An EM approach to variable selection and its application to audio restoration

Rubén M. Clavería, and Simon J. Godsill, *Senior Member, IEEE*

***Abstract*—Sparse representations have played a key role in the development of signal processing over the last decades, and have been widely applied to audio signal processing tasks such as denoising or compression. Heavy-tailed priors and ℓ_1-norm regularization schemes are a popular tool to promote sparsity on the model coefficients, among other reasons due to their suitability for efficient estimation strategies. On the other hand, spike-and-slab priors explicitly model the inclusion of each coefficient within the sparse set, facilitating the representation of arbitrarily complex networks of dependencies between variables—a desirable feature in many applications. However, devising fast optimization strategies to compute this class of models has typically been an elusive task; in most cases, practitioners must resort to costly Monte Carlo sampling algorithms. Relying on a conjugate mixture prior formulation that allows for closed-form Expectation-maximization updates, we propose here a framework to perform efficient joint variable selection and coefficient estimation in time-frequency dictionaries (such as Gabor frames) under problem-specific prior constraints—namely, structured *a priori* variable inclusion probabilities intended to reflect the typical time-frequency persistence patterns found in audio signals, and heavy-tailed priors on synthesis coefficients aimed at capturing the wide range of values expected in non-stationary processes. Applications to audio denoising, missing data interpolation and transient identification are presented.**

***Index Terms*—Expectation-maximization, Bayesian variable selection, Gabor frames, sparse regression, denoising, missing data**

I. Introduction

Sparse coding is a well-established paradigm in the field of signal processing and has been successfully applied to restoration tasks such as signal denoising [1], [2] or image inpainting [3], also underlying the compressive sensing reconstruction technique [4]. In these and other related signal processing applications, the target signal is typically modelled as a weighted sum of atoms from a fixed or learned dictionary, and some type of sparsity-inducing constraint is imposed on the synthesis coefficients. Conventional tools to encode the sparsity assumption include the ℓ_1-norm and related regularization schemes [5], [6], heavy-tailed distributions [7], [8] (some of which are equivalent to certain classic forms of regularization), spike-and-slab priors [9], [10], [11], or a combination of them (e.g. [12]).

Intuitively, the sparsity-promoting property of heavy-tailed priors can be understood as a consequence of their very function shape: such distributions are markedly peaked around zero and have wider tails than a Gaussian distribution; thus, many regression coefficients are shrunk to zero in the *a posteriori* estimate, whereas a significantly smaller set of larger coefficients is employed to approximate the target signal. Spike-and-slab priors constitute a different approach to variable selection in which inclusion is explicitly modelled by an activation coefficient—a variable indicating whether the respective regressor is included to the model's support set or not. Conditional on these activation variables, the prior of each regression coefficient takes either the form of a *spike*–a distribution strongly concentrated around zero, typically a Dirac delta or a zero-centered Gaussian with small variance–or a *slab*–a diffuse distribution covering the range of values the selected regression coefficients are likely to take. Activation variables, in turn, are assigned their own prior. If no further knowledge is available about the dependencies that may exist between model coefficients, the standard choice is an independent Bernoulli prior (e.g., the one adopted in seminal works like [10]), whose parameter roughly controls the level of sparsity of the model.

An appealing aspect of the hierarchical nature of spike-and-slab models is their suitability for structured sparsity. The standard Bernoulli prior for activation coefficients can be readily replaced by Markov chain, Ising or other discrete priors in order to incorporate structural information in the model, without necessarily sacrificing tractability. Relevant examples include [13], where a Bayesian variable selection model endowed with an Ising prior is employed to reflect the known structure of genomic data, or [12], where Markov chain and Ising priors are adopted to induce time-frequency persistence over a grid of audio synthesis coefficients. Although different in their mechanisms to carry out variable selection, spike-and-slab priors and heavy-tailed distributions are not mutually exclusive: in the study of non-stationary processes may such as audio time series, prior structural knowledge of the signal can be modelled with (structured) spike-and-slab priors, and at the same time a heavy-tailed prior on regression coefficients can be kept in order to capture the wide range of values expected in non-stationary processes (rather than just for the sake of sparsity). Works like [12], [14], for example, benefit from this hybrid approach in the context of audio noise reduction. Typically, computation of spike-and-slab models relies on Markov chain Monte Carlo strategies—that is the case of the aforementioned examples and pioneering works like [10]. While Markov chain Monte Carlo (MCMC) methods offer remarkable benefits in the context of statistical inference (e.g., principled computation of otherwise intractable models, complete characterization of posterior distributions), they also tend to be more computationally demanding than optimization-based alternatives, when available—hence many efforts to boost the efficiency of MCMC methods have been

Rubén M. Clavería and Simon J. Godsill are with the Department of Engineering, University of Cambridge, CB2 1PZ, UK (e-mail:rmc83@cam.ac.uk; sjg@eng.cam.ac.uk).

made in recent years, including works specifically aimed at Bayesian variable selection [15], [16].

In this work, the problem of variable selection and coefficient estimation in time-frequency dictionaries is addressed with an Expectation-Maximization (EM) approach. Unlike LASSO and its variations, Bayesian variable selection through spike-and-slab priors is not easily addressed with fast optimization methods, and although attempts to tackle the problem through EM have been proposed (e.g., [17]), their applicability is limited due to the unavailability of a closed-form solution for the E-step. In the present work, this limitation is circumvented by setting a small-variance Gaussian spike distribution instead of a Dirac delta, which leads to closed-form EM updates. Related Gaussian mixture formulations have been previously exploited in [18], where an EM algorithm for wavelet-based signal processing is proposed, and [19], where a general framework for EM-based variable selection is introduced. Although a Gaussian mixture spike-and-slab prior is a modification of the ideal case where irrelevant coefficients are shrunken to the exact value of zero, it provides significant gains in efficiency with little compromise in the quality of the results, and in fact it is possible to argue that no basis coefficient should ever be precisely zero when modelling real data.

This work leverages the hierarchical nature of spike-and-slab priors to promote the time-frequency persistence patterns observed in real audio signals through Markov chain priors on the activation coefficients, for which exact E-step updates are derived and efficiently computed with the Forward-backward algorithm [20]. Additionally, a rule to estimate Markov chain parameters (transition probabilities) in the M-step is presented. A key novel aspect of the algorithms proposed in this work enhances computational performance by solving the relevant linear systems (see Section III-B) with the Fast Iterative Shrinkage/Thresholding Algorithm [21], [22], which is facilitated by the block-sparse structure of the time-frequency dictionaries used. In a manner akin to [12], [14], the models proposed here have both structured spike-and-slab priors for activation variables and independent heavy-tailed priors for regression coefficients —traits that are well suited to modelling of audio signals and which can easily be implemented using the proposed framework. Applications to audio denoising, missing data interpolation and transient identification are presented.

This article is organized as follows: Section II describes the specific generative model adopted. Section III presents the proposed Expectation-maximization procedure, the variable selection scheme and an extension to multi-resolution dictionary representations. Section IV presents denoising, transient estimation and missing data interpolation results with the proposed algorithm. Finally, conclusions and prospective extensions to this work are discussed Section V.

II. A PROBABILISTIC MODEL FOR AUDIO SIGNALS

In a Bayesian setting, the mechanism through which a signal $\mathbf{y}$ is produced is described in terms of a probabilistic model, which must specify both how $\mathbf{y}$ is generated from a set of latent variables $\mathbf{x}$–the emission probability $p(\mathbf{y}|\mathbf{x})$– and a description of the latent variables—the prior $p(\mathbf{x})$. In turn, the dependency of both probabilities on a fixed set of parameters $\boldsymbol{\theta}$ can be stated explicitly: $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$, $p(\mathbf{x}|\boldsymbol{\theta})$. Aside from providing a theoretically sound approach to uncertainty characterization, this class of modelling benefits from the diverse and well-established apparatus of Bayesian inference: Expectation-maximization [23], different forms of MCMC [24], [25], variational approximation [26], etc. can be employed to calculate the relevant integrals, estimates or statistics. Depending on the specific objectives of a problem, practitioners can choose, adapt and combine these well tested approaches.

The main objective of audio restoration is generating clean, realistic sounding versions of a degraded input signal. Unlike fields such genomic analysis, where identifying the relevant covariates associated to a genetic trait or the multiple peaks of the posterior is of critical importance, in audio restoration the complete characterization of the posterior distribution is typically unnecessary and most techniques focus primarily on obtaining reasonable point estimates. In that context, MCMC strategies can be leveraged to explore the feature space and may have better convergence properties then their EM- or gradient-based counterparts [14], but the desired output is still a point-estimate. Furthermore, given the focus on the perceptual quality of the output, sub-optimal solutions are acceptable as long as they generate high quality sound reconstructions.

On these grounds, and bearing in mind the potential that computationally efficient strategies may have in online settings, we devise a fast EM-based algorithm for signal restoration, extending the framework presented in [19]. Owing to the arguments presented before, the usual drawbacks of EM –namely, not providing a complete characterization of the posterior distribution nor guaranteeing convergence to the global optimum– are overlooked. Nevertheless, results in both section IV and [19] support the potential of Expectation-Maximization Variable Selection for practical tasks.

A. Model description

Here we provide full details of the proposed audio time-frequency synthesis regression, noting that the framework of the earlier Expectation-Maximization Variable Selection (EMVS) method [19] is significantly simpler, being essentially the well-known Bayesian Variable Selection model for linear regression of [11] [1], but here a much more efficient EM alternative to the Gibbs sampler-based Stochastic Search Variable Selection of [10], [11] is proposed.

The model used in this work is based upon the *Gabor linear regression* framework of [12], where audio signals are represented as a weighted sum of Gabor atoms,

$$x(t) = \sum_{m=0}^{\mathrm{M}-1} \sum_{n=0}^{\mathrm{N}-1} \tilde{c}_{m,n} g_{m,n}(t) \tag{1}$$

$$g_{m,n}(t) = w_n(t) \exp(-2\pi \mathrm{i} \frac{m}{\mathrm{M}} t). \tag{2}$$

[1] Which in in turn can be regarded as an updated version of [10], endowed with a conjugate-prior on the regression coefficients that facilitate marginalization.

Function $w_n(t)$ in Eq. 2 denotes a window function centered at some time value t_n (window length W). In this work, centers t_n are placed so that a 50% overlap between adjacent frames occurs—a choice based on the Balian-Low theorem, which states that no orthonormal Gabor basis can achieve good localization properties in both time and frequency at the same time [27]. Since the resulting dictionary leads to an undetermined linear system, a high level of regularization is needed to avoid overfitting—hence the relevance of incorporating strategies for variable selection.

