SEQUENTIAL MCMC METHODS FOR AUDIO SIGNAL ENHANCEMENT

Rubén M. Clavería,[1*] *Simon J. Godsill,*[1]

[1] Department of Engineering, University of Cambridge, CB2 1PZ, UK
rmc83@cam.ac.uk; sjg@eng.cam.ac.uk

ABSTRACT

With the aim of addressing audio signal restoration as a sequential inference problem, we build upon Gabor regression to propose a state-space model for audio time series. Exploiting the structure of our model, we devise a sequential Markov chain Monte Carlo algorithm to explore the sequence of filtering distributions of the synthesis coefficients. The algorithm is then tested on a series of denoising examples. Results suggest that the sequential approach is competitive with batch strategies in terms of perceptual quality and signal-to-noise ratio, while showing potential for real-time applications.

Index Terms— Audio denoising, Gabor frames, Bayesian filtering, Sequential MCMC

1. INTRODUCTION

Sparse dictionary representations [1, 2] are at the core of many of the advances in the field of signal processing in the last decades, such as signal compression [3] and denoising [4]. Bayesian extensions, in which assumptions about regression coefficients are incorporated through prior distributions, have been successfully applied to the analysis of images, e.g. [5], and audio time series [6]. Crucially, structural information of the transform domain, such as wavelet hierarchies in image analysis, can be readily included in these models, yielding advantages over simpler unstructured sparsity approaches.

1.1. Problem statement and contribution

Gabor regression [6, 7, 8] is a dictionary representation endowed with a prior structure designed to favour smoothness of the estimated function and sparseness of the coefficients. Inference in these models is typically performed in a batch fashion through Markov chain Monte Carlo [9] (MCMC). With real time applications in mind, we investigate sequential alternatives to these batch algorithms. Taking the assumptions of Gabor regression as a starting point (namely, an atomic decomposition of the signal, structured sparsity and heavy-tailed priors on synthesis coefficients), we formulate a state-space model for the parameters of interest. We propose a sequential Markov chain Monte Carlo strategy [10, 11, 12] to approximate the filtering distributions of time-frequency coefficients and reconstruct a degraded input signal. The scheme is then tested in background noise reduction tasks, where it proves to be competitive with batch techniques for similar models [6, 13] in terms of perceptual quality. Thus, our contribution is the foundation and validation of the sequential approach rather than an attempt to outperform state-of-the-art batch denoising methods (e.g. [14]). Potential enhancements of this setting are discussed in Section 5.

1.2. Related work

Relevant work in sequential time-frequency inference can be found, for example, in [15], where a bank of switching oscillators is used to model the occurrence of musical notes, or [16], where the Particle Filter [17, 18] is used to carry out frequency tracking in audio signals. In both cases, inference relies on the existence of a state-space model and is performed sequentially. Importantly, [19] presents a general state-space model for audio time series (referred to as *Gaussian time-frequency representation*) that, under specific assumptions, yields probabilistic analogues of classic notions like filter banks and Short-term Fourier transform or recover previous models such as the Probabilistic Phase Vocoder [20]. Links with classic methods such as Wiener and Kalman filtering are also explored. Lastly, [21] provides an up-to-date review of time-frequency state-space models.

2. A SEQUENTIAL APPROACH TO GABOR REGRESSION

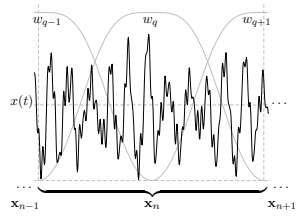
A window function $w(t)$ is a symmetric smooth function that is zero-valued outside an interval of length W, e.g. the Hann window. We define Gabor atoms as frequency- and time-shifted replicas of $w(t)$, as in $g_{m,q}(t) = w(t - h \cdot q)\ \exp(2\pi \mathrm{i} \frac{m}{M} t)$, where h a fixed "hop" in the time domain. In this work, we adopt $h = \frac{W}{2}$, generating an overcomplete dictionary with a 50% overlap in the time domain. Shifts are determined by indices m (frequency) and q (time). In Gabor regression [6], signals are represented as a weighted sum of Gabor atoms: $x(t) = \sum_{m=0}^{M-1} \sum_{q=0}^{Q-1} \tilde{c}_{m,q} g_{m,q}(t)$. Due to Hermitian symmetry arguments of atoms $g_{m,q}$, it can be shown that the range of m can be limited to $m = 0, \ldots, \frac{M}{2}$. Moreover, the original complex coefficients ($m \neq 0$) can be replaced by the real representation $c_{m,q} = (2\,\mathrm{Re}\{\tilde{c}_{m,q}\}, -2\,\mathrm{Im}\{\tilde{c}_{m,q}\}) \in \mathbb{R}^2$ so that the contribution of atom $g_{m,q}$ is written as $c_{m,q}(1)\,\mathrm{Re}\{g_{m,q}\} + c_{m,q}(2)\,\mathrm{Im}\{g_{m,q}\}$.

