases sound "crisper" than with the tonal-only model used in approach 3). The lack of transients model in 3) also creates some pre-echo at the beginning of the notes.

The results tend to show that the values of ν_1 and ν_2 do not have a strong impact on the results, especially in terms of output SNRs. More atoms are indeed selected in the high-frequency range with approach 2) as compared with approach 1), as can be seen on plots (b1) and (b2) of Fig. 4, but listening to the audio samples does not reveal a large perceptual difference. One might find that the source estimate obtained with approach 2) sounds slightly "brighter" than the other. However, we noticed that in low-input SNRs conditions, setting a high value of ν_2 could help detecting some transients that would have been undetected with a low value. For example, in the 0-dB input SNR case, the audio samples reveal that the attack of the second note is not captured by 1), while it is detected by 2). Note also that both approaches miss the attack of the third note.

On the computational side, modeling the transients does lead to an important increase of the computational burden, which is multiplied by 4 between approaches (1,2,4) and (3). This is because of the MDCT and inverse MDCT (IMDCT) operations required at each step of the Gibbs sampler in approaches (1,2,4): the computations of $\tilde{x}_{1|2}$ and $\tilde{x}_{2|1}$ each require one MDCT operation and one IMDCT operation. On the opposite, approach 3) only requires one MDCT operation at the beginning to obtain the input data to the Gibbs sampler and one IMDCT operation at the end to reconstruct a source estimate. However, using structured priors in approaches 1) and 2) instead of unstructured priors as in approach 4) has little cost, only 68 min of CPU time for 1000 iterations instead of 65 min (4% increase).

The Gibbs sampling strategies used for (1–4) are of course computationally more demanding than EM approaches such as the one used in [8]. However, they do not suffer from problems of convergence to local minima, problems that we did encounter

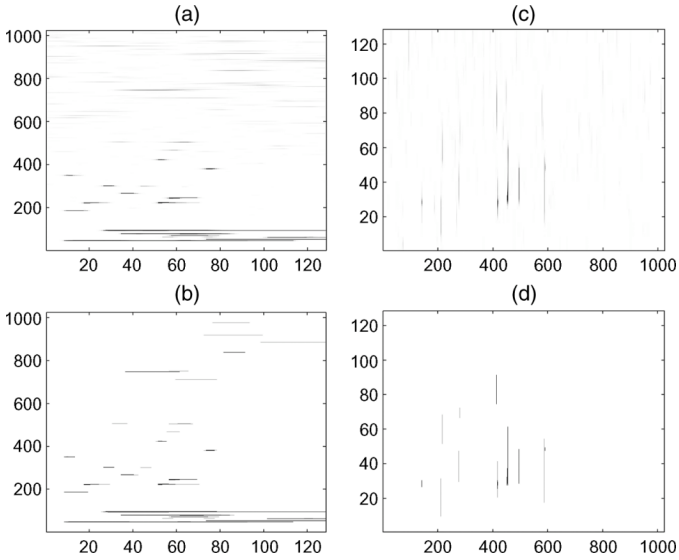

Fig. 5. MMSE and MAP estimates of γ_1 and γ_2 obtained with approach 2) for the 0-dB input SNR case. (a) MMSE estimate of γ_1. (b) MAP estimate of γ_1. (c) MMSE estimate of γ_2. (d) MAP estimate of γ_2.

in our earlier trials of using EM with the source model described in (6) and (7). Because MCMCs strategies yield a full description of the posterior distribution and not only one point estimate [typically a maximum *a posteriori* (MAP) estimate], they can be used to compute a wide range of point estimates, including uncommon ones. As such, in order to further eliminate the residual artifacts in the MMSE estimates, we computed the following source estimates:

$$s^{\mathrm{MIX}} = \Phi_1\left(\tilde{s}_1^{\mathrm{MMSE}} \otimes \gamma_1^{\mathrm{MAP}}\right) + \Phi_2\left(\tilde{s}_2^{\mathrm{MMSE}} \otimes \gamma_2^{\mathrm{MAP}}\right) \quad (23)$$

where $\otimes$ denotes vector element-wise multiplication, and where γ_1^{MAP} and γ_2^{MAP} are MAP estimates of γ_1 and γ_2.[6] This leads to slightly lower output SNRs and slightly "MIDI-like" sound quality, but removes all of the artifacts. The output SNRs values and corresponding sound samples are available in [22]. MMSE and MAP estimates of γ_1 and γ_2 obtained with approach 2) can be compared in Fig. 5, for the 0-dB input SNR case. Note that for this latter case, the output SNRs for s^{MMSE} and s^{MIX} are, respectively, 16.0 and 15.8, but that the total number of selected atoms is, respectively, 12467 (4.8%) and 1407 (0.5%), so that our proposed model and inference technique could be relevant to simultaneous denoising and very-low bit-rate coding of noisy musical signals.