As our study is restricted to signals in the real domain, we adopt the real representation of Gabor atoms and coefficients derived in [28] (see also [12], Appendix A.1). Under such a representation, complex coefficients $\tilde{c}_{m,n}$ are replaced by a bivariate vector $\mathbf{c}_{m,n} = [c_{m,n,\mathrm{r}}, c_{m,n,\mathrm{i}}]^\top$, whose elements are the real and imaginary parts of the original complex coefficient. Accordingly, each Gabor atom is represented by the real and the imaginary parts of complex function $g_{m,n}(t)$. If the observed signal $y(t),\ t = 0, \ldots, T-1$ is modelled as a clean signal $x(t)$ affected by additive Gaussian noise of mean zero and variance σ^2 (for all time instants t), the resulting likelihood function can be expressed as

$$p(\boldsymbol{y}|\mathbf{c}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^{\mathrm{T}}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{G}\mathbf{c}\|^2\right) \quad (3)$$

where Eqs. (1, 2) are put in matrix form[2] for the sake of conciseness. Different of forms of sparsity–both structured and unstructured–can be promoted by setting spike-and-slab priors on coefficients $\mathbf{c}_{m,n}$. Given a set of activation coefficients $\gamma_{m,n}$ (referred to as *indicators* or *activation variables* subsequently), the prior density of each coefficient pair $\mathbf{c}_{m,n}$ is given by a bivariate uncorrelated Gaussian whose variance is $\sigma^2 v_0$ when $\gamma_{m,n} = 0$ and $\sigma^2 v_{m,n}$ when $\gamma_{m,n} = 0$ (Eq. 4). Scaling factors v_0 and $v_{m,n}$ control the relevance of each coefficient: a small quantity v_0 is assigned to the negligible coefficients, whereas larger values $v_{m,n}$ are assigned to variables that are 'selected' in the regression model. This form of Gaussian mixture prior, having non-zero variance v_0 for insignificant coefficients, is adopted in [19] and used in the present work for its mathematical tractability rather than necessarily for its physical modelling accuracy (although it may be argued that all coefficients should be assigned some small variance in real data modelling) —the derivations shown in Section III strongly rely on it, and are not possible in the case where the 'unselected' coefficients are set to be exactly 0. Note that $p(\mathbf{c}_{m,n}|\sigma^2, \gamma_{m,n}, v_{m,n})$ collapses to the case studied in [12], $(1-\gamma_{m,n})\delta_{\mathbf{0}_2}(\mathbf{c}_{m,n}) + \gamma_{m,n}\mathcal{N}(\mathbf{c}_{m,n}; \mathbf{0}_2, \sigma^2 v_{m,n}\mathbf{I}_2)$ when $v_0 \to 0$. Since coefficient pairs $\mathbf{c}_{m,n}$ are independent given an indicator vector $\boldsymbol{\gamma}$, the joint coefficient prior $p(\mathbf{c}|v_0, \mathbf{v}, \boldsymbol{\gamma}, \sigma^2)$ reduces to Eq. 5.

$$\begin{aligned} p(\mathbf{c}_{m,n}|\sigma^2, \gamma_{m,n}, v_{m,n}) = \ldots \\ \mathcal{N}\Big(\mathbf{c}_{m,n}; \mathbf{0}, \sigma^2(\gamma_{m,n}v_{m,n} + [1-\gamma_{m,n}]v_0)\boldsymbol{I}_2\Big) \end{aligned} \quad (4)$$

$$p(\mathbf{c}|v_0, \mathbf{v}, \boldsymbol{\gamma}, \sigma^2) = \prod_{m=0}^{\mathrm{M}/2} \prod_{n=0}^{\mathrm{N}-1} p(\mathbf{c}_{m,n}|\sigma^2, \gamma_{m,n}, v_{m,n}) \quad (5)$$

Following [12] and related examples of Gabor-based restoration [29], [30], dependencies between the different time-frequency elements are modelled at the *indicator* level through structured priors on activation vector $\boldsymbol{\gamma}$ rather than on coefficients $\mathbf{c}_{m,n}$ directly. Although different priors on $\boldsymbol{\gamma}$ can be deployed, as exemplified by the logistic regression prior or Markov random field case studies of [19], [12], Markov chain priors are emphasized in this work (Eq. 6) due to their suitability in representing time-frequency persistence patterns [31], [12], [14]. In turn, hyperparameters $\boldsymbol{\phi}$ (transition probabilities) are assigned their own priors (Eq. 7; definition of Beta distribution in Appendix A). Although each chain has an independent set of transition probabilities $\phi_{00,m}, \phi_{11,m}$ in our work here, a pair ϕ_{00}, ϕ_{11} that is common to all chains can also lead to good results (e.g., [14]). The priors for activation variables and transition probabilities are given as:

$$\begin{aligned} p(\boldsymbol{\gamma}_m|\phi_{00}, \phi_{11}) = p(\gamma_{m,0}; \phi_{00,m}, \phi_{11,m}) \cdot \ \ldots \\ \prod_{n=1}^{\mathrm{N}-1} p(\gamma_{m,n}|\gamma_{m,n-1}; \phi_{00,m}, \phi_{11,m}) \end{aligned} \quad (6)$$

$$p(\boldsymbol{\phi}) = \prod_{m=0}^{\mathrm{M}/2} \mathcal{B}(\phi_{00,m}; \alpha_{00}, \beta_{00})\mathcal{B}(\phi_{11,m}; \alpha_{11}, \beta_{11}) \quad (7)$$

Eq. (6) contains the primary model used in this work, where each frequency m is assigned an independent Markov chain along frame index n as a means of capturing the persistence of tonal components along the time axis. Chains are assumed to be in equilibrium, so that $p(\gamma_{m,0}; \phi_{00}, \phi_{11})$ is the stationary probability of the Markov chain[3]. A similar prior applied across frequency rather than time is proposed in Section III-D to model transient components. Unstructured sparsity is also possible, e.g., by setting independent Bernoulli priors on each activation variable: $p(\boldsymbol{\gamma}) = \phi^{|\gamma|}(1-\phi)^{P-|\gamma|}$, $p(\phi) = \mathcal{B}(\phi; \alpha_\phi, \beta_\phi)$, but leads to lower-quality results for the kind of signals studied in this work.

The dependency of $\mathbf{c}$ on σ^2 (Eq. 4) is known as the *conjugate formulation*–in contrast to the non-conjugate case where an independent variance σ_k^2 is assigned to each coefficient c_k, e.g., in [10], [12]–and leads to analytical simplification through marginalization, since $\mathbf{c}$ and σ^2 can be eliminated from the analysis and a closed-form expression for $p(\boldsymbol{\gamma}|\mathbf{y})$ can be used instead. That trait is conveniently exploited in the MCMC-based algorithms for varable selection such as the Stochastic Search Variable Selection of [11] or Tempered Gibbs Sampler of [16]. In EMVS, though, the main advantage of such a formulation is that it decouples the M-steps of $\mathbf{c}$ (Eq. 21) and σ^2 (Eq. 22).

In their simplest form, frameworks for Bayesian variable selection consider a common variance for all the *activated* variables (e.g., [10]). Setting an independent random variance

[2]Matrix $\mathbf{G}$ is simply the Gabor atoms expressed as column vectors and arranged in an orderly way. Similarly $\mathbf{y}$ and $\mathbf{c}$ correspond to the signal $y(t)$ and coefficients $\mathbf{c}_{m,n}$ stacked in columns.

[3]Concretely, $p(\gamma_{m,0}; \phi_{00}, \phi_{11}) = \frac{(1-\phi_{00,m})^{\gamma_{m,1}}(1-\phi_{11,m})^{1-\gamma_{m,1}}}{2-\phi_{00,m}-\phi_{11,m}}$. For chains that are not in equilibrium, initial probabilities π_0, $\pi_1 = 1-\pi_0$ must be defined explicitly.

for each coefficient does not however lead to any major compromise in tractability. Moreover, independent heavy-tailed priors on regression coefficients can be easily incorporated by this means into the model using the Scale mixtures of normals (SMiN) approach [32], which allows a wide range of densities (including heavy-tailed distributions of interest like Laplace or the Student's t) as an integral $p(x) = \int \mathcal{N}(x; 0, \mathrm{s}) p(\mathrm{s})\mathrm{ds}$ where the form of $p(x)$ is controlled by through the specific distribution $p(\mathrm{s})$ assigned to variance s. In this work, the conditionally Gaussian structure of SMiN is employed to induce a Student's t prior on the activated regression coefficients by setting an inverse-Gamma prior (definition in Appendix A) on scaling factors $v_{m,n}$: $p(v_{m,n}) = \mathcal{IG}\left(v_{m,n}; \kappa, \eta\right)$. The rate parameter η can be made frequency-dependent to reflect the decaying profile of the power spectrum of typical audio signals along the frequency axis:

$$\eta_m = \frac{\eta_0}{\sqrt{1 + \left[\frac{2\pi m/M}{\omega_0}\right]^p}}, \tag{8}$$

where a Butterworth-like gain function with parameters p (order) and ω_0 (cut-off frequency) has been used. The rationale behind using independent Student's t priors on active regression coefficients is that heavy-tailed priors can capture the broad range of values of non-stationary processes in a way that purely Gaussian priors cannot. Note that $v_{m,n}$ is defined for all (m, n), even if they are not used in Eq. 4 when $\gamma_{m,n} = 0$. The dependency of η on all coefficients $c_{m,n}$ complicates its online estimation through EM; however, manual tuning is possible since a fairly wide range of values lead to good results in practice. Moreover, since the model described is also amenable to analysis using MCMC schemes, appropriate values for η can be calculated from the data in a way akin to [12] (where η is assigned its own prior and estimated online within the sampling-based algorithm) and then used in the EM routine.