2.1. Observation model

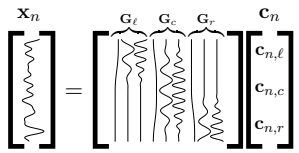
Consider a partition of $x(t)$ into $N = \frac{T}{W}$ segments $\mathbf{x}_1, \ldots, \mathbf{x}_N$, $\mathbf{x}_n \in \mathbb{R}^W$. Due to the limited support of $w(t)$ in the time domain, each chunk $\mathbf{x}_n$ is determined by the contributions of a small subset of atoms rather than the whole Gabor dictionary (Figure 1a). With $M = W$ and a 50% window overlap, sinusoids $\exp(2\pi \mathrm{i} \frac{m}{W} t)$ are periodic with period W. Consequently, atoms $g_{m,q}$ and $g_{m,q+2}$ are simply time-shifted replicas of the same set of non-zero values, and a similar argument holds for atoms "in between" (i.e. $g_{m,q-1}$, $g_{m,q+1}$). Thus, a single set of basis functions of length W can be used to construct all segments $\mathbf{x}_n$. This point is illustrated in Figure 1b, where $\mathbf{x}_n$ is represented as the product of $\mathbf{G} = [\mathbf{G}_\ell, \mathbf{G}_c, \mathbf{G}_r]$, a matrix of basis functions common to all n, and a column vector $\mathbf{c}_n$ containing the relevant coefficients.

With $\tilde{\mathbf{g}}_m \in \mathbb{C}^{W \times 1}$ the non-zero values of $g_{m,0}(t) = w(t) \exp(-2\pi \mathrm{i} \frac{m}{M} t)$ arranged in column form, we match the real-valued representation previously introduced by defining the atom $\mathbf{g}_{m,c} \in \mathbb{R}^{W \times 2}$ as $\mathbf{g}_{m,c} = [\mathrm{Re}(\tilde{\mathbf{g}}_m), \mathrm{Im}(\tilde{\mathbf{g}}_m)]$. The central matrix in Figure 1b is then defined as $\mathbf{G}_c = [\mathbf{g}_{0,c}, \ldots, \mathbf{g}_{\frac{M}{2},c}]$. To account for the contribution of the adjacent atoms, we define matrices $\mathbf{G}_\ell$ and $\mathbf{G}_r$ (*left* and *right*, respectively) as time-shifted versions of $\mathbf{G}_c$ padded with zeros, as shown in Figure 1b. Denoting $c_{m,n,p} \in \mathbb{R}^2$ the coefficient associated to frequency m, segment n, position $p \in \{\ell, c, r\}$, we construct vectors $\mathbf{c}_{n,\ell}$, $\mathbf{c}_{n,c}$, $\mathbf{c}_{n,r}$ by stacking coefficients $c_{m,n,p}$ in column form. Assuming the observed signal $y(t)$ is a version of $x(t)$ corrupted by additive Gaussian white

*Thanks to ANID Chile - Beca Doctorado en el Extranjero for funding.



(a) $\mathbf{x}_n$ and adjacent segments. Windows $w_q(t) = w(t - h \cdot q)$ (plotted alongside signal x) determine the support of atoms $g_{m,q}$.



(b) Chunks $\mathbf{x}_n$ are aligned with the central set of atoms $\mathbf{G}_c$ and receive contributions from $\mathbf{G}_\ell$, $\mathbf{G}_c$, $\mathbf{G}_r$. Vector $\mathbf{c}_n$ determines how much each atom contributes to $\mathbf{x}_n$.

noise, the observation equation is:

$$p(\mathbf{y}_n|\mathbf{c}_n, \sigma_n^2) = \mathcal{N}\Big(\mathbf{y}_n; \overbrace{[\mathbf{G}_\ell\ \mathbf{G}_c\ \mathbf{G}_r]}^{\mathbf{G}}\mathbf{c}_n, \sigma_n^2\mathbf{I}_W\Big) \quad (1)$$

where terms $\mathbf{y}_n$ are chunks of $y(t)$ (analogous to $\mathbf{x}_1, \mathbf{x}_2, \dots$), σ_n^2 is the noise level in $\mathbf{y}_n$ and $\mathbf{I}_W \in \mathbb{R}^{W\times W}$ is the identity matrix.

2.2. State model

In this section, we propose a prior structure for coefficients $\mathbf{c}_n$ and nuisance parameters (e.g. σ_n^2) that inherits the desirable qualities of batch Gabor regression, such as structured sparsity, but has the advantage of being suitable for sequential estimation.

Coefficient structure: Due to atom overlap, association between $\mathbf{y}_n$ and an underlying state is not straightforward, as each signal chunk $\mathbf{y}_n$ "shares" atoms with the adjacent segments (Figure 1a). It must be noted, though, that the definition of a distinct set of coefficients $\mathbf{c}_n$ for each segment (as described in Eq. 1) with the restriction $p(\mathbf{c}_{n,\ell}|\mathbf{c}_{n-1,r}) = \delta(\mathbf{c}_{n,\ell} - \mathbf{c}_{n-1,r})$ facilitates the formulation of a state-space model, as in that case each $\mathbf{c}_n$ is unambiguously associated to a single observation $\mathbf{y}_n$. We describe that structure as *redundant* in the sense that, in a batch setting, the definition of variable $\mathbf{c}_{n,\ell}$ as a copy of $\mathbf{c}_{n-1,r}$ would be unnecessary.