B. Denoising of a Long Polyphonic Audio Excerpt

1) Experimental Setup: We now present denoising results of a long polyphonic excerpt. The data is a 24-s-long excerpt of the song *Mama Vatu* from Susheela Raman, sampled at 44.1 kHz with $T = 2^{20} = 1048576$. The excerpt starts with drums only, then enters an acoustic guitar and then the voice. White Gaussian noise was added to the excerpt with $\sigma = 0.03$. The data was segmented in n_S "superframes" of length l_S and each superframe was processed separately with approach 2). Three values of l_S were considered, as shown in Table II. In every case, the superframes are overlapping over 1024 samples, where a sinebell window was used for analysis and overlap-add reconstruction of the full denoised signals.[7] The sampler was now run on a more recent computer, a Mac Pro clocked at 3 GHz with 4-GB RAM, and computation time is divided by 4, supporting the possibility that MCMC approaches get more and more popular as computational power increases.

TABLE II
STATISTICS RELATED TO THE DENOISING OF THE 24-S-LONG POLYPHONIC MUSICAL EXCERPT

l_S (samples)	32768	65536	131072
l_S (seconds)	0.75	1.5	3.0
n_S	33	16	8
Computation time (min)	105	101	107
Percentage of selected atoms (MMSE)	40.7	36.0	30.4
Percentage of selected atoms (MIX)	5.2	5.4	5.0
Overall output SNR (MMSE) (dB)	20.5	20.2	20.3
Overall output SNR (MIX) (dB)	20.1	19.9	20.0

The input and output SNRs in each superframe and for each value of l_S are represented on Fig. 6. For comparison, we also applied to the whole signal the standard MMSE short-time spectral amplitude (STSA) estimator under uncertainty of signal presence of Ephraim and Malah [21]. The short-time Fourier transform of the signal was computed with same time tiling as the first MDCT basis: sinebell window of length 2048 $(2l_{\mathrm{frame1}})$, 50% overlap. The noise variance was fixed to its true value, the signal variance at each time–frequency point was estimated through moving average over the three precedent frames. The signal presence probability was arbitrarily fixed to 0.1 (which seemed to give a good tradeoff between perceptual quality and overall output SNR). Running the Ephraim and Malah STSA estimator only takes seconds and yields 19.3-dB overall output SNR. Sound samples can be found in [22].

2) Discussion: As can be seen in Fig. 6, the input SNR ranges roughly around 10 dB before the voice enters and then around 15 dB. The output SNRs range around 20 dB throughout, with a low variance in cases $l_S = 65536$ and $l_S = 131072$. The global estimates are of acceptable audio quality with best results obtained to our opinion in the $l_S = 65536$ case. The denoising of the first part of the signal, containing music only, is especially good. The denoising of the last part, containing voice and music, is less satisfying, probably because the signal is "richer" (and thus less sparse) but also because our model does not take into account the specificities of voice: vibrato/glissando, unvoiced phonemes, etc.

If the local estimates in each superframe in particular have a good quality, the reconstructed global estimates however suffer from changes of "regime" from one superframe to another, which result in slight changes of loudness and timbre. Again, even though it has a lower output SNR, s^{MIX} contains less artifacts than s^{MMSE} and is more pleasant to listen too. Note also that, as shown in Table II, s^{MIX} employs in every case only $\approx$5% of the total numbers of atoms $K = 2T = 2^{21}$ while

[6]The MAP estimate of γ_i is simply computed by thresholding to 0 all the values of γ_i^{MMSE} lower that 0.5 and thresholding to 1 all the values greater than 0.5. Note that other threshold values could also be considered.

[7]In every case the last superframe was dropped because it contained mainly zeros originating from the prior zero-padding of the signal. The denoised signal is thus slightly shorter than the original noisy signal, the missing bit is replaced by light noise in the audio results .