Finally, as in many Bayesian variable selection frameworks, an inverse-Gamma prior is set for observation noise variance σ^2: $p(\sigma^2) = \mathcal{IG}(\sigma^2; \frac{\nu}{2}, \frac{\nu\lambda}{2})$ due to its conjugacy properties. Thus, the joint density for all parameters can be written as:

$$\begin{aligned} p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi, \mathbf{y}) = p(\mathbf{y}|\mathbf{c}, \sigma^2)\ p(\mathbf{c}|v_0, \mathbf{v}, \boldsymbol{\gamma}, \sigma^2) \\ p(\boldsymbol{\gamma}|\boldsymbol{\phi})\ p(\mathbf{v})\ p(\boldsymbol{\phi})\ p(\sigma^2). \end{aligned} \tag{9}$$

III. Expectation-Maximization for Variable Selection

Rather than calculating the maximizer of Eq. (9) directly (given a fixed observation vector $\mathbf{y}$)[4], EMVS treats the indicator vector $\boldsymbol{\gamma}$ as a latent variable and calculates the MAP estimate of the other model parameters by maximizing the log-posterior of the marginal $\mathcal{L}_{\mathbf{y}}(\mathbf{c}, \mathbf{v}, \sigma^2, \phi) = \log \int p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi|\mathbf{y})\mathrm{d}\boldsymbol{\gamma}$ through an Expectation-maximization scheme. Activation coefficients could, for example, then be recovered *ex post* by solving:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathbb{P}(\boldsymbol{\gamma}|\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}}), \tag{10}$$

[4]Since $p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi|\mathbf{y}) = p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi, \mathbf{y})/p(\mathbf{y})$, maximizing the likelihood $p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi|\mathbf{y})$ or the joint probability $p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \phi, \mathbf{y})$ is equivalent if $\mathbf{y}$ is fixed.

where $\boldsymbol{\theta}$ denotes all the model parameters excluding the regression coefficients—σ^2, $\boldsymbol{\phi}$ and $\mathbf{v}$ in the case of the audio signal model considered here, but this can vary depending on the specific problem. The estimate $\hat{\boldsymbol{\gamma}}$ of Eq. (10) is not equivalent to that obtained with the posterior inclusion probability $p(\boldsymbol{\gamma}|\mathbf{y})$ (e.g., [11]), which may have different implications depending on the application.

In EM, $p(\mathbf{c}, \mathbf{v}, \sigma^2, \phi|\mathbf{y})$ is indirectly optimized by maximizing the auxiliary function $Q(\mathbf{c}, \mathbf{v}, \sigma^2, \boldsymbol{\phi}|\cdot)$ (defined in Eq. (11)) iteratively through the E and M steps.

$$\begin{aligned} &Q(\mathbf{c}, \mathbf{v}, \sigma^2, \boldsymbol{\phi}|\hat{\mathbf{c}}, \hat{\mathbf{v}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}) \\ &\quad = \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\log p(\mathbf{c}, \boldsymbol{\gamma}, \mathbf{v}, \sigma^2, \boldsymbol{\phi}|\mathbf{y})|\hat{\mathbf{c}}, \hat{\mathbf{v}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}\right] \\ &\quad = Q_1(\mathbf{c}, \sigma^2, \mathbf{v}|\cdot) + Q_2(\phi_{00,m}, \phi_{11,m}|\cdot) + \mathrm{C}, \end{aligned} \tag{11}$$

The symbol $\cdot$, used for conciseness', denotes the set of estimates $\hat{\mathbf{c}}, \hat{\mathbf{v}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}$ on which the latent variable $\boldsymbol{\gamma}$ is conditioned. In order to facilitate algebraic manipulation, $Q(\mathbf{c}, \mathbf{v}, \sigma^2, \boldsymbol{\phi}|\cdot)$ is separated into terms $Q_1(\mathbf{c}, \mathbf{v}, \sigma^2|\cdot)$ (Eq. (12), with $\mathrm{P} = 2(\mathrm{M}/2 + 1)\mathrm{N}$) and $Q_2(\boldsymbol{\phi}|\cdot) = \sum_m Q_{2,m}(\phi_{00,m}, \phi_{11,m}|\cdot)$ (Eq. (13)),

$$\begin{aligned} &Q_1(\mathbf{c}, \sigma^2, \mathbf{v}, \boldsymbol{\phi}|\mathbf{c}^{(k)}, [\sigma^2]^{(k)}, \mathbf{v}^{(k)}, \phi^{(k)}) \\ &= \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left[\log p(\mathbf{c},\ \sigma^2, \mathbf{v}, \phi|\boldsymbol{y})|\mathbf{c}^{(k)}, [\sigma^2]^{(k)}, \mathbf{v}^{(k)}, \phi^{(k)}, \mathbf{y})\right] \\ &= -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{G}\mathbf{c}\|^2 - \frac{\nu\lambda}{2\sigma^2} - \left[\frac{\mathrm{T} + \mathrm{P} + \nu}{2} + 1\right]\log\sigma^2 \\ &\quad - \frac{1}{2\sigma^2}\sum_{m,n}\|\mathbf{c}_{m,n}\|^2 \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left(\frac{1}{v_{m,n}\gamma_{m,n} + (1 - \gamma_{m,n})v_0}\Big|\cdot\right) \\ &\quad - \sum_{m,n}\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left(\gamma_{m,n}\log v_{m,n} + (1 - \gamma_{m,n})\log v_0\Big|\cdot\right) \\ &\quad - (\kappa + 1)\sum_{m,n}\log v_{m,n} - \sum_{m,n}\frac{1}{v_{m,n}}\eta_m \end{aligned} \tag{12}$$

Defining $\boldsymbol{\gamma}_m = \{\gamma_{m,n}\}_{n=0,\ldots,\mathrm{N}-1}$ and using $\#00$ to denote the "transition count" (i.e. the count of how many times the chain goes from state 0 to state 0; and similarly for $\#01$, $\#11$, $\#10$), Q_2 can be written as follows:

$$\begin{aligned} &Q_{2,m}(\phi_{00,m}, \phi_{11,m}|\mathbf{c}^{(k)}, [\sigma^2]^{(k)}, \mathbf{v}^{(k)}, \phi^{(k)}) \\ &\quad = \ [\alpha_{\phi_{11,m}} + \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#11) - 1]\log\phi_{11,m} \\ &\quad + [\beta_{\phi_{11,m}} + \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#10) - 1]\log(1 - \phi_{11,m}) \\ &\quad + [\alpha_{\phi_{00,m}} + \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#00) - 1]\log\phi_{00,m} \\ &\quad + [\beta_{\phi_{00,m}} + \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#01) - 1]\log(1 - \phi_{00,m}) \\ &\quad + \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\gamma_{m,0}) \cdot \log(1 - \phi_{00,m}) \\ &\quad + [1 - \mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\gamma_{m,0})] \cdot \log(1 - \phi_{11,m}) \\ &\quad - \log(2 - \phi_{00,m} - \phi_{11,m}). \end{aligned} \tag{13}$$

A. E-step

Since most terms in Eqs. (12), (13) do not depend on the latent variable $\boldsymbol{\gamma}$ directly, the main challenge of the E-step lies in the calculation of the expressions involving operator $\mathbb{E}_{\boldsymbol{\gamma}|\cdot}$ within Q_1 and Q_2. Regardless of the underlying probability structure of indicators $\gamma_{m,n}$ (i.e. unstructured, Markov chain, etc.), the required expectations can be obtained as follows:

$$
\begin{aligned}
p^*_{m,n} &= \mathbb{E}_{\boldsymbol{\gamma}|\cdot}(\gamma_{m,n}|\mathbf{c}^{(k)}, \sigma^{2(k)}, \mathbf{v}^{(k)}, \boldsymbol{\phi}^{(k)}) \\
&= \sum_{\substack{\gamma_{m,\ell}\in\{0,1\} \\ \ell=0,\ldots,\mathrm{N}-1}} \gamma_{m,n} p(\boldsymbol{\gamma}_{m,0:\mathrm{N}-1}|\mathbf{c}^{(k)}, \sigma^{2(k)}, \mathbf{v}^{(k)}, \boldsymbol{\phi}^{(k)}) \\
&= \sum_{\substack{\gamma_{m,\ell}\in\{0,1\} \\ \ell\neq n}} 1 \cdot p(\boldsymbol{\gamma}_{m,0:\mathrm{N}-1}|\mathbf{c}^{(k)}, \sigma^{2(k)}, \mathbf{v}^{(k)}, \boldsymbol{\phi}^{(k)}) \\
&= \mathbb{P}(\gamma_{m,n} = 1|\mathbf{c}^{(k)}, \sigma^{2(k)}, \mathbf{v}^{(k)}, \boldsymbol{\phi}^{(k)}),
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
d^*_{m,n} &= \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left(\frac{1}{v_{m,n}\gamma_{m,n} + (1-\gamma_{m,n})v_0}\Big|\cdot\right) \\
&= \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left(\frac{\gamma_{m,n}}{v_{m,n}} + \frac{1-\gamma_{m,n}}{v_0}\Big|\cdot\right) \quad (\ldots \text{for } \gamma_{m,n}\in\{0,1\}) \\
&= \frac{p^*_{m,n}}{v_{m,n}} + \frac{1-p^*_{m,n}}{v_0},
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\gamma}|\cdot}\left(\gamma_{m,n}\log v_{m,n} + (1-\gamma_{m,n})\log v_0\Big|\cdot\right) \\
&\quad = p^*_{m,n}\log v_{m,n} + (1-p^*_{m,n})\log v_0,
\end{aligned} \tag{16}
$$

where terms $p^*_{m,n}$ and $d^*_{m,n}$ are introduced to ease notation. However, calculating of the probability of $\gamma_{m,n} = 1$ may involve more or less complexity depending on the kind of spike-and-slab prior adopted. In the simple case where independent Bernoulli priors are adopted for $\gamma_{m,n}$, this probability can be easily computed using that $\mathbb{P}(\gamma_{m,n} = 1|\mathbf{c}^{(k)}, [\sigma^2]^{(k)}, \mathbf{v}^{(k)}, \boldsymbol{\phi}^{(k)}) \propto \mathbb{P}(\gamma_{m,n} = 1|\phi^{(k)})\ p(c^{(k)}_{m,n}|\gamma_{m,n} = 1, v^{(k)}_{m,n}, [\sigma^2]^{(k)})$, where $\mathbb{P}(\gamma_{m,n} = 1|\phi^{(k)}) = \phi^{(k)}$. For the prior used in this work, the hidden Markov model structure of the posterior probability (Eq. 17) can be exploited to calculate the exact solution of $\mathbb{P}(\gamma_{m,n} = 1|\cdot)$. Denoting $\mathrm{b}^0_{m,n} = \mathcal{N}(\boldsymbol{c}^{(k)}_{m,n}|\mathbf{0}, [\sigma^2]^{(k)}v_0\boldsymbol{I}_2)$, $\mathrm{b}^1_{m,n} = \mathcal{N}(\boldsymbol{c}^{(k)}_{m,n}|\mathbf{0}, [\sigma^2]^{(k)}v^{(k)}_{m,n}\boldsymbol{I}_2)$, it is clear that the posterior probability used in $\mathbb{E}_{\boldsymbol{\gamma}|\cdot}$ satisfies:

$$
\begin{aligned}
&p(\boldsymbol{\gamma}_m|\mathbf{c}^{(k)}, \mathbf{v}^{(k)}, [\sigma^2]^{(k)}, \phi^{(k)}_{00,m}, \phi^{(k)}_{11,m}) \propto \\
&\quad \left(p(\gamma_{m,0}|\phi^{(k)}_{00,m}, \phi^{(k)}_{11,m})\mathrm{b}^{\gamma_{m,0}}_{m,0}\right) \cdot \prod_{n=1}^{\mathrm{N}-1}\left(\phi^{(k)}_{\gamma_{m,n}\gamma_{m,n-1},m}\mathrm{b}^{\gamma_{m,n}}_{m,n}\right),
\end{aligned} \tag{17}
$$

where $\mathrm{b}^{\gamma_{m,n}}_{m,n}$ can be regarded as the emission or observation probability of the states of the hidden Markov chain $p(\boldsymbol{\gamma}_m|\phi^{(k)}_{00,m}, \phi^{(k)}_{11,m})$. The Forward-backward algorithm (see [20] for a review) provides a means to calculate both the single marginals (i.e., the probability of the chain of being in a specific state at time step n) and the joint marginals (the probability of a certain sequence of states starting at time step n).