Structured sparsity: Audio signals are remarkably structured in the spectral domain. Typically, only a reduced subset of frequencies are present, and they appear in the form of highly distinct patterns in the time-frequency surface. Moreover, as the spectral content varies over time, most frequencies just remain active for a short time interval. Hence it is expected that only a subset of time-frequency Gabor atoms are used in the reconstruction task. We model this assumption by assigning each $c_{m,n,p}$ a prior probability of being exactly 0. To this end, we define the indicator variables $\gamma_{m,n,p}$ and adopt a spike-and-slab density [22, 23] on $c_{m,n,p}$:

$$p(c_k|\gamma_k, s_k) = (1-\gamma_k)\delta(c_k) + \gamma_k\,\mathcal{N}\left(c_k; \mathbf{0}_2, s_k\mathbf{I}_2\right), \quad (2)$$

where k is used to represent an index set m, n, p concisely. Eq. 2 establishes that $c_k = 0$ if $\gamma_k = 0$ (the *spike* distribution $\delta(c_k)$), and drawn from a Gaussian distribution with covariance $s_k\mathbf{I}_2$ if $\gamma_k = 1$ (the *slab* distribution). Variance s_k is discussed later in this section.

Empirically observed time-frequency patterns can be incorporated into the model by adopting different priors on $\boldsymbol{\gamma}$ (e.g. Ising, Markov chains, etc.). In this work, however, we limit our attention to Markov chain densities along n, as they mirror the steadiness of tonal components of music over time. We define Markov chain dependencies along the n axis for each frequency m, with each chain being independent *a priori*. Formally, for each segment n we have that $p(\gamma_{m,n,\ell}|\gamma_{m,n-1,r}) = \delta(\gamma_{m,n,\ell} - \gamma_{m,n-1,r})$ (analogous to the redundant structure of regression coefficients), whereas $p(\gamma_{m,n,c}|\gamma_{m,n,\ell})$ and $p(\gamma_{m,n,r}|\gamma_{m,n,c})$ are governed by the Markov chain transition probabilities ϕ_{00}, ϕ_{11}. For example, $p(\gamma_{m,n,c} = 1|\gamma_{m,n,\ell} = 1) = \phi_{11}$, $p(\gamma_{m,n,r} = 1|\gamma_{m,n,c} = 0) = 1 - \phi_{00}$, etc. For simplicity, ϕ_{00}, ϕ_{11} are treated as hyperparameters, hence assumed fixed during sequential inference.

Heavy-tailedness: Adopting eavy-tailed priors on coefficients is a widely utilised device to induce sparsity in linear regression models (e.g., [24, 25]) and can be used in combination with spike-and-slab priors. In audio enhancement tasks, this assumption is also justified on the grounds of reflecting the wide range of values that time-frequency coefficients take in real audio signals [6]. We impose a heavy-tailed prior on c_k through the conditionally Gaussian structure known as *Scaled Mixture of Normals* [26] (SMiN). In SMiN, a heavy-tailed density $p(x)$ is represented as the marginal of a Gaussian with variable variance v: $p(x) = \int_{\mathbb{R}^+}\mathcal{N}(x; 0, v)p(v)\mathrm{d}v$, where $p(v)$ is known a *mixing distribution*. We impose a Student's t-prior on $c_{m,n,p}$ by adopting an Inverse Gamma prior $p(s_{m,n,p}|\nu_n) = \mathcal{IG}(s_{m,n,p}; \kappa, \nu_n)$ on $s_{m,n,p}$ as the mixing distribution. Due to our redundant coefficient structure, the restriction $s_{m,n,\ell} = s_{m,n-1,r}$ is imposed, akin to $c_{m,n,\ell}$ and $\gamma_{m,n,\ell}$.

Global parameters: Scale parameter ν is related to the level of dispersion of synthesis coefficients. In practice, it is crucial to tune ν automatically, as the range of values of $c_{m,n,p}$ varies largely depending on the spectral characteristics of signal $x(t)$, its intensity, and the specific choices of $w(t)$ and W. The formulation of ν and σ^2 as global parameters (as in [6, 13]) complicates sequential inference due to their dependency upon on the whole sequence of states and observations. Moreover, in many real-life settings noise level σ^2 and signal intensity–indirectly modelled by ν–may exhibit temporal variations for which the assumption of a global ν and σ^2 is overly inflexible. To facilitate online adaptation to these variations, we replace global variables ν, σ^2 with slowly varying sequences $\nu_{1:N}$ and $\sigma_{1:N}^2$. Parameters $\nu_{1:N}$, $\sigma_{1:N}^2$ play essentially the same role as ν, σ^2, but are segment-dependent—each element is associated to a signal chunk $\mathbf{y}_n$. This mapping is arbitrary, as variations in volume and noise may occur at any time and are not necessarily aligned with the partition $\mathbf{y}_{1:N}$. However, as variations are expected to occur at a steady pace, we argue that ν_n and σ_n^2 can represent local levels of signal intensity and noise appropriately. To encourage smoothness along the time axis, we set Gamma and inverse-Gamma Markov chain priors on $\nu_{1:N}$ and $\sigma_{1:N}^2$, respectively: $p(\nu_n|\nu_{n-1}) = \mathcal{G}(\nu_n|\alpha_\nu, \frac{\alpha_\nu}{\nu_{n-1}})$, $p(\sigma_n^2|\sigma_{n-1}^2) = \mathcal{IG}\left(\sigma_n^2|\alpha_\sigma, \sigma_{n-1}^2(\alpha_\sigma - 1)\right)$. Values $(\frac{\alpha_\nu}{\nu_{n-1}})$ and $\sigma_{n-1}^2(\alpha_\sigma - 1)$ are chosen so that $\mathbb{E}_{\nu\sim p(\nu_n|\nu_{n-1})}(\nu_n) = \nu_{n-1}$ and $\mathbb{E}_{\sigma_n^2\sim p(\sigma_n^2|\sigma_{n-1}^2)}(\sigma_n^2) = \sigma_{n-1}^2$, to encourage continuity.