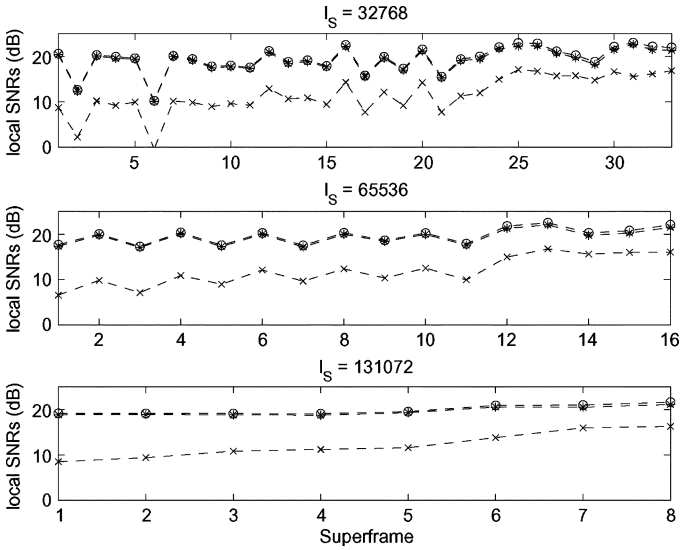

Fig. 6. Evolution of the local input SNR (−x−), MMSE output SNR (−o−) and MIX output SNR (− ∗ −) for each value of l_S. The x-axis represents the superframe index (ranging from 1 to n_S).

s^{MMSE} employs around six times more of them. In comparison, the STSA estimate sounds the same throughout, contains less musical noise but still contains quite a lot of noise. It also sounds "flatter," mainly because the transients are attenuated.

V. Conclusion

In this paper, a new approach for audio signal denoising has been presented and demonstrated, based upon prior probabilistic modeling of the signal. The main two aspects of the model are:

- the overcompleteness of the waveform dictionary that is used for expanding the signal;
- the introduction of (various sorts of) dependencies in the *transform domain*, i.e., between the coefficients of the expansion.

These two aspects may be seen as attempts to move towards models that are more realistic than the usual waveform models. Different types of waveforms (here, broad and narrow MDCT atoms) are used to capture different components in the signal (here, tonals and transients). Since these waveforms are still not sufficient to model directly such components, the introduction of dependencies between coefficients (structures) provides a way to improve the modeling. Components may be viewed as *chains* (modeled here as Markov chains) of dependent time–frequency atoms, that could be called *time–frequency molecules*.

The modeling involving two layers proposed in this paper seems rather accurate, as advocated in [10] and [14]. In addition, the proposed algorithm avoids most problematic parameter tunings that were present in [14] and has the advantage of simultaneous estimation of the two layers, unlike the algorithm of [10]. The model is versatile enough to allow other ingredients, such as for example the frequency profiles in (8).

The numerical results presented here demonstrate the efficiency of our approach in the framework of a denoising problem. Since it provides a fairly simple description of signals in terms of limited numbers of coefficients, one may also think of using this approach as a preprocessing step for further applications, such as tempo identification, segmentation, or source separation. However, let us stress that in the current version of the algorithm, the signal model of (5) is in fact driven by the noise term e, which implies that the approach is bound to *denoising* problems, and not directly transposable to other tasks such as coding. Indeed, running the algorithm on "clean" signal results in poor signal decompositions. This is due to the fact that unlike noisy signals, "clean signals" have a sparse expansion in the dictionary, and the number of degrees of freedom to be taken into account is unknown *a priori*. As a result, the algorithm may produce very sparse representations of signals when a small noise is added, but the quality of the reconstruction may be problematic if high precision is needed.

Let us stress that stationary colored noise can be considered as well, but destroys the conditional independence structure of the coefficients and thus impairs computational efficiency. In the general case where $\mathbb{E}\{ee^T\} = \Sigma_e$, $\tilde{e}_i$ [defined in (12)] is still a Gaussian process but with covariance $\Phi_i^T \Sigma_e \Phi_i$. Thus, data $\tilde{x}_{i|-i}$ is not i.i.d anymore, and inferring $\tilde{s}_i$ now requires inverting $(\mathrm{diag}(v_i)^{-1} + (\Phi_i^T \Sigma_e \Phi_i)^{-1})$. In fact, as in the general Bayesian variable selection setting, full block update of γ_i now requires computing 2^N posterior probabilities corresponding to every possible value of γ_i, the computation of each requiring itself to invert a matrix of the latter form. However, it is still fairly cheap to implement a component-by-component Gibbs sampler where one expansion coefficient is updated conditionally upon the others and the indicator variables, like in [23] (where a parametric AR model of the noise is used), or to some extent like in [12] where the noise is white but where the coefficients are updated pairwise. Another possibility to keep the conditional independence structure of the coefficients, is to approximate $\Phi_1^T \Sigma_e \Phi_1$ and $\Phi_2^T \Sigma_e \Phi_2$ by diagonal matrices.

As to the applicability of our approach to the denoising of long musical excerpts, if the strategy that we propose in Section IV-B yields encouraging results, it is yet not optimal. A much better strategy would consist of taking an online approach of the problem, in which frames (of size l_{frame1} or a multiple) of the noisy signal would be processed sequentially, using dynamic models of the parameters λ_i, $P_{i,00}$, and $P_{i,11}$, and possibly σ^2. The classical approach to such updating problems is the Kalman filter. Here, however, we have intractable updates that will require numerical computations. Particle filters are a state-of-the-art method that might be used to deal with a complex model such as this (see [24]–[26] for introductory material and some audio noise-reduction applications). Such an approach should prevent the audible "changes of regime" encountered in our results here.