Since there is no interaction between chains, terms $p^*_{m,n}$ and $\mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#s_1s_2),\ s_1, s_2 \in \{0,1\}$ are obtained by running Algorithm 1 for each frequency index m separately, and are related to the Feedback-Backward Algorithm output (α_n, β_n) through the following equations (the dependency on index m is omitted for ease of notation):

$$
p^*_n = \frac{\alpha^1_n\beta^1_n}{\alpha^0_n\beta^0_n + \alpha^1_n\beta^1_n}, \tag{18}
$$

Algorithm 1: Forward-backward algorithm (2 states)

Input: $\pi_1, \phi_{00}, \phi_{11}, \{(\mathrm{b}^0_n, \mathrm{b}^1_n)\}_{n=0,\ldots,\mathrm{N}-1}$
Output: $\{(\alpha_n, \beta_n)\}_{n=0,\ldots,\mathrm{N}-1}$

Forward algorithm:
$\alpha^0_0 = (1-\pi_1)\mathrm{b}^0_0$; $\alpha^1_0 = \pi_1\mathrm{b}^1_0$;
for $n = 1, \ldots, N-1$ **do**
  $\alpha^0_n = \alpha^0_{n-1}\phi_{00}\mathrm{b}^0_n + \alpha^1_{n-1}\phi_{10}\mathrm{b}^0_n$;
  $\alpha^1_n = \alpha^0_{n-1}\phi_{01}\mathrm{b}^1_n + \alpha^1_{n-1}\phi_{11}\mathrm{b}^1_n$;
end
Backward algorithm:
$\beta^0_{\mathrm{N}-1} = 1$; $\beta^1_{\mathrm{N}-1} = 1$;
for $n = N-2, \ldots, 0$ **do**
  $\beta^0_n = \beta^0_{n+1}\phi_{00}\mathrm{b}^0_{n+1} + \beta^1_{n+1}\phi_{01}\mathrm{b}^1_{n+1}$;
  $\beta^1_n = \beta^0_{n+1}\phi_{10}\mathrm{b}^0_{n+1} + \beta^1_{n+1}\phi_{11}\mathrm{b}^1_{n+1}$;
end

Defining $q_{s_ns_{n+1}} = \alpha^{s_n}_n\phi_{s_ns_{n+1}}\mathrm{b}^{s_{n+1}}_{n+1}\beta^{s_{n+1}}_{n+1}$, the probability of a transition s_1s_2 at frame index n is given by:

$$
\begin{aligned}
q^*_{s_ns_{n+1}} &= \mathbb{P}(\gamma_n = s_n, \gamma_{n+1} = s_{n+1}|\mathbf{c}, \mathbf{v}, \sigma^2, \phi_{00}, \phi_{11}) \\
&= \frac{q_{s_ns_{n+1}}}{\sum_{t_n}\sum_{t_{n+1}} q_{t_nt_{n+1}}},
\end{aligned} \tag{19}
$$

from which the expected "transition count" is obtained as follows:

$$
\mathbb{E}_{\boldsymbol{\gamma}_m|\cdot}(\#s_ns_{n+1}) = \sum_{n=0}^{\mathrm{N}-2} q^*_{s_ns_{n+1}} \tag{20}
$$

B. M-step

As noted in [19], maximization of parameters (just $\mathbf{c}$, σ^2 and $\boldsymbol{\phi}$ in the simpler model of [19]) is facilitated by the separability of $Q(\cdot)$. In the extended model presented here, which incorporates parameters $v_{m,n}$, that claim does not hold anymore, since $v^{(k+1)}_{m,n}$ depends on $c^{(k+1)}_{m,n}$. It must be noted, though, that if the algorithm proceeds by sequentially maximizing $\mathbf{c}$, σ^2, $\boldsymbol{\phi}$ and $\boldsymbol{v}$ conditional on the updated values of the rest of the parameters, the resulting sequence of $Q^{(k)}$, $k = 1, \ldots, \mathrm{K}_{\max}$ is still monotone, guaranteeing convergence. By virtue of elementary multivariate Gaussian distribution results, the update of regression coefficients obtained from Eq. (12) is given by:

$$
\begin{aligned}
\mathbf{c}^{(k+1)} &= \arg\min_{\mathbf{c}} \frac{1}{2}\|\mathbf{y} - \mathbf{G}\mathbf{c}\|^2 + \frac{1}{2}\sum_{m,n} d^*_{m,n}\|\mathbf{c}_{m,n}\|^2 \\
&= (\mathbf{G}^\top\mathbf{G} + \mathbf{D}^*)^{-1}\mathbf{G}^\top\mathbf{y},
\end{aligned} \tag{21}
$$

where D^* is the diagonal matrix containing terms $d^*_{m,n}$ (Eq. (15)) in the appropriate order. Despite of its size–just one second of audio sampled at 44,110 Hz involves the calculation of 88,200 coefficients–, the linear system posed by Eq. 21 is efficiently solved through the Fast Iterative Shrinkage/Thresholding Algorithm of [21], enhanced with the acceleration introduced in [33]. Initially proposed for models consisting of a data-fidelity term (in the manner of $\frac{1}{2}\|\mathbf{y} - \mathbf{G}\mathbf{c}\|^2$) plus a non-smooth regularization (e.g., ℓ_1-norm or total variation), FISTA has proved useful to solve certain forms of ridge regression as well. In particular, the scheme

Algorithm 2: FISTA for linear systems

Input: $\mathbf{y}, \mathbf{G}, \mathbf{c}^{(0)} = \mathbf{z}^{(0)} = \mathbf{x}_{\text{init}}, \mathrm{d} = 4, L, \mathbf{D}^*$
Output: $\hat{\mathbf{c}}$.
1 **for** $k = 0, \ldots, K_{\max} - 1$ **do**
2 $\quad \mathbf{z}^* = \mathbf{z}^{(k)} + \frac{1}{L}\mathbf{G}^\top \left(\mathbf{y} - \mathbf{G}\mathbf{z}^{(k)}\right)$;
3 $\quad \mathbf{c}_{m,n}^{(k+1)} = \left(\mathbf{I} + \frac{1}{L}\mathbf{D}^*\right)^{-1} \mathbf{z}^*$;
4 $\quad \mathbf{z}^{(k+1)} = \mathbf{c}^{(k+1)} + \frac{k}{k+\mathrm{d}}(\mathbf{c}^{(k+1)} - \mathbf{c}^{(k)})$;
5 **end**
6 $\hat{\mathbf{c}} = \mathbf{c}^{(K_{\max})}$;

described in Algorithm 2 has been used previously in [34] for coefficient estimation in the context of a non-negative matrix factorization model for audio signals. Performance is further improved by computing *analysis* ($\mathbf{G}^\top\mathbf{x}$) and *synthesis* ($\mathbf{G}\mathbf{x}$) operations through of Fast Fourier Transform—something facilitated by the structure of the dictionary matrix $\mathbf{G}$ and not possible in more general variable selection models. To ensure convergence, constant L in Algorithm 2 must be greater or equal than the Lipschitz constant of gradient $\nabla\frac{1}{2}(\mathbf{y} - \mathbf{G}\mathbf{z})$, which equals the largest eigenvalue of $\mathbf{G}^\top\mathbf{G}$. Note that the shrinkage step of Algorithm 2 (code line 3) involves the inversion of a diagonal matrix, i.e., it is just a element-wise product between the inverses of the elements in the diagonal and vector $\mathbf{z}^*$.

The update rule for $[\sigma^2]^{(k+1)}$ is obtained by grouping the terms containing σ^2 in Q_1 and identifying the formula of $\log \mathcal{IG}(\sigma^2|\alpha', \beta')$. Then the maximizer is simply the mode of that distribution: $\frac{\beta'}{\alpha'+1}$.

$$[\sigma^2]^{(k+1)} = \frac{\|\mathbf{y} - \mathbf{G}\mathbf{c}^{(k+1)}\|^2 + \sum_{m,n} d_{m,n}^* \|\mathbf{c}_{m,n}^{(k+1)}\|^2 + \nu\lambda}{\mathrm{T} + \mathrm{P} + \nu + 2} \tag{22}$$

Following the same reasoning, grouping terms containing the scaling factors $v_{m,n}$ in Q_1 (given the updated values of $\mathbf{c}$ and σ^2) leads to the following update rule:

$$v_{m,n}^{(k+1)} = \left[\frac{p_{m,n}^* \|\mathbf{c}_{m,n}^{(k+1)}\|^2}{2[\sigma^2]^{(k+1)}} + \eta_m\right] \Big/ \left[\kappa + p_{m,n}^* + 1\right] \tag{23}$$

Finally, transition probabilities $\phi_{00,m}$, $\phi_{11,m}$ should be updated by solving $\arg\max_{\phi_{00,m},\phi_{11,m}} Q_{2,m}(\phi_{00,m}, \phi_{11,m}|\cdot)$. Rather than optimizing for $\phi_{00,m}$ and $\phi_{11,m}$ jointly, the procedure presented next is a sequential maximization scheme that leads to good results in practice. First, $\phi_{00,m}^{(k+1)}$ given $\phi_{11,m}^{(k)}$ is solved; then, the update $\phi_{11,m}^{(k+1)}$ given $\phi_{00,m}^{(k+1)}$ is obtained. For the first step, it suffices to calculate the following gradient:

$$\frac{\partial Q_{2,m}(\phi_{00,m}, \phi_{11,m}^{(k)})}{\partial \phi_{00,m}} = \frac{N_{00}}{\phi_{00,m}} - \frac{N_{01}}{1 - \phi_{00,m}} + \frac{1}{2 - \phi_{00,m} - \phi_{11,m}^{(k)}},$$

and make that quantity equal to zero. $N_{00} = \alpha_{\phi_{00,m}} + \mathbb{E}_{\boldsymbol{\gamma}|\cdot}(\#00) - 1$ and $N_{01} = \beta_{\phi_{00,m}} + \mathbb{E}_{\boldsymbol{\gamma}|\cdot}(\#01) + \mathbb{E}_{\boldsymbol{\gamma}|\cdot}\gamma_{m,1} - 1$ simply denote the terms accompanying $\log\phi_{00,m}$ and $\log(1 - \phi_{00,m})$, respectively, in Eq. 13. The equality $\frac{\partial Q_{2,m}}{\partial \phi_{00,m}} = 0$ leads to a quadratic polynomial, so that the update rule for $\phi_{00,m}^{(k+1)}$ is given by the solution of:

$$\phi_{00,m}^{(k+1)} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}{2a_2} \tag{24}$$

that lies in the interval $[0, 1]$. In Eq. 24, the auxiliary quantities $b = 2 - \phi_{11,m}^{(k)}$, $a_0 = N_{00} \cdot b$, $a_1 = 1 - N_{00}(1 + b) - N_{01}b$ and $a_2 = N_{00} + N_{01} - 1$ were used, a_0, a_1 and a_2 being polynomial coefficients of the respective quadratic problem. The update of transition probability $\phi_{11,m}^{(k+1)}$ is obtained using analogous arguments.