Joint model: Putting all densities together, we have:

$$p(\mathbf{c}_{1:N}, \mathbf{s}_{1:N}, \boldsymbol{\gamma}_{1:N}, \sigma_{1:N}^2, \nu_{1:N}, \mathbf{y}_{1:N}) =$$
$$\left[p(\mathbf{y}_1|\mathbf{c}_1, \sigma_1^2)p(\mathbf{c}_1|\boldsymbol{\gamma}_1, \mathbf{s}_1)p(\mathbf{s}_1|\nu_1)p(\sigma_1^2)p(\gamma_1)p(\nu_1)\right]\cdot\dots \quad (3)$$
$$\prod_{n=2}^{N}\left[p(\mathbf{y}_n|\mathbf{c}_n, \sigma_n^2)p(\mathbf{c}_n|\mathbf{c}_{n-1}, \boldsymbol{\gamma}_n, \mathbf{s}_n)p(\mathbf{s}_n|\nu_n)p(\sigma_n^2|\sigma_{n-1}^2)\right.$$
$$\left.p(\gamma_n|\gamma_{n-1})p(\nu_n|\nu_{n-1})\right],$$

which has a state-space model structure. In particular, coefficients can be described in terms of a state evolution equation as in:

$$\mathbf{c}_n = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{I}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{c}_{n-1} + \boldsymbol{\omega}_n, \quad \boldsymbol{\omega}_n = \begin{bmatrix}\boldsymbol{\omega}_{n,\ell}\\ \boldsymbol{\omega}_{n,c}\\ \boldsymbol{\omega}_{n,r}\end{bmatrix},$$

where process noise $\boldsymbol{\omega}_n$ satisfies $\omega_{m,n,p} \sim \mathcal{N}(\mathbf{0}, s_{m,n,p}\mathbf{I}_2)$ when $\gamma_{m,n,p} = 1$ and $p \in \{c, r\}$; and $\omega_{m,n,p} = 0$ if $p = \ell$ or $\gamma_{m,n,p} = 0$.

3. COMPUTING THE MODEL

We define the state $q_n = \{\mathbf{c}_n, \boldsymbol{\gamma}_n, \mathbf{s}_n, \nu_n, \sigma_n^2\}$ and consider each segment $\mathbf{y}_n$ as an observation. Our aim is to devise a scheme to estimate the sequence of filtering distributions $p(q_n|\mathbf{y}_{1:n})$. As the coefficient prior (Eq. 2) is of little predictive value, using it as an importance sampling distribution (e.g. [17]) would hardly place samples in a zone of high probability. Moreover, due to the high dimensionality of the problem, the importance sampling-based Particle filter

may struggle to provide a reliable estimation of the target distribution. The sequential MCMC method outlined in [10] and later popularised as *MCMC-based Particle filter* [11] has proved useful for high-dimensional problems, outperforming traditional Particle filters in certain scenarios [12].

Sequential MCMC (sMCMC hereinafter) relies on MCMC rather than importance sampling to explore the filtering distribution. To avoid the high expense of marginalisation with respect to q_{n-1} [11], the non-marginalised version [12] targets the joint distribution $p(q_{n-1}, q_n|y_{1:n})$ and subsequently extracts the values q_n—an alternate way to approximate $p(q_n|y_{1:n})$. To this end, the non-marginalised sMCMC [12] exploits the factorisation:

$$p(q_{n-1}, q_n|y_{1:n}) \propto p(q_{n-1}|y_{1:n-1})p(q_n|q_{n-1})p(y_n|q_n)$$

and $\hat{p}(q_{n-1}|y_{1:n-1}) = \frac{1}{P}\sum_{j=1}^{P}\delta(q_{n-1} - q_{n-1}^{(j)})$ (a particle representation of the filtering distribution) to sample $p(q_{n-1}, q_n|y_{1:n})$ approximately. Conditionals $p(q_n|q_{n-1}, y_{1:n}) \propto p(q_n|q_{n-1}) \cdot p(y_n|q_n)$ and $\hat{p}(q_{n-1}|q_n, y_{1:n}) \propto p(q_n|q_{n-1}) \cdot \hat{p}(q_{n-1}|y_{1:n-1})$ are sampled alternately for $N_{\text{burn-in}}+P$ iterations, generating a collection $\{q_{n-1}^{(j)}, q_n^{(j)}\}_{j=1,\dots,P}$ approximately drawn from $p(q_{n-1}, q_n|y_{1:n})$. As $p(q_n|q_{n-1}, y_{1:n})$ does not have a closed-form expression in the general case, practical implementations resort to Gibbs sampling (either component-wise or blocked), Metropolis-Hastings steps or a combination of them. Optionally, refinement moves can be incorporated in order to improve the mixing of the Markov chain [10, 12].