Further work can extend the models in several useful ways. First, it will be natural to extend the framework to use multiple bases of different resolutions, rather than the two proposed here. Then, one can envisage models for long, slowly varying tonals as well as shorter, more rapidly varying, tonals. One can also readily extend the framework to include other types of bases, especially wavelet bases (and corresponding wavelet tree prior models [27]). Open questions remain about how best to construct the structured priors in such settings: for example, one might expect dependencies of indicators both *within* a single basis and *between* different bases. As such, one might want to lift the independence assumption between transients and tonals

of our model, and model the fact that in real signals, tonals are most often preceded by a transient (the attack). More general spatial Markov random field structures can also be envisaged for modeling of these dependencies in a tractable and physically meaningful way. Finally, one may consider the fixed time–frequency grids imposed in this framework as too constraining altogether, and more general multiresolution frameworks that allow arbitrary time–frequency locations and resolutions for the atoms, accompanied by appropriate structured priors, may be considered (see, e.g., the work of [28] for some advances in this direction). However, while the latter models are not difficult to think of, the corresponding estimation and denoising algorithms will no doubt require greater computational sophistication.

APPENDIX I
STANDARD DISTRIBUTIONS

Gaussian:
$\mathcal{N}(u|\mu,\sigma^2) = (2\pi\sigma^2)^{-(1/2)} \exp -((u-\mu)^2/2\sigma^2)$.
Student t:
$t(u|\alpha,\lambda) = (\Gamma((\alpha+1)/2)/\lambda\sqrt{\alpha\pi}\Gamma(\alpha/2))\ (1+(1/\alpha)\ (u/\lambda)^2)^{-((\alpha+1)/2)}$.
Beta:
$\mathcal{B}(u|\alpha,\beta) = (\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta))u^{\alpha-1}(1-u)^{\beta-1}, u\in[0,1]$.
Gamma:
$\mathcal{G}(u|\alpha,\beta) = (\beta^\alpha/\Gamma(\alpha))u^{\alpha-1}\exp(-\beta u), u\geq 0$.
inv-Gamma:
$\mathcal{IG}(u|\alpha,\beta) = (\beta^\alpha/\Gamma(\alpha))u^{-(\alpha+1)}\exp(-(\beta/u)), u\geq 0$.

The inverted-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.

APPENDIX II
CONDITIONAL POSTERIOR DENSITIES

A. Prior Weight

This section gives the expression of the prior weight $\tau_{i,k}^{\text{prior}} = p(\gamma_{i,k}=1|\gamma_{i,-k})/p(\gamma_{i,k}=0|\gamma_{i,-k})$ required in (15).

1) Horizontal Markov Chains: $\forall q = 1,\ldots,l_{\text{frame1}}$*:*

- $n=1$

$$\tau_{1,k}^{\text{prior}} = \frac{p\left(\gamma_{1,(q,2)}|\gamma_{1,(q,1)}=1\right)p\left(\gamma_{1,(q,1)}=1\right)}{p\left(\gamma_{1,(q,2)}|\gamma_{1,(q,1)}=0\right)p\left(\gamma_{1,(q,1)}=0\right)} = \begin{cases} \frac{(1-P_{1,11})\pi_1}{P_{1,00}(1-\pi_1)} = \frac{(1-P_{1,00})}{P_{1,00}}, & \text{if } \gamma_{1,(q,2)}=0 \\ \frac{P_{1,11}\pi_1}{(1-P_{1,00})(1-\pi_1)} = \frac{P_{1,11}}{(1-P_{1,11})}, & \text{if } \gamma_{1,(q,2)}=1 \end{cases}$$

- $n=2,\ldots,n_{\text{frame1}}-1$

$$\tau_{1,k}^{\text{prior}} = \frac{p\left(\gamma_{1,(q,n+1)}|\gamma_{1,(q,n)}=1\right)p\left(\gamma_{1,(q,n)}=1|\gamma_{1,(q,n-1)}\right)}{p\left(\gamma_{1,(q,n+1)}|\gamma_{1,(q,n)}=0\right)p\left(\gamma_{1,(q,n)}=0|\gamma_{1,(q,n-1)}\right)} = \begin{cases} \frac{(1-P_{1,00})(1-P_{1,11})}{P_{1,00}^2} & \text{if } \gamma_{1,(q,n-1)}=0 \text{ and } \gamma_{1,(q,n+1)}=0 \\ \frac{P_{1,11}}{P_{1,00}} & \text{if } \gamma_{1,(q,n-1)}=0 \text{ and } \gamma_{1,(q,n+1)}=1 \\ \frac{P_{1,11}}{P_{1,00}} & \text{if } \gamma_{1,(q,n-1)}=1 \text{ and } \gamma_{1,(q,n+1)}=0 \\ \frac{P_{1,11}^2}{(1-P_{1,11})(1-P_{1,00})} & \text{if } \gamma_{1,(q,n-1)}=1 \text{ and } \gamma_{1,(q,n+1)}=1 \end{cases}$$