C. Variable selection

A direct estimate of the reconstructed signal may be obtained from the Expectation-maximization scheme as $\mathbf{G}\hat{\mathbf{c}}$. However, in experimentation it was observed this procedure did not produce the most successful denoising performance, perhaps as a result of all the small but non-zero coefficients estimated when indicator variables are zero. Instead we found it preferable to prune to zero spurious coefficients whose estimated activity is zero, where indicator $\hat{\boldsymbol{\gamma}}$ is obtained by solving Eq. (10) for $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\phi}}, \hat{\sigma}^2, \hat{\mathbf{v}})$, as now detailed through use of the Viterbi Algorithm.

Under the specific hierarchical models used in [19] – namely, Bernoulli and logistic regression priors, or a Markov Random Field prior approximated through independent mean field estimates of each γ_i– the solution of Eq. 10 is facilitated by the independence of activation coefficients: in order to maximize $\mathbb{P}(\boldsymbol{\gamma}|\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}})$, it suffices to set $\hat{\gamma}_i = 1$ whenever $\mathbb{P}(\gamma_i|\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}}) \geq 0.5$ (i denotes a generic index); thus, variable selection is reduced to a simple thresholding rule. Since the independence assumption does not hold for the Markov chain prior used in this work, each chain in Eq. 10 (which in turn describes a probability with the same formula as Eq. 17) is solved through the Viterbi algorithm [35], which finds the highest probability sequence of states using a scheme arguably similar to Algorithm 1. Thus, the restored signal is computed as:

$$\hat{\mathbf{x}} = \mathbf{G}(\hat{\boldsymbol{\gamma}}_{\text{Viterbi}} \odot \hat{\mathbf{c}}), \tag{25}$$

where $\odot$ denotes element-wise matrix product. In practice, replacing the Viterbi solution with the simple thresholding rule used in [19] (i.e., using the $\mathbb{P}(\gamma_{m,n} = 1|\hat{\mathbf{c}}, \hat{\sigma}^2, \hat{\mathbf{v}}, \hat{\boldsymbol{\phi}}) > 0.5$ criterion) also leads to good results. Alternatively, restoration can be carried out by recalcultating the synthesis coefficients using the limited dictionary $\mathbf{G}_{\hat{\boldsymbol{\gamma}}=1}$.

D. Multi-layer signal representation

The methodology proposed in previous sections is not limited to the Gabor signal representation of Eq. 1. Besides alternative choices for the basis (e.g., within this framework a Modified discrete cosine transform could serve a similar purpose as Gabor atoms), our basic formulation can be extended to multi-dictionary representations intended to capture different aspects of the signal of interest. In particular, a multi-resolution Gabor approach to simultaneously identify tonal and transient components in audio signals has been

proposed in [14], [36]. In lapped transform representations, typical window lengths of 1024 or 2048 samples (assuming a sampling frequency of 44,100 Hz) are successful in capturing time-persistent tonal components such as sustained piano notes but struggle to characterize short-lived events like percussive hits or note onsets–the transients–, which are typically absent in the reconstruction, resulting in a somewhat smoothed version of the original signal. Complementing the original long time resolution dictionary with a shorter time resolution basis not only fills the gaps left by missing transients but also facilitates the sharp separation of tonal and transient components—each of them is captured in a separate layer. Denoting the original, unobserved signal by $\mathbf{x}$, this extended model is summarized in Eq. 26, where $\mathbf{G}$ and $\mathbf{G}'$ are the long and short time resolution dictionaries, respectively, and $\mathbf{c}$, $\mathbf{c}'$ the corresponding coefficient vectors.

$$\mathbf{x} = \begin{bmatrix}\mathbf{G}\ \mathbf{G}'\end{bmatrix} \begin{bmatrix}\mathbf{c}\\ \mathbf{c}'\end{bmatrix} = \mathbf{G}\mathbf{c} + \mathbf{G}'\mathbf{c}' \tag{26}$$

This enhancement comes at the expense of a higher computational cost, since the number of model parameters roughly doubles in this setting. However, inference can be carried out with essentially the same procedure described in sections III-A, III-B, III-C—only a few minor modifications must be made in order to account for the new features. Specifically, the model of Eq. 26 must be plugged in the likelihood function (Eq. 3), and the newly introduced coefficients $\mathbf{c}'$ must be endowed with a spike-and-slab prior controlling their inclusion in the model (as in Eq. 4). In turn, an indicator vector $\boldsymbol{\gamma}'$ with its own structured prior (in the manner of Eq. 6) must be defined, as well as scaling factors $v'_{m',n'}$ controlling the variance of $\mathbf{c}'|\boldsymbol{\gamma}'$, with prior distributions $\mathcal{IG}(v'_{m',n'}; \kappa', \eta')$. Optionally, a distinct v'_0 can be set for coefficients $c_{m',n'}$ for which $\gamma'_{m',n'} = 0$.

Akin to tonal components, which exhibit a noticeable time persistence trait in real audio signals, transients show distinctive patterns in the time-frequency surface, tending to group in vertical lines—persistence along the frequency axis. Although the incorporation of a short time resolution dictionary can be enough to detect and estimate transients (e.g., [36] just uses unstructured priors for coefficients in each layer), inference can be improved by setting Markov chain priors on the indicator variable $\boldsymbol{\gamma}'$ along frequency index m'. Following [14], where a similar two-resolution model is used, prior $p(\boldsymbol{\gamma}'|\pi_1, \phi'_{00}, \phi'_{11})$ is defined such that all chains share the same transition probabilities ϕ'_{00}, ϕ'_{11}. The initial probability $\pi_1 = \mathbf{P}(\gamma'_{0,n'} = 1)$ ($n' = 0, \ldots, \mathrm{N}' - 1$) is given its own beta prior (vertical chains in the transient layer are not assumed to be in equilibrium as those in the tonal layer, so that the π_1 must be estimated explicitly). In the E-step, expected values involving $\boldsymbol{\gamma}$ remain the same, whereas those involving $\boldsymbol{\gamma}'$ can be readily calculated with the Forward-Backward algorithm (Algorithm 1, provided that it is fed with the appropriate inputs $\pi_1, \phi'_{00}, \phi'_{11}$, and considering chains in the frequency index m'). In the M-step, the update rule for $[\mathbf{c}; \mathbf{c}']$ can be easily re-derived (this involves adapting Algorithm 2 for the multi-resolution dictionary), and the same holds for σ^2. The update rule for $v'_{m',n'}$ is analogue to Eq. 23, whereas the updates for $\pi_1, \phi'_{00}, \phi'_{11}$, although not exactly equal to those for $\phi'_{00,m}, \phi_{11,m}$ (Eq. 24), can be obtained through the same reasoning. Formulae for updates in the multilayer case is included in Appendix B.

IV. Experimental results

In this section the denoising capabilities of the algorithms proposed in Section III are tested on a series of short single-instrument excerpts (Section IV-A) and on a longer, more complex piece of polyphonic music (Section IV-B). Additionally, the transient estimation property of the two-layer representation introduced in Section III-D is assessed. Section IV-C presents an extension of this framework to the problem of missing data interpolation. Audible results are available on https://github.com/matclaveria/EMVS_audio_restoration. Algorithms are implemented in MATLAB, and all experiments were run on an laptop with an Intel i7 2.3 GHz processor and 16 GB RAM.

A. Denoising: short single-instrument excerpts

In order to evaluate how the proposed algorithm performs depending on the specific traits that may be present in a musical piece, we carry out noise removal experiments on three different short samples: a fast piano excerpt, a glockenspiel recording and a string quartet extract. Input signals are generated by adding white Gaussian noise to the clean signal, and noise variances are chosen in such a way that target signal-to-noise ratios ($\mathrm{SNR_{in}}$) of 0, 10, 20 (dB) are obtained. Three different prior structures are tested for each case: **i**. the unstructured case where the prior inclusion probabilities of $\boldsymbol{\gamma}$ are modelled by independent Bernoulli distributions with a common parameter ϕ; **ii**. the one-layer "horizontal" Markov chain prior specified in Section II (Eq. 6); **iii**. the multi-resolution extended model described in Section III-D. All excerpts are sampled at 44.1 kHz. A Gabor dictionary with 50% frame overlap and window length 2048 is adopted for the tonal layer in all cases, whereas the transient layer of the multi-resolution case uses an overlapped Gabor dictionary with window length 256.