3.1. Sequential MCMC for audio signal enhancement

The approach adopted in this work is similar to [12] in that we sample $p(q_{n-1}, q_n|\mathbf{y}_{1:n})$ rather than $p(q_n|\mathbf{y}_{1:n})$ directly, but differs in how the sampling is carried out. Firstly, we leverage the conjugacy of the priors adopted in Section 2 to derive exact Gibbs updates, rather than Metropolis-Hastings steps, to sample the components of q_n. As for q_{n-1}, an appealing feature of the non-marginalised sMCMC is that q_{n-1} does not need to be sampled analytically from its conditional. Instead, particles in $\hat{p}(q_{n-1}|y_{1:n-1}) = \frac{1}{P}\sum_{j=1}^{P}\delta(q_{n-1} - q_{n-1}^{(j)})$ can be re-used: to sample from the approximate conditional $\hat{p}(q_{n-1}|q_n, y_{1:n}) \propto p(q_n|q_{n-1}) \cdot \hat{p}(q_{n-1}|y_{1:n-1})$, it suffices to sample particles $q_{n-1}^{(j)}$ proportional to $p(q_n|q_{n-1}^{(j)})$, thus reducing the sampling of q_{n-1} to a weighting procedure. The naive application of this approach to our model (Eq. 3), though, leads to poor signal reconstructions for the reasons detailed below:

1. As the rightmost coefficients $\mathbf{c}_{n-1,r}(= \mathbf{c}_{n,\ell})$ obtained at time n–1 do not incorporate the next observation $\mathbf{y}_n$, the pool of values available for $\mathbf{c}_{n,\ell}$ at time n is particularly poor. Thus, making joint draws of $\mathbf{c}_{n-1,\{\ell,c,r\}}$ in the manner described above (i.e. taking values from $\hat{p}(q_{n-1}|y_{1:n-1})$) leads to unsatisfactory reconstructions. This can be alleviated by recalculating the probabilities of $\mathbf{c}_{n-1,r} = \mathbf{c}_{n,\ell}$ in the light of new observation $\mathbf{y}_n$ at time n—a form of partial smoothing.

2. As restriction $p(\mathbf{c}_{n,\ell}|\mathbf{c}_{n-1,r}) = \delta(\mathbf{c}_{n,\ell} - \mathbf{c}_{n-1,r})$ affects the leftmost coefficients of $\mathbf{c}_n$, the "weighting" of particles $q_{n-1}^{(j)}$ does not only depend on $p(q_n|q_{n-1})$ but also on the likelihood $p(\mathbf{y}_n|\mathbf{c}_n, \sigma_n^2)$. As $\frac{1}{\sigma_n^2}\|\mathbf{y}_n - \mathbf{G}\mathbf{c}_n\|^2$ takes large values, it renders the exponential term in Eq. 1 almost infinitesimally small. This hampers particle diversity in the way described as follows: let W_j be the raw weights of particles $q_{n-1}^{(j)}$ (i.e. the relevant densities, including $p(\mathbf{y}_n|\mathbf{c}_n, \sigma^2)$, evaluated at $q_{n-1}^{(j)}$, q_n) and $j^\star$ the index of the particle with the maximum raw weight. As even small shifts in $\mathbf{c}_{n-1,\ell}$ result in a change of several orders of magnitude of weights W_j, we have that $\sum_{j=1}^{P}\mathrm{W}_j = \mathrm{W}_{j^\star} + \epsilon$ (ϵ negligible) and $\hat{p}(q_{n-1}^{(j^\star)}|q_n, \mathbf{y}_{1:n}) = \frac{\mathrm{W}_{j^\star}}{\sum_{j=1}^{P}\mathrm{W}_j} \approx 1$, whereas the probabilities of the rest of the particles are nearly 0. Updates of q_n over the MCMC iterations barely affect this dynamics: if a best fitting particle $q_{n-1}^{(j^\star)}$ is sampled once, it is unlikely that the Markov chain escapes that value in the subsequent iterations.

To address these issues, we split the state q_{n-1} into two subsets: q'_{n-1}, whose elements are taken from the pre-existing collection of particles $\hat{p}(q_{n-1}|\mathbf{y}_{1:n-1})$; and q''_{n-1}, whose elements are sampled via exact Gibbs updates alongside the current state q_n. This scheme improves the sampling of $p(q_{n-1}, q_n|\mathbf{y}_{1:n})$ but entails an increased computational cost due to the extra Gibbs steps. A desirable condition for q''_{n-1} is to contain $(\boldsymbol{\gamma}_{n-1,r}, \mathbf{c}_{n-1,r}) = (\boldsymbol{\gamma}_{n,\ell}, \mathbf{c}_{n,\ell})$ (as these are the variables most affected by $\mathbf{y}_n$) while avoiding to sample the whole state q_{n-1} analytically. Setting $q'_{n-1} = \boldsymbol{\gamma}_{n-1,\{\ell,c\}}$, $q''_{n-1} = \{\mathbf{c}_n, \boldsymbol{\gamma}_{n-1,r}, \sigma_{n-1}^2, \nu_{n-1}\}$, the sampling strategy described as follows offers a good compromise between computational efficiency and quality of the generated samples. At time step n:

1. **Initialisation**. Given $\hat{p}(q_{n-1}|\mathbf{y}_{1:n-1})$, marginal distributions of any subset of q_{n-1} can be approximated simply by picking the relevant coordinates of particles $\{q_{n-1}^{(j)}\}_{j=1,\dots,P}$. In particular, $\hat{p}(\boldsymbol{\gamma}_{n-1,\{\ell,c\}}|y_{1:n-1}) = \frac{1}{P}\sum_{j=1}^{P}\delta(\boldsymbol{\gamma}_{n-1,\{\ell,c\}} - \boldsymbol{\gamma}_{n-1,\{\ell,c\}}^{(j)})$. The initial value of $\boldsymbol{\gamma}_{n-1,\{\ell,c\}}$ is set by choosing a random particle from this approximate distribution, then $\sigma_{n-1:n}^2$ and $\nu_{n-1:n}$ are drawn from their priors. Coefficients $\mathbf{c}$, $\boldsymbol{\gamma}$ in q''_{n-1} and q_n are initialised at 0.