- $n=n_{\text{frame1}}$

$$\tau_{1,k}^{\text{prior}} = \frac{p\left(\gamma_{1,(q,n_{\text{frame1}})}=1|\gamma_{1,(q,n_{\text{frame1}}-1)}\right)}{p\left(\gamma_{1,(q,n_{\text{frame1}})}=0|\gamma_{1,(q,n_{\text{frame1}}-1)}\right)} = \begin{cases} \frac{(1-P_{1,00})}{P_{1,00}} & \text{if } \gamma_{1,(q,n_{\text{frame1}}-1)}=0 \\ \frac{P_{1,11}}{(1-P_{1,11})} & \text{if } \gamma_{1,(q,n_{\text{frame1}}-1)}=1 \end{cases}$$

2) Vertical Markov Chains: $\forall n = 1,\ldots,n_{\text{frame2}}$*:*

- $q=1$

$$\tau_{2,k}^{\text{prior}} = \frac{p\left(\gamma_{2,(2,n)}|\gamma_{2,(1,n)}=1\right)p\left(\gamma_{2,(1,n)}=1\right)}{p\left(\gamma_{2,(2,n)}|\gamma_{2,(1,n)}=0\right)p\left(\gamma_{2,(1,n)}=0\right)} = \begin{cases} \frac{(1-P_{2,11})\pi_2}{P_{2,00}(1-\pi_2)} & \text{if } \gamma_{2,(2,n)}=0 \\ \frac{P_{2,11}\pi_2}{(1-P_{2,00})(1-\pi_2)} & \text{if } \gamma_{2,(2,n)}=1 \end{cases}$$

- $q=2,\ldots,l_{\text{frame2}}-1$

$$\tau_{2,k}^{\text{prior}} = \frac{p\left(\gamma_{2,(q+1,n)}|\gamma_{2,(q,n)}=1\right)p\left(\gamma_{2,(q,n)}=1|\gamma_{2,(q-1,n)}\right)}{p\left(\gamma_{2,(q+1,n)}|\gamma_{2,(q,n)}=0\right)p\left(\gamma_{2,(q,n)}=0|\gamma_{2,(q-1,n)}\right)} = \begin{cases} \frac{(1-P_{2,00})(1-P_{2,11})}{P_{2,00}^2} & \text{if } \gamma_{2,(q-1,n)}=0 \text{ and } \gamma_{2,(q+1,n)}=0 \\ \frac{P_{2,11}}{P_{2,00}} & \text{if } \gamma_{2,(q-1,n)}=0 \text{ and } \gamma_{2,(q+1,n)}=1 \\ \frac{P_{2,11}}{P_{2,00}} & \text{if } \gamma_{2,(q-1,n)}=1 \text{ and } \gamma_{2,(q+1,n)}=0 \\ \frac{P_{2,11}^2}{(1-P_{2,11})(1-P_{2,00})} & \text{if } \gamma_{2,(q-1,n)}=1 \text{ and } \gamma_{2,(q+1,n)}=1 \end{cases}$$

- $q=l_{\text{frame2}}$

$$\tau_{2,k}^{\text{prior}} = \frac{p\left(\gamma_{2,(l_{\text{frame2}},n)}=1|\gamma_{1,(l_{\text{frame2}}-1,n)}\right)}{p\left(\gamma_{2,(l_{\text{frame2}},n)}=0|\gamma_{1,(l_{\text{frame2}}-1,n)}\right)} = \begin{cases} \frac{(1-P_{2,00})}{P_{2,00}} & \text{if } \gamma_{2,(l_{\text{frame2}}-1,n)}=0 \\ \frac{P_{2,11}}{(1-P_{2,11})} & \text{if } \gamma_{2,(l_{\text{frame2}}-1,n)}=1 \end{cases}$$