While the hyperparameters associated to σ^2 (ν, λ) and $v_{m,n}$ (κ, η) do not make much of a difference in the final results[5], the relevance of choosing appropriate hyperparameters for $\boldsymbol{\phi}$ (Eq. 7) must be emphasized: in order to obtain good reconstructions, models using Markov chain priors in either the time or the frequency axis must assume transition probabilities ϕ_{00} strongly concentrated in the upper feasible range (i.e., approaching 1); otherwise, the resulting reconstructions tend to exhibit a large presence of artifacts. Similarly, and as pointed out in [14], the prior of the initial probability π_1 of the transient layer must be strongly biased towards 0 to ensure reasonable results (hence $\alpha_\pi = 2$, $\beta_\pi = 5000$ are adopted in the Beta prior). Akin to [12], [14], it was found that setting a frequency-dependent rate parameter η_m for scaling factors $v_{m,n}$ leads to better solutions than flat frequency profiles. This feature is

[5]Through non-exhaustive trial-and-error tests it was found that $\nu = 3, \lambda = 0.05$ and $\kappa = \kappa' = 3, \eta = \eta' = 0.6$ were adequate for all the cases studied, and results remain unaltered by minor changes in these values.

incorporated by means of Eq. 8, setting values $\omega_0 = 2\pi \cdot 0.33$, $p = 6$ for the tonal layer and $p' = 3$ for the transient layer of case **iii** (since transients are expected to be rich in high frequency components, the decay is expected to be slower). A downside of EMVS is that the scaling factor v_0 (for the non-selected coefficients) must be tuned empirically; fortunately, for the setting described in this work the search interval can be limited to $v_0 \in [0.4 \cdot 10^{-3}, 1.5 \cdot 10^{-3}]$. In particular, experiments using the one-layer models (Bernoulli and and temporal Markov chain priors) were run with $v_0 = 10^{-3}$, whereas values $v_0 = 0.7 \cdot 10^{-3}$, $v_0' = 2.8 \cdot 10^{-3}$ were chosen for the the two-layer model. It was found, in practice, that setting $v_0' = \mathrm{f} \cdot v_0$, with $\mathrm{f} > 1$ a small factor leads to better results. All experiments are run for only 40 EM iterations. Results are summarized in Table I

1) Discussion of results: : Results in Table I suggest that the adoption of suitable structured priors and multilayer representations can lead to significant benefits in the quality of the output signal: in most cases, the one layer representation with Markov chain prior obtains higher $\mathrm{SNR}_{\mathrm{out}}$ values than its unstructured counterpart, whereas the two-layer representation widens that gap even further, scenarios of high degradation ($\mathrm{SNR}_{\mathrm{in}}$ = 0 dB) being the only exception (the gain in $\mathrm{SNR}_{\mathrm{out}}$ is either negligible or negative). From a qualitative standpoint, structured priors tend to generate clean, almost artifact-free reconstructions, even in cases where $\mathrm{SNR}_{\mathrm{out}}$ is lower than that obtained with an unstructured prior (e.g., glockenspiel excerpt, $\mathrm{SNR}_{\mathrm{in}}$ = 20 dB), making the former more pleasant to hear. Despite the higher $\mathrm{SNR}_{\mathrm{out}}$ values obtained when structured representations are used, results from the strings example suggest that the assumptions adopted in this work might be inadequate or too constraining for certain types of waveform: since bowed string sounds do not have distinguishable transient components[6], the transient layer acts more as an algorithmic burden (?) than as a helpful dictionary in this scenario. Accordingly, the reconstructed transients of the strings examples do not capture any meaningful property: they consist of a continuous "musical noise" which fits in smoothly in the full mix (possibly supplementing the waveform formed by the tonal layer) but do not sound good on their own. This suggests that the higher $\mathrm{SNR}_{\mathrm{out}}$ value simply comes more from a richer dictionary basis representation. Conversely, reconstructed transients of the piano and glockenspiel examples feature very distinct "clicks" aligned with the onset of musical notes (see Figure 1, and upper panel in Figure 2); and in those cases the increase in $\mathrm{SNR}_{\mathrm{out}}$ values can be directly traced to the incorporation of short-length transient components that the tonal-only representation fails to capture (Figure 2, lower panel, illustrates the "gap" filled in by the transient component).

The glockenspiel results in Table I are comparable with those obtained with the Gibbs sampler proposed in [14], both in terms of $\mathrm{SNR}_{\mathrm{out}}$ and subjective auditory criteria. The estimator used here (Eq. 25) is less prone to artifacts than the empirical average estimate of the coefficients (minimum mean squares estimator, MMSE) and achieves similar performance to the "MIX" estimate proposed in the aforementioned work (i.e., MMSE $\mathbf{c}$ values thresholded by the MAP estimate of the activation coefficients $\boldsymbol{\gamma}$). It must be noted that the scheme presented here is weaker in scenarios of low $\mathrm{SNR}_{\mathrm{in}}$: although the overall reconstructions sound arguably similar, the two layer model reported in Table I fails to deliver a sharp characterization of the transient when $\mathrm{SNR}_{\mathrm{in}}$ = 0 dB. This may be attributable to the fact the scheme presented here is less adaptive (since parameters such as v_0 and η were manually fixed in this experiment) and the use of a Gaussian spike prior for the "unselected" variables: since variable selection essentially depends on the magnitude of coefficients $c_{m,n}$ (Eq. 10), actual signal components with very low energy (such as transients) might be discarded. Fine-tuning of parameters η, v_0 for each different input (either manually or using a theoretically sound adaptive strategy) might improve these results.

2) Comparison with classic noise suppression methods: The rightmost column of Table I contains the output SNR of

[6]Instead, note onsets are marked by a gradual transition from silence to the tonal regime.

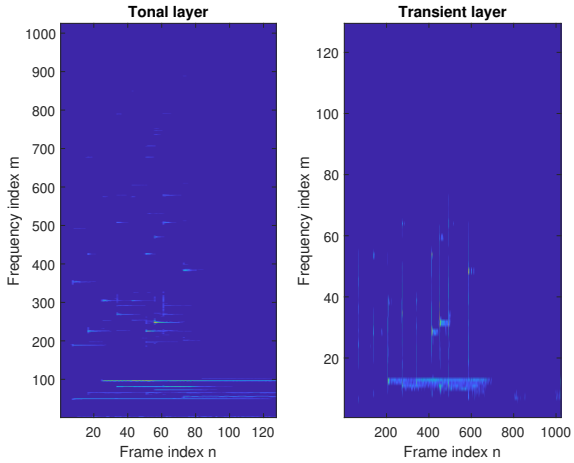

Fig. 1. Power spectra of tonal and transient layers, glockenspiel example, input $\mathrm{SNR}_{\mathrm{in}}$ = 10 dB. Tonal layer exhibits continuity along the time axis whereas the transient layer privileges vertical lines

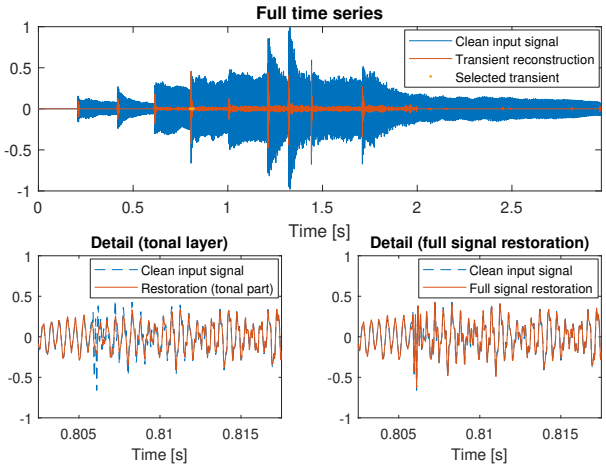

Fig. 2. Transient reconstruction of a degraded glockenspiel example, $\mathrm{SNR}_{\mathrm{in}}$ = 10 dB

.

TABLE I
RESULTS FOR SHORT SINGLE-INSTRUMENT PIECES

	SNR_{in} [dB]	One layer - unstructured		One layer - structured		Two layers - structured		Wiener
		SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]
Piano (∼1.5 s)	0	13.27	1.24	14.14	1.77	14.39	3.38	11.52
	10	19.89	1.49	20.57	1.76	21.32	4.38	19.97
	20	26.31	1.58	27.18	1.63	28.16	4.03	28.95
Glockenspiel (∼3 s)	0	14.74	3.51	15.80	3.99	15.48	10.43	11.80
	10	20.88	3.53	20.86	3.96	22.40	12.71	20.28
	20	27.94	3.52	27.46	3.99	29.53	13.78	28.35
Strings (∼4 s)	0	10.27	6.21	11.34	7.27	11.34	17.06	10.17
	10	17.37	6.26	18.00	7.44	18.43	20.18	18.17
	20	24.59	6.33	24.87	7.28	25.45	21.48	26.30

reconstructions using the Wiener subtraction rule described in [37], Chapter 6. Alternatives such as spectral subtraction and power subtraction were also tested, but results on music excerpts were inferior both in terms of output SNR and perceptual quality. All results were obtained assuming knowledge of the actual background noise level σ^2 (unlike the EM approach proposed in this work, the Wiener subtraction rule does not automatically estimate σ^2). As expected, the Wiener rule tends to be much more prone to audio artifacts (the only exception being the strings excerpt), even in high SNR_{in} regimes, despite obtaining comparable SNR_{out} values. In low SNR_{in} scenarios, the EM algorithm tends to obtain higher SNR_{out} with few or no artifacts; however, reconstructions tend to be sound more artificial, MIDI-like than their artifact-affected Wiener counterparts. In medium noise regimes (SNR_{in} = 10 [dB]) EM reconstructions exhibit significantly better perceptual quality and normally superior SNR_{out} values than Wiener reconstructions.

This comparison points out the limitations of SNR_{out} as a quality measure of signal reconstruction techniques. Furthermore, although classic spectral subtraction techniques such as [38], [39], [40] admit probabilistic interpretations (estimators derived through a range of different probabilistic criteria can be seen, for example, in [41]), the fact that the they are analysis-based rather than synthesis-based makes difficult to propose a full generative model of waveform $\mathbf{y}$ based on them. Thus, they do not provide a means to address tasks such as multiresolution decomposition (e.g., as in transient identification) or missing data interpolation in any obvious way. Conversely, the EM framework presented in this work can be easily adapted to tackle those tasks.

3) Convergence and efficiency: It is in efficiency where the major advantage of our EM scheme is found: whereas the Gibbs sampler strategy of [14] (where a similar model is computed) has a reported runtime of over an hour[7] for the same glockenspiel excerpt, the algorithm proposed here takes seconds to process the sequence, obtaining comparable results. In particular, runtimes of experiments adopting one layer representations are similar to the duration of the input sequence, suggesting potential for real-time applications (e.g., using this EM approach to process mini-batches of a streamed audio sequence). Unlike other EM-based approaches such as [42], the scheme proposed here is robust to random initialization: although the estimates obtained cannot be proven to be global optima, the algorithm consistently delivers reasonable signal reconstructions, regardless of the initial values of $\mathbf{v}$, σ^2, $\boldsymbol{\phi}$ ($\mathbf{c}$ is initialized to zero). The fast convergence of the algorithm can be observed in Figure 3, where the values of coefficients and scaling factors $v_{m,n}$ of a randomly chosen frame are plotted over the successive EM iterations. Since scaling factors $v_{m,n}$ are estimated explicitly rather than margined out, the proposed algorithm does not properly conform to the the Student's t-distribution prior assumption. Instead, it operates in a way reminiscent of the *automatic relevance determination* ideas underpinning [7], [8], carrying out variable selection through the adjustment of coefficient variances.