2. **Gibbs sweep**. Elements σ_{n-1}, ν_{n-1}, $\mathbf{s}_{n-1,r} = \mathbf{s}_{n,\ell}$ $\mathbf{c}_{n-1,r} = \mathbf{c}_{n,\ell}$, $\boldsymbol{\gamma}_{n-1,r} = \boldsymbol{\gamma}_{n,\ell}$ (subset q''_{n-1}) and $\mathbf{c}_{\{c,r\}}$, $\boldsymbol{\gamma}_{n,\{c,r\}}$, $\mathbf{s}_{n,\{c,r\}}$, ν_n, σ_n^2 (state q_n) are sampled from their full conditionals in a sequential manner, given $\mathbf{y}_{1:n}$ and the current value of q'_{n-1}. The component-wise Gibbs updates derived from the model of Eq. 3 are standard and to a great extent akin to those of [6, 13]. While $\mathbf{s}$, ν, σ^2 can be sampled from their conditionals with little computational effort, the exploration of the coefficient space $(\boldsymbol{\gamma}, \mathbf{c})$ in positions (n, ℓ), (n, c), (n, r) and along all frequencies m is the bulk of the algorithm, as it involves the sequential sampling of $3(\frac{M}{2} + 1)$ pairs $(\gamma_{m,n,p}, c_{m,n,p})$ ($p \in \{\ell, c, r\}$) per iteration.

3. **Joint draws of $\boldsymbol{\gamma}_{n-1,\ell}$, $\mathbf{c}_{n-1,\ell}$ and $\boldsymbol{\gamma}_{n-1,c}$, $\mathbf{c}_{n-1,c}$.** While probability $p(\mathbf{c}_{n-1,r}, \boldsymbol{\gamma}_{n-1,r}|\mathbf{y}_{1:n-1})$ changes significantly if $\mathbf{y}_n$ is incorporated, it is empirically observed that the effect of $\mathbf{y}_n$ on the leftmost and central indicators is minor: $p(\boldsymbol{\gamma}_{n-1,\{\ell,c\}}|\mathbf{y}_{1:n-1}) \approx p(\boldsymbol{\gamma}_{n-1,\{\ell,c\}}|\mathbf{y}_{1:n})$. Moreover, given $\boldsymbol{\gamma}_{n-1,c}$, coefficients $\mathbf{c}_{n-1,c}$ can be easily sampled from their conditional Gaussian distribution: $p(\mathbf{c}_{n-1,c}|\boldsymbol{\gamma}_{n-1,c}, \mathbf{s}_{n-1,c}, \mathbf{y}_{n-1}, \dots) = \mathcal{N}(\mathbf{c}_{n-1,c}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are functions of $\boldsymbol{\gamma}_{n-1,c}$, $\mathbf{c}_{n-1,\{\ell,r\}}$, $\mathbf{s}_{n-1,c}$ $\mathbf{y}_{n-1}$ and can be calculated easily. As most indicators are 0 in typical cases, the conditional Gaussian deals with a set of coefficients of reduced dimension. Sampling from this conditional is then dramatically more efficient than sweeping over m sequentially. Hence we adopt the proposal:

$$Q(\boldsymbol{\gamma}_{n-1,c}, \mathbf{c}_{n-1,c}) = \frac{1}{P}\sum_{j=1}^{P}\delta(\boldsymbol{\gamma}_{n-1,c} - \boldsymbol{\gamma}_{n-1,c}^{(j)})\mathcal{N}(\mathbf{c}_{n-1,c}|\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)}).$$

to make joint draws of $\boldsymbol{\gamma}_{n-1,c}$, $\mathbf{c}_{n-1,c}$ (analogous for $\boldsymbol{\gamma}_{n-1,\ell}$, $\mathbf{c}_{n-1,\ell}$). The proposed values are then accepted or rejected according to a Metropolis-Hastings acceptance probability using Eq. 3. As refreshing $\mathbf{c}_{n-1,\{\ell,c\}}$ on every iteration introduces variability, the problem of particle diversity is overcome. Since we leverage the knowledge of $\hat{p}(\boldsymbol{\gamma}_{n-1,\{\ell,c\}}|\mathbf{y}_{1:n-1})$ to propose feasible values for the indicator $\boldsymbol{\gamma}$, the sampling of $\mathbf{c}$ is done with little computational effort.

4. **Termination condition**: if the iteration count is less than the maximum number of MCMC steps, go to step **2.**, otherwise stop.