B. Markov Transition and Initial Probabilities

1) Horizontal Markov Chain: We have

$$\begin{aligned} &p(\gamma_1|P_{1,00},P_{1,11},\pi_1) \\ &= \prod_{q=1}^{l_{\text{frame1}}}\prod_{n=2}^{n_{\text{frame1}}} p\left(\gamma_{1,(q,n)}|\gamma_{1,(q,n-1)},P_{1,00},P_{1,11}\right) \\ &\quad \times p\left(\gamma_{1,(q,1)}|\pi_1\right) \\ &= P_{1,00}^{\#\gamma_1(00)}(1-P_{1,00})^{\#\gamma_1(01)}P_{1,11}^{\#\gamma_1(11)}(1-P_{1,11})^{\#\gamma_1(10)} \\ &\quad \times \left(\frac{1-P_{1,00}}{2-P_{1,00}-P_{1,11}}\right)^{\#\gamma_{1,(q,1)}} \\ &\quad \times \left(\frac{1-P_{1,11}}{2-P_{1,00}-P_{1,11}}\right)^{l_{\text{frame1}}-\#\gamma_{1,(q,1)}} \end{aligned}$$

where $\#\gamma_1(ij)$ is defined as the cardinality of the set $\{\gamma_{1,(q,n)} = j|\gamma_{1,(q,n-1)} = i,\ q = 1,\ldots,l_{\text{frame1}},\ n = 2,\ldots,n_{\text{frame1}}\}$

$$p\left(P_{1,11}|\gamma_1,P_{1,00},\alpha_{P_{1,11}},\beta_{P_{1,11}}\right)\propto\frac{\mathcal{B}\left(P_{1,11}|\#\gamma_1(11)+\alpha_{P_{1,11}},\#\gamma_1(10)+l_{\text{frame1}}-\#\gamma_{1,(q,1)}+\beta_{P_{1,11}}\right)}{(2-P_{1,00}-P_{1,11})^{l_{\text{frame1}}}}$$

and $\#\gamma_{1,(q,1)}$ is the cardinality of the set $\{\gamma_{1,(q,1)}=1, q=1,\ldots,l_{\text{frame1}}\}$. Hence, we have

$$\begin{aligned}&p\left(P_{1,00}|\gamma_1,P_{1,11},\alpha_{P_{1,00}},\beta_{P_{1,00}}\right)\\&\quad\propto p(\gamma_1|P_{1,00},P_{1,11},\pi_1)p\left(P_{1,00}|\alpha_{P_{1,00}},\beta_{P_{1,00}}\right)\\&\quad\propto\frac{\mathcal{B}\left(P_{1,00}|\#\gamma_1(00)+\alpha_{P_{1,00}},\#\gamma_1(01)+\#\gamma_{1,(q,1)}+\beta_{P_{1,00}}\right)}{(2-P_{1,00}-P_{1,11})^{l_{\text{frame1}}}}\end{aligned}$$

$P_{1,00}$ can be updated using a M-H step, and we used the proposal distribution

$$\begin{aligned}&q\left(P_{1,00}|\gamma_1,\alpha_{P_{1,00}},\beta_{P_{1,00}}\right)\\&\quad=\mathcal{B}\left(P_{1,00}|\#\gamma_1(00)+\alpha_{P_{1,00}},\#\gamma_1(01)+\#\gamma_{1,(q,1)}+\beta_{P_{1,00}}\right).\end{aligned}$$

The acceptance probability $\alpha(P^{\star}_{1,00}|P_{1,00})$ of candidate $P^{\star}_{1,00}$ is simply

$$\alpha\left(P^{\star}_{1,00}|P_{1,00}\right)=\left(\frac{2-P_{1,00}-P_{1,11}}{2-P^{\star}_{1,00}-P_{1,11}}\right)^{l_{\text{frame1}}}\tag{24}$$

Similarly we have the following, as shown by the equation at the top of the page. $P_{1,11}$ can be updated using a M-H step, for which we use the proposal distribution

$$\begin{aligned}q\left(P_{1,11}|\gamma_1,\alpha_{P_{1,11}},\beta_{P_{1,11}}\right)=\mathcal{B}\left(P_{1,11}|\#\gamma_1(11)+\alpha_{P_{1,11}},\right.\\\left.\#\gamma_1(10)+l_{\text{frame1}}-\#\gamma_{1,(q,1)}+\beta_{P_{1,11}}\right)\end{aligned}$$

The acceptance probability $\alpha(P^{\star}_{1,11}|P_{1,11})$ of candidate $P^{\star}_{1,11}$ is simply

$$\alpha\left(P^{\star}_{1,11}|P_{1,11}\right)=\left(\frac{2-P_{1,00}-P_{1,11}}{2-P_{1,00}-P^{\star}_{1,11}}\right)^{l_{\text{frame1}}}\tag{25}$$

However, because of the exponent l_{frame1} in (24) and (25), the acceptance ratios can stay very low for long periods of time, yielding poorly mixing chains and long burn-in periods. Instead, we found very satisfying in practice to update the transitions probabilities $P_{1,00}$ and $P_{1,11}$ to the modes of their posterior distributions. After calculations of their derivatives, this simply amounts to root polynomials of order two and to choose the root with value lower to one. We favored this latter option in practice.