In the solutions obtained, $\hat{\sigma}^2$ does not deliver an estimate of the background noise but rather a mere measure of discrepancy between $\mathbf{y}$ and $\mathbf{Gc}$. Without variable selection (Eq. 25), product $\mathbf{G}\hat{\mathbf{c}}$ tends to follow the raw signal $\mathbf{y}$ rather than provide a clean version of it. Through the histograms of Figure 4 (upper panel), the accumulation pattern of each set of coefficients (selected according to the criteria described in Section III-C) can be observed. Whereas selected coefficients take values that may depart significantly from the peak around zero (in coherence with the heavy-tailed assumption incorporated in the model), unselected coefficients form a small variance Gaussian-like profile (again conforming to the modelling assumptions). The signal resulting from discarded synthesis coefficients, $\mathbf{G}[(1-\hat{\boldsymbol{\gamma}}) \odot \hat{\mathbf{c}}]$, contains essentially white noise. Figure 4, lower panel, plots the posterior inclusion probabilities $\mathbb{P}(\boldsymbol{\gamma}|\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}})$ of all coefficients[8], showing a sharp distinction between selected and discarded variables—probabilities tend to concentrate towards to the edges of the interval [0, 1].

B. Denoising: polyphonic music excerpt

The denoising algorithm is now tested on a long polyphonic excerpt featuring percussive sounds, plucked strings,

[7]It must be mentioned that the computational power of a regular laptop at the time of publication was inferior to what is now available. For the sake of comparison, we ran our implementation of the Gibbs-based method of [12] on the Glockenspiel sequence, SNR_{in} = 10 [dB], sampling frequency 44,100 [Hz], window size 2048 samples. We ran 2,000 iterations of the algorithm, obtaining a runtime of 57 [min], which is still dramatically longer than that of the EM algorithm proposed here. Code and result are available in the git repository of this work.

[8]Since the MAP trajectories (Viterbi algorithm) and the heuristic criterion $\mathbb{P}(\boldsymbol{\gamma}|\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}}) > 0.5$ obtain similar solutions in this particular context, the marginal posterior inclusion probabilities serve as a proxy of the actual variable selection.

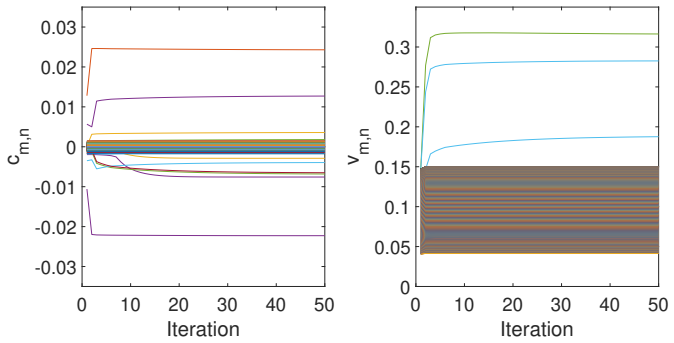

Fig. 3. Value of coefficients $c_{m,n}$ and scaling factors $v_{m,n}$ over iterations.

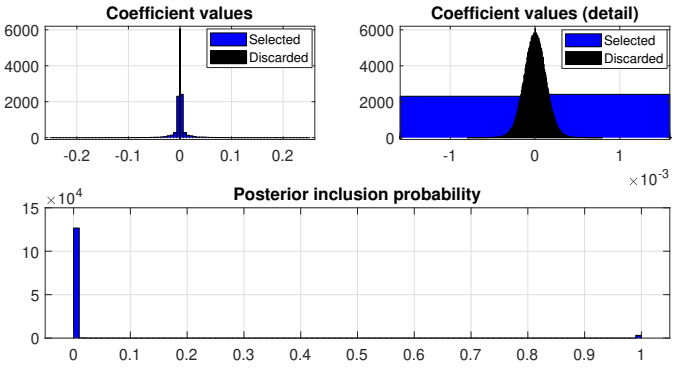

Fig. 4. Histograms of synthesis coefficients (upper panel); posterior inclusion probability (lower panel)

bass guitar and voice (∼24 second long, sampled at 44.1 kHz), contaminated with additive Gaussian white noise with $\sigma^2 = 0.03$ (yielding an SNR_{in} of 13 dB, approximately). Due to its rich variety of textures and overall complexity, this piece[9] has been previously used to test the robustness of other audio processing techniques (e.g., [14], [34]). The same models tested in Section IV-A are adopted (namely, the unstructured and structured one-layer models, and the two-layer model of Section III-D). Due to efficiency of the proposed approach, there is no need to separate the time series into shorter chunks to be processed independently (e.g., as done in [14]): the fast computation provided by Algorithm 2, coupled with the fast convergence of EMVS, facilitates the treatment of the whole batch problem in just a few minutes, even for long audio time series. Other than the scaling factors of unselected variables, which are fixed to $v_0 = 0.8 \cdot 10^{-3}$, $v_0' = 2.4 \cdot 10^{-3}$ in the two-layer case, all the hyperparameter values of Section IV-A are kept. The number of iterations is set to 80. Results are summarized in Table II.

Results obtained on the polyphonic music excerpt are consistent with those presented in Section IV-A: as in the simpler short-length examples, structured priors obtain better quantitative results than their unstructured counterpart, with the two-layer representation achieving the highest SNR_{out} and superior perceptual qualities. Also in line with results of Section IV-A, the two layer model proves adequate to characterize percussive sounds, with the transient layer capturing predominantly the onset of drum hits. Although the SNR_{out} value of 19.47 dB is slightly lower than those obtained for the same example in [14], the audible quality can be regarded as superior, with less noticeable artifacts and a particularly clean reconstruction of the vocals and plucked string sounds. Runtimes are significantly shorter than those of algorithms computing similar models: in [14], the same piece of music is divided into shorter chunks to be processed in parallel using a Gibbs sampling scheme, taking slightly above one hour per each fragment.

[9]The original song is titled *Mama Vatu* by Susheela Raman.

C. Missing data interpolation

In scenarios with missing data, only a subset of the original measurements y_t are accessible. As commonly done in the image inpainting literature (e.g., [22]), data loss can be modelled by a decimation matrix $\mathbf{H}$ that pre-multiplies the "synthesized" signal $\mathbf{Gc}$, as in:

$$\log p(\mathbf{y}|\mathbf{c}, \sigma^2) = -\frac{1}{2\sigma^2}\|\mathbf{y}_{\mathcal{T}} - \mathbf{HGc}\|^2 + \text{Const.} \qquad (27)$$

where $\mathbf{y}_{\mathcal{T}}$ denotes the vector containing the observed samples. Thus the framework described throughout the previous sections can be readily adapted to address this problem simply by replacing the descent step of Algorithm 2 with $\mathbf{z}^* = \mathbf{z}^{(k)} + \frac{1}{L}\mathbf{G}^\top\mathbf{H}^\top\left(\mathbf{y} - \mathbf{HGz}^{(k)}\right)$, where, due to the sparse structure of $\mathbf{H}$–it contains merely one non-zero value per row–, matrix products $\mathbf{HG}$ and $\mathbf{G}^\top\mathbf{H}^\top$ can be computed using standard point-wise matrix operations of any numerical computing environment (MATLAB, NumPy, etc.), entailing minor overhead.

In this experiment, a jazz trumpet (plus band) excerpt is corrupted with a series of gaps so as to emulate the impulsive noise affecting old recordings (e.g., gramophone discs, magnetic tapes). Specifically, the input is generated by selecting a 16 second long extract sampled at 44.1 kHz, adding white Gaussian noise so as to reach a target value of SNR_{in}–this value is referred to as "SNR_{in} (before data loss)" in Table III–, and introducing random gaps of length 2-5 ms (approximately 80 to 240 samples at the chosen sampling frequency), separated by bits of 2-7 ms. A one-layer Gabor dictionary with 50% overlap and window length $W = 4096$ is used (smaller values of W lead to lower SNR_{out} values in this particular example). The input signal is analyzed with both the unstructured and the temporal Markov chain one-layer models. Transient analysis is dismissed in this context, since in practice many transients fall partially or entirely in the missing bits of the signal, complicating their characterization. Results are shown in Table III, where data loss rates, SNR_{out} and runtimes were averaged over 10 realizations (since the randomness of the data loss mechanism leads to more variability in the outcomes than the sole random additive noise). For comparison purposes, analogue results for the glockenspiel example of Section IV-A are included (same setting, except that Gabor window length is set back to $W = 2048$). As in Section IV-B, for the jazz trumpet extract a longer EM run was found to produce marginally better results, so that the number of iterations is set to 80 (that number is kept for the glockenspiel example).

TABLE II
RESULTS FOR POLYPHONIC EXAMPLE

	SNR_{in} [dB]	One layer - unstructured		One layer - structured		Two layers - structured	
		SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]	Runtime [s]
Mama Vatu (∼24 s)	13.07	18.76	42.28	18.89	55.02	19.47	138.10

TABLE III
RESULTS FOR MISSING DATA INTERPOLATION

Recording	Duration [s]	SNR_{in} [dB] (before data loss)	SNR_{in} [dB] (after data loss)	Data loss rate	Unstructured		Markov chain prior	
					SNR_{out} [dB]	Runtime [s]	SNR_{out} [dB]	Runtime [s]
Jazz trumpet	16	20	4.08	38.67%	11.74	62.52	12.61	69.82
Glockenspiel	∼3	15	3.85	38.84%	16.65	8.75	17.12	9.55

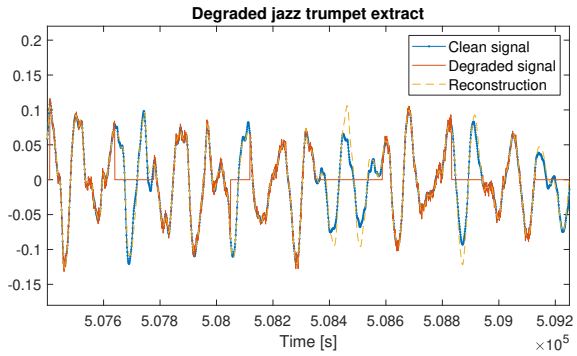

Fig. 5. Detail of reconstructed missing segments (jazz trumpet example, structured prior)

The methodology proposed in this subsection provides a framework for simultaneous noise reduction and missing data interpolation. Results are competitive with those of previous approaches, reaching higher SNR_{out} values than works like [29], where a similar setting is explored[10]. Again, the use of structured priors leads to better results, both quantitatively and qualitatively. Figure 5 shows how the missing bits are filled in with the signal estimate. Even though the interpolation does not always mimic the original signal accurately, the imputed segments smoothly match the neighboring bits, providing a convincing and pleasant-sounding reconstruction, especially in the light of the degraded input signals. This is achieved by virtue of smooth basis functions of the Gabor dictionary and the structured prior of $\boldsymbol{\gamma}$; through these mechanisms the model indirectly extracts information from the neighboring measurements. Attempts to reproduce the good results obtained with window length $W = 4096$ (jazz example) with other configurations were unsuccessful, which points out the relevance of considering different resolutions and time scales in audio analysis.