The approximate filtering distribution $\hat{p}(\mathbf{c}_{n,\ell}|\mathbf{y}_{1:n})$ is used to estimate $\mathbf{c}_{n,\ell}$ as $\hat{\mathbf{c}}_{n,\ell} = \frac{1}{P}\sum_{j=1}^{P}\mathbf{c}_{n,\ell}^{(j)}$, on the grounds that this marginal does not change much once observation $\mathbf{y}_{n+1}$ arrives. The

Table 1. Batch inference

	SNR_{in} [dB]	Length ($N_{\text{it}}/N_{\text{burn-in}}$)	SNR_{out} [dB] (σ^2, ν)	SNR_{out} [dB] ($\sigma^2_{1:N}, \nu_{1:N}$)
Glockenspiel ($\sim$ 3 seconds)	10	40/20	20.58	19.88
		800/100	21.42	20.44
Trumpet ($\sim$ 4 seconds)	10	40/20	18.35	17.12
		800/100	19.14	18.75
Vibraphone ($\sim$ 4 seconds)	10	40/20	20.34	20.56
		800/100	22.23	21.89
Oratorio ($\sim$ 4 seconds)	10	40/20	19.15	17.21
		800/100	19.31	19.12

estimate $\mathbf{c}_{n,r} = \mathbf{c}_{n+1,\ell}$ is not updated until the next time step $n+1$. As $\mathbf{c}_{n,c}$ is included in q_n'' and refreshed at time $n+1$, $\hat{c}_{n,c}$ is obtained from its "smoothed" version (i.e., values obtained at time $n+1$ rather than samples from $\hat{p}(q_n|\mathbf{y}_{1:n})$), leading to better results in practice.

4. EXPERIMENTAL RESULTS

Two sets of experiments are carried out: in Section 4.1, the proposed model (Eq. 3) is tested against the original model where parameters σ^2, ν are global (e.g., as in [6]), with both models computed in a batch fashion. Section 4.2 compares the results obtained using batch methods with the sequential approach proposed in Section 3. Experiments are implemented on MATLAB and run on a personal computer with Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz and 16 GB RAM. Hyperparameters were set to $\phi_{00} =$ 1–$1/2000$, $\phi_{11} =$ 1–$1/50$, $\alpha_{\sigma^2} = 10$, $\alpha_\nu = 10$, $\alpha_\nu = 10$ $\kappa = 3$, and it was found that minor changes in these values do not alter the results noticeably. Audible results are available at https://rclaveria.github.io/sMCMC_audio/index.html.

4.1. Validation of the proposed model

While Eq. 3 has a state-space model structure that allows for sequential inference, batch MCMC strategies are still feasible, as the newly introduced parameters $\sigma^2_{1:N}, \nu_{1:N}$ can be updated through exact Gibbs steps. In this section, we study the impact of posing ν, σ^2 as slowly varying sequences instead of global parameters. The two models are tested on 4 audio excerpts affected by stationary additive noise, sampled at 44,100 Hz. Two different chain lengths are tested for each example. With the aim of studying the best performance achievable by each model, a long chain (800 iterations with 100 burn-in samples) is run, as little change is observed when the length surpasses that threshold. In turn, a short chain (40 iterations discarding the first 20 samples) is employed to assess how rapidly these algorithms produce a perceptually acceptable output. A window size of $W = 2048$ samples is used for the time-frequency atoms.

Results in Table 1 suggest that the formulation of ν, σ^2 as time-varying parameters implies a slight loss in terms of waveform signal-to-noise ratio (SNR), as the best performances achieved by the model with global parameters (fourth column) are better than those obtained by the time-varying model (fifth column). However, from a perceptual point of view this gap is almost unnoticeable, proving that the model provides a good compromise between output quality and model tractability—it is, after all, the introduction of time-varying parameters $\sigma^2_{1:N}, \nu_{1:N}$ that facilitates sequential processing. Moreover, as can be noted from the audible results, the model with $\sigma^2_{1:N}, \nu_{1:N}$ requires shorter runs to generate perceptually satisfactory reconstructions, with as few as 40 iterations being enough to virtually remove background noise, with little (if any) artifacts. Conversely, reconstructions based on the model with global ν, σ^2 may still contains some residual background noise at the same stage (check glockenspiel, trumpet and vibraphone examples). We conjecture that the dependency of global parameter σ^2 on all data points $y(t)$ entails slower convergence.

While noticeable, the loss associated to the formulation of σ^2 and ν as frame-dependent parameters is inconsequential in many scenarios of interest, for example in real-time settings where the aim is to ensure the overall intelligibility of the signal rather than a high-fidelity reconstruction. Moreover, the newly introduced model could be more suitable for signals with slowly varying background noise (quasi-stationary) or significant variations of the signal's volume—recall that ν controls the *a priori* "size" of coefficients, therefore the amplitude of the signal.

Table 2. Sequential inference

	SNR_{in} [dB]	Length ($N_{\text{it}}/N_{\text{burn-in}}$)	SNR_{out} [dB] (Batch)	SNR_{out} [dB] (Sequential)
Glockenspiel ($\sim$3 seconds)	10	80/20	20.13	19.29
Trumpet ($\sim$12 seconds)	15	80/20	21.46	20.03
Vibraphone ($\sim$8.5 seconds)	15	80/20	25.05	23.18
Vibraphone ($\sim$8.5 seconds)	20	80/20	28.79	26.96
Oratorio ($\sim$8.5 seconds)	15	80/20	21.45	19.84

4.2. Sequential audio signal processing

In this section, the sMCMC algorithm proposed in Section 3 is compared with batch MCMC. Both algorithms use the model of Eq. 3. Tests are carried out on 5 different audio examples of varying length and input signal-to-noise ratio. As in Section 4.1, all audio excerpts are sampled at 44,100 Hz and a window length of 2048 is adopted for the dictionary atoms. The analysis is limited to short chains (80 samples with a burn-in period of 20) as longer chains do not yield significantly higher SNR values nor noticeably better perceptual qualities for any of the algorithms tested (note that sMCMC does not need a large number of samples to produce good quality outputs). Audible and quantitative results (Table 2) indicate that, while there is a systematic gap between the the output SNR values obtained with both methods, reconstructions obtained via sequential MCMC still provide good quality audible results, as can be checked in the website (the vibraphone excerpt with SNR_{out} = 20 dB is a particularly good example). Moreover, extensions such as online impulse removal or interpolation of missing values are relatively straightforward.