2) Vertical Markov Chain: We have

$$\begin{aligned}&p(\gamma_2|P_{2,00},P_{2,11},\pi_2)\\&\quad=\prod_{n=1}^{n_{\text{frame2}}}\prod_{q=2}^{l_{\text{frame2}}}p\left(\gamma_{2,(q,n)}|\gamma_{2,(q-1,n)},P_{2,00},P_{2,11}\right)\\&\qquad\times p\left(\gamma_{2,(1,n)}|\pi_2\right)\\&\quad=P_{2,00}^{\#\gamma_2(00)}(1-P_{2,00})^{\#\gamma_2(01)}P_{2,11}^{\#\gamma_2(11)}\\&\qquad\times(1-P_{2,11})^{\#\gamma_2(10)}\pi_2^{\#\gamma_{2,(1,n)}}(1-\pi_2)^{n_{\text{frame}}2-\#\gamma_{2,(1,n)}}\end{aligned}$$

where $\#\gamma_2(ij)$ is defined as the cardinality of the set $\{\gamma_{2,(q,n)}=j|\gamma_{2,(q-1,n)}=i,\ q=2,\ldots,l_{\text{frame2}},\ n=1,\ldots,n_{\text{frame2}}\}$ and $\#\gamma_{2,(1,n)}$ is the cardinality of the set $\{\gamma_{2,(1,n)}=1, n=1,\ldots,n_{\text{frame2}}\}$. Hence, we have

$$\begin{aligned}&p\left(P_{2,00}|\gamma_2,\alpha_{P_{2,00}},\beta_{P_{2,00}}\right)\\&\quad\propto p(\gamma_2|P_{2,00},P_{2,11},\pi_2)p\left(P_{2,00}|\alpha_{P_{2,00}},\beta_{P_{2,00}}\right)\\&\quad=\mathcal{B}\left(P_{2,00}|\#\gamma_2(00)+\alpha_{P_{2,00}},\#\gamma_2(01)+\beta_{P_{2,00}}\right).\end{aligned}$$

Similarly

$$\begin{aligned}&p\left(P_{2,11}|\gamma_2,\alpha_{P_{2,11}},\beta_{P_{2,11}}\right)\\&\qquad=\mathcal{B}\left(P_{2,11}|\#\gamma_2(11)+\alpha_{P_{2,11}},\#\gamma_2(10)+\beta_{P_{2,11}}\right)\end{aligned}$$

and finally

$$\begin{aligned}&p\left(\pi_2|\gamma_{2,(1,n)},\alpha_{\pi_2},\beta_{\pi_2}\right)\\&\qquad=\mathcal{B}\left(\pi_2|\#\gamma_{2,(1,n)}+\alpha_{\pi_2},n_{\text{frame2}}-\#\gamma_{2,(1,n)}+\beta_{\pi_2}\right)\end{aligned}$$

REFERENCES

[1] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 6, pp. 969–978, Jun. 1990.

[2] *ISO/IEC 13818-7 (MPEG-2 Advanced Audio Coding, AAC)*, ISO/IEC 13818-7 , Int. Org. Standard., 1997.

[3] S. Mallat and S. Zhang, "Matching pursuits with time–frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[4] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[5] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[6] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, Mar. 2003.

[7] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, Sep. 2003.

[8] M. E. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Process.*, vol. 86, no. 3, pp. 457–470, Mar. 2006.

[9] C. Févotte and S. J. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Process. Lett.*, vol. 13, no. 7, pp. 441–444, Jul. 2006.

[10] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Appl. Comput. Harmonic Anal.*, vol. 18, pp. 137–166, 2005.

[11] C. Tantibundhit, J. R. Boston, C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and A. A. El-Jaroudi, "Speech enhancement using transient speech components," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP06)*, 2006, pp. I-833–I-836.

[12] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularisation for time–frequency surface estimation," *J. R. Statist. Soc. Ser. B*, vol. 66, no. 3, pp. 575–589, 2004, read paper (with discussion).

[13] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, Mar. 2005, pp. V-517–V-520.

[14] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Process.*, vol. 82, no. 11, pp. 1595–1617, 2002.

[15] C. Févotte and S. Godsill, "A Bayesian approach to blind separation of sparse sources," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.

[16] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani, "Sparse regression with structured priors: Application to audio denoising," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, 2006, pp. III-57–III-60.
[17] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proc. Int. Conf. Signal Process. (ICSP'00)*, 2000, pp. 21–24.
[18] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, ser. Texts in Statistical Science, 2nd ed. London, U.K.: Chapman & Hall/CRC, 2004.
[19] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[20] J. Geweke, *Variable Selection and Model Comparison in Regression*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Swith, Eds., 5th ed. Oxford, U.K.: Oxford Univ. Press, 1996, pp. 609–620.
[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[22] [Online]. Available: http://www.tsi.enst.fr/fevotte/Samples/ieee_asl_denoising/
[23] M. Davy and S. Godsill, "Bayesian harmonic models for musical signal analysis," in *Seventh Valencia Int. Meeting (Bayesian Statistics VII)*. Oxford, U.K.: Oxford Univ. Press, 2002.
[24] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, pp. 197–208, 2000.
[25] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 173–185, Mar. 2002.
[26] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for non-linear time series," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 156–168, 2004.
[27] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Appl. Comput. Harmonic Anal.*, vol. 18, no. 2, pp. 137–166, Mar. 2005.
[28] M. A. Clyde and R. L. Wolpert, "Nonparametric function estimation using overcomplete dictionaries," in *Bayesian Statistics VIII*, J. Bernardo, J. Berger, A. Dawid, A. Smith, M. West, and D. Heckerman, Eds. Oxford, U.K.: Oxford Univ. Press, 2007, pp. 1–24.

Cédric Févotte was born in Laxou, France, in 1977. He graduated from the French engineering school École Centrale de Nantes, Nantes, France, received the Diplôme d'Études Approfondies en Automatique et Informatique Appliquée (M.Sc.) degree and the Diplôme de Docteur en Automatique et Informatique Appliquée (Ph.D.) degree jointly from the École Centrale de Nantes and the Université de Nantes in 2000 and 2003, respectively.

From November 2003 to March 2006, he was a Research Associate with the Signal Processing Laboratory, University of Cambridge, Cambridge, U.K., working on Bayesian approaches to many audio signal processing tasks such as audio source separation, denoising, and feature extraction. From May 2006 to February 2007, he was a Researcher with the startup company Mist-Technologies, Paris, working on mono/stereo to 5.1 surround sound upmix solutions. In March 2007, he joined the Département Signal/Images, GET/Télécom Paris (ENST), where his interests generally concern statistical signal processing and unsupervised machine learning with audio applications.

Bruno Torrésani was born in Marseilles, France, in 1961. He received the Ph.D. degree in theoretical physics from the Université de Provence, Marseille, France, in 1986 and the Habilitation degree from the Université de la Méditerranée, Marseille, in 1993.

From 1989 to 1997 he was a Researcher at the French Centre National de la Recherche Scientifique (CNRS). He is currently a Professor at Université de Provence, with joint appointment in the physics and mathematics departments. His current research interests are mainly in the domains of mathematical signal processing, with emphasis on harmonic analysis and probability-based approaches, and applications in audio signal processing and computational biology. His other domains of expertise are mathematical physics and electromagnetic scattering theory. He is currently Associate Editor of *Applied and Computational Harmonic Analysis*, the *International Journal of Wavelets and Multiresolution Information Processing*, and the *Journal of Signal, Image and Video Processing*.

Laurent Daudet a former physics student at the École Normale Supérieure, Paris, France, received the Ph.D. degree in mathematical modeling from the Université de Provence, Marseille, France, in 2000, in audio signal representations.

In 2001 and 2002, he was a Marie Curie Postdoctoral Fellow in the Centre for Digital Music at Queen Mary, University of London, London, U.K. Since 2002, he has been working as an Assistant Professor at the Pierre and Marie Curie University—Paris 6, where he joined the Laboratoire d'Acoustique Musicale, now part of the D'Alembert Institute for Mechanical Engineering. His research interests include audio coding, time–frequency and time-scale transforms, and sparse representations for audio.

Simon J. Godsill is a Professor of Statistical Signal Processing in the Engineering Department, Cambridge University, Cambridge, U.K. He runs a research group in Signal Inference and its Applications, with special interest in Bayesian and statistical methods for signal processing, Monte Carlo algorithms for Bayesian inference, modeling and enhancement of audio and musical signals, tracking, and high-frequency finance. He has published extensively in journals, books and conferences. He has coedited several journal special issues on topics related to Monte Carlo methods and Bayesian inference.

Prof. Godsill was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and on the IEEE Signal Processing Theory and Methods Committee. He was Scientific Chair and local host of the recent IEEE NSSPW workshop in Cambridge (Sep. 2006) and will coorganize a year-long SAMSI Institute research program on Sequential Monte Carlo methods from 2008 to 2009. He has been on the scientific committees of numerous international conferences and workshops, including EUSIPCO 2007 and 2008, and he has organized special sessions ranging from audio processing topics to probability and inference.