5. FINAL REMARKS

A state-space model for audio signals is presented and tested on a series of music excerpts. The proposed model draws inspiration from existing work on Gabor regression [6] and keeps many of its desirable traits, such as structured sparsity. A sequential MCMC strategy is proposed to approximate the filtering distributions of the coefficients and calculate an estimate of the signal. Although this approach is proved useful to process the signal in an online fashion (unlike traditional batch methods found in the literature) and produces good quality reconstructions, many challenges are yet to be addressed, such as real-time performance (which the current version of the algorithm is still far from achieving) or online learning of transition probability parameters ϕ. Additional strategies to fill the performance gap between sequential and batch are also in demand, since differences in SNR and perceptual quality, although not critical, are still noticeable. This gap is partially explained by the fact that reconstructions are based on filtering distributions rather than whole state trajectories or fully smoothed states—after all, filtering distributions use incomplete information. In this regard, implementing larger sliding windows for online smoothing is a promising direction of research. However, this would require additional computational effort. Exploring the indicator space $\boldsymbol{\gamma}$–arguably the conundrum of the proposed algorithm–with enhanced sampling strategies such as the promising [27] may increase the efficiency of the algorithm. The use of approximate schemes like variational Bayes or expectation propagation may also be feasible alternatives.

6. REFERENCES

[1] Stéphane G Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] Shaobing Chen and David Donoho, "Basis pursuit," in *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1994, vol. 1, pp. 41–44.

[3] Ralph Neff and Avideh Zakhor, "Very low bit-rate video coding based on matching pursuits," *IEEE Transactions on circuits and systems for video technology*, vol. 7, no. 1, pp. 158–171, 1997.

[4] Felix Abramovich, Theofanis Sapatinas, and Bernard W Silverman, "Wavelet thresholding via a Bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 4, pp. 725–749, 1998.

[5] Matthew S Crouse, Robert D Nowak, and Richard G Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Transactions on signal processing*, vol. 46, no. 4, pp. 886–902, 1998.

[6] Patrick J Wolfe, Simon J Godsill, and Wee-Jing Ng, "Bayesian variable selection and regularization for time–frequency surface estimation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 3, pp. 575–589, 2004.

[7] Patrick J Wolfe and Simon J Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2005, vol. 5, pp. v–517.

[8] James Murphy and Simon Godsill, "Joint Bayesian removal of impulse and background noise," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 261–264.

[9] CP Robert and G Casella, "Monte Carlo statistical methods," *Springer, New York*, 2004.

[10] Carlo Berzuini, Nicola G Best, Walter R Gilks, and Cristiana Larizza, "Dynamic conditional independence models and Markov chain Monte Carlo methods," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1403–1412, 1997.

[11] Zia Khan, Tucker Balch, and Frank Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[12] François Septier, Sze Kim Pang, Avishy Carmi, and Simon Godsill, "On MCMC-based particle methods for Bayesian filtering: Application to multitarget tracking," in *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2009, pp. 360–363.

[13] Cédric Févotte, Bruno Torrésani, Laurent Daudet, and Simon J Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 174–185, 2007.

[14] Francois Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[15] Ali Taylan Cemgil, Hilbert J Kappen, and David Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 679–694, 2006.

[16] Richard Geoffrey Everitt, Christophe Andrieu, and Manuel Davy, "Online Bayesian inference in some time-frequency representations of non-stationary processes," *IEEE transactions on signal processing*, vol. 61, no. 22, pp. 5755–5766, 2013.

[17] Neil J Gordon, David J Salmond, and Adrian FM Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE proceedings F (radar and signal processing)*. IET, 1993, vol. 140, pp. 107–113.

[18] Simon Godsill, "Particle filtering: the first 25 years and beyond," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7760–7764.

[19] Richard E Turner and Maneesh Sahani, "Time-frequency analysis as probabilistic inference," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6171–6183, 2014.

[20] Ali Taylan Cemgil and Simon J Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *2005 13th European Signal Processing Conference*. IEEE, 2005, pp. 1–4.

[21] William J Wilkinson, Michael Riis Andersen, Joshua D Reiss, Dan Stowell, and Arno Solin, "Unifying probabilistic models for time-frequency analysis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3352–3356.

[22] Toby J Mitchell and John J Beauchamp, "Bayesian variable selection in linear regression," *Journal of the american statistical association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[23] Edward I George and Robert E McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[24] Trevor Park and George Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[25] Michael E Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[26] David F Andrews and Colin L Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[27] Michalis K Titsias and Christopher Yau, "The Hamming ball sampler," *Journal of the American Statistical Association*, vol. 112, no. 520, pp. 1598–1611, 2017.