V. DISCUSSION

This work presents an Expectation-maximization algorithm for signal reconstruction using fixed Gabor dictionaries. A Gaussian mixture spike-and-slab prior for regression coefficients is adopted, facilitating the derivation of closed-form EM steps. Conditional on their activation variables, regression coefficients are given independent Student's t priors, which are conveniently represented as scale mixtures of normals. Accordingly, coefficient variances are estimated as additional parameters of the model within the EM algorithm. Persistence over the axes of the time-frequency surface is modeled through Markov chain prior on the activation coefficients, for which an ad-hoc E-step is derived: leveraging the structure of the corresponding conditional density, marginal probabilities of activation coefficients can be exactly calculated by means of the Forward-backward algorithm. An M-step for Markov chain's parameters (transition probabilities) is also derived. The ridge regression sub-problem associated to the regression coefficients is efficiently solved through the Fast Iterative Shrinkage/Thresholding Algorithm. Variable selection is carried out after the EM procedure, calculating the most likely state sequences (given the estimated regression coefficients) by means of the Viterbi algorithm, and signal reconstruction is performed by keeping the coefficients thus selected. The scheme is proved suitable for multi-resolution representations, and applied to transient estimation tasks and missing data interpolation.

The scheme is significantly more efficient than sampling-based alternatives for computing similar models, and comparable in terms of the quality of the results, both measurably and subjectively. Whereas reconstructions using the proposed algorithm obtains SNR_{out} values in the same range as those obtained computing similar models through MCMC and exhibits similar audible quality (arguably with less artifacts), the computational time is reduced from the order of hours to seconds or minutes. Furthermore, it is empirically observed that the proposed EM scheme generates consistent results (regardless of initial conditions) and does not get stuck in non-meaningful local maxima. To a great extent, efficiency relies on the FISTA technique to compute the M-step of $\mathbf{c}$. The FISTA scheme can be extended to address problems with different covariance structures, since it suffices to replace term $\frac{1}{2}\|\mathbf{y} - \mathbf{G}\mathbf{c}\|^2$ with the (also convex) function $\frac{1}{2}[\mathbf{y} - \mathbf{G}\mathbf{c}]^\top \mathbf{M}[\mathbf{y} - \mathbf{G}\mathbf{c}]$, for which tractable "descent" steps can be derived in certain scenarios, e.g., if $\mathbf{M}$ is a banded matrix.

Although most results were obtained using the same, or very similar, hyperparameter values, signal restoration is not completely adaptive or "blind": hyperparameters (most notably v_0, the scaling factor of discarded variables) require some degree

[10]In that work, though, an extract from the same recording is downsampled (due to computational efficiency reasons) and degraded with gaps. This and the use of a shorter window length make a direct comparison difficult.

of fine-tuning. Strategies to automatically tune v_0, either using statistical criteria (e.g., comparing different models in terms of their marginal likelihoods, including v_0 as a model parameter, etc.) or adopting an audio-oriented approach (e.g., assessing different outputs with perceptually motivated measures), are topics of interest for future research. Another downside of the proposed framework is its intrinsic dependence on the noise term σ^2 required for the conjugate model for the coefficient variances $\sigma^2 v_{m,n}$. Although in theory this may hinder the suitability of our method for compression or encoding tasks (i.e., finding sparse representations of clean signals), adding very small noise[11] works well in practice, generating sparse reconstructions that are often perceptually indistinguishable from the clean signal. In those cases structured priors do not entail noticeable advantages and might even be detrimental.

The framework is not limited to the dictionary representation and structured priors adopted in this work. Instead, it could be readily adapted for other bases (e.g., wavelets, discrete cosine transform, learned dictionaries, and uneven grid of frequencies, etc.) or hierarchical structures. As shown in [19], models of this type are suitable for arbitrary Gaussian Markov Random Field priors on activation coefficients $\boldsymbol{\gamma}$, which may be a more accurate for certain signals—sources such as human voice tend to form "patches" in the time-frequency surface, rather than mere horizontal patterns. The comparatively poorer results obtained for string sounds, as well as noticeable presence of artifacts found in the polyphonic example, support that a finer modelling of audio signals may be beneficial in certain contexts. The proposed scheme is also suitable for other multi-resolution models. Other potential applications include using the proposed EM scheme in higher level tasks (for example, the identification of transients may be helpful for tempo tracking) or embedding it in other signal restoration tasks. For example, the EM solution could be used to feed an MCMC algorithm to compute a more complex model, or exploited in a real-time setting where efficiency is more crucial than high precision (e.g., applying the EM scheme to sequentially process mini-batches of a data stream).

APPENDIX

A. Probability densities: formulae reminder

Beta distribution:

$$\mathcal{B}(\phi;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

Gamma distribution:

$$\mathcal{G}(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\exp\left(-\beta x\right)$$

Inverse-Gamma distribution:

$$\mathcal{IG}(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right).$$

B. EM updates for the multilayer case

FISTA algorithm to solve the M-step upgrade of $\mathbf{c}$ (coefficients of tonal layer) and $\mathbf{c}'$ (coefficients of transient layer):

[11]Noise values such that SNR_{in} = 60 dB were tested.

Algorithm 3: FISTA for the two-resolution model

Input: $\mathbf{y}, \mathbf{G}, \mathbf{G}', \mathbf{c}^{(0)} = \mathbf{z}^{(0)} = \mathbf{x}_{\text{init}}, \mathrm{d} = 4, L$
$\mathbf{c}'^{(0)} = \mathbf{z}'^{(0)} = \mathbf{x}'_{\text{init}}, \mathbf{D}^*, \mathbf{D}'^*$

Output: $\hat{\mathbf{c}}, \hat{\mathbf{c}}'$.

for $k = 0, \ldots, K_{\max} - 1$ **do**
- $\Delta = \mathbf{y} - \mathbf{G}\mathbf{z}^{(k)} - \mathbf{G}'\mathbf{z}'^{(k)}$;
- $\mathbf{z}^* = \mathbf{z}^{(k)} + \frac{1}{L}\mathbf{G}^\top\Delta$;
- $\mathbf{z}'^* = \mathbf{z}'^{(k)} + \frac{1}{L}\mathbf{G}'^\top\Delta$;
- $\mathbf{c}_{m,n}^{(k+1)} = \left(\mathbf{I} + \frac{1}{L}\mathbf{D}^*\right)^{-1}\mathbf{z}^*$;
- $\mathbf{c}_{m,n}'^{(k+1)} = \left(\mathbf{I} + \frac{1}{L}\mathbf{D}'^*\right)^{-1}\mathbf{z}'^*$;
- $\mathbf{z}^{(k+1)} = \mathbf{c}^{(k+1)} + \frac{k}{k+\mathrm{d}}(\mathbf{c}^{(k+1)} - \mathbf{c}^{(k)})$;
- $\mathbf{z}'^{(k+1)} = \mathbf{c}'^{(k+1)} + \frac{k}{k+\mathrm{d}}(\mathbf{c}'^{(k+1)} - \mathbf{c}'^{(k)})$;

end

$\hat{\mathbf{c}} = \mathbf{c}^{(K_{\max})}, \hat{\mathbf{c}}' = \mathbf{c}'^{(K_{\max})}$;

With $\mathrm{P} = 2(\mathrm{M}/2+1)\mathrm{N} + 2(\mathrm{M}'/2+1)\mathrm{N}'$ being the total number of coefficients, the update of σ^2 is:

$$[\sigma^2]^{(k+1)} = \Bigg[\|\mathbf{y} - \mathbf{G}\mathbf{c}^{(k+1)} - \mathbf{G}'\mathbf{c}'^{(k+1)}\|^2 + \nu\lambda + \ldots$$

$$\sum_{m,n} d^*_{m,n}\|\mathbf{c}^{(k+1)}_{m,n}\|^2 + \sum_{m,n} d'^*_{m,n}\|\mathbf{c}'^{(k+1)}_{m,n}\|^2\Bigg] \Big/ \left[\mathrm{T} + \mathrm{P} + \nu + 2\right]$$

Denoting $\#00$ the count of transitions from state 0 to 0 (analogous for any other transitions), we define quantities $N'_{00} = \alpha_{\phi'_{00}} + \mathbb{E}_{\boldsymbol{\gamma}'|\cdot}(\#00)$ and $N'_{01} = \beta_{\phi'_{00}} + \mathbb{E}_{\boldsymbol{\gamma}'|\cdot}(\#01)$. Note that the expected count $\mathbb{E}_{\boldsymbol{\gamma}'|\cdot}(\#00)$ is carried out across all frames since, unlike the tonal part, common parameters ϕ'_{00}, ϕ'_{11} control transitions in all the vertical chains of the transient layer. With N'_{11} and N'_{10} defined analogously, M-step update rules for ϕ'_{00} and ϕ'_{11} are given by:

$$\phi_{00}'^{(k+1)} = \frac{N'_{00} - 1}{N'_{00} + N'_{01} - 2}, \quad \phi_{11}'^{(k+1)} = \frac{N'_{11} - 1}{N'_{11} + N'_{10} - 2}.$$

M-step update of Markov chain initial probability π_1 is:

$$\pi_1^{(k+1)} = \frac{\alpha_{\pi_1} + \sum_{n'} p'^*_{1,n'} - 1}{\alpha_{\pi_1} + \beta_{\pi_1} + \mathrm{N}' - 2}$$

Rules for π_1, ϕ'_{00} and ϕ'_{11} follow from observing that the additional functional $Q_2(\pi_1, \phi'_{00}, \phi'_{11}|\cdot)$ (similar to Eq. 13, but for the newly included parameters) and recognizing the forms of a the logarithm of a Beta density.

REFERENCES

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

[2] S. Ghael, A. M. Sayeed, and R. G. Baraniuk, "Improved wavelet denoising via Empirical Wiener Filtering," in *SPIE Technical Conference on Wavelet Applications in Signal Processing*, 1997.

[3] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *arXiv preprint arXiv:1411.3230*, 2014.

[4] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 433–440.

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

[8] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

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

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

[11] ——, "Approaches for Bayesian variable selection," *Statistica sinica*, pp. 339–373, 1997.

[12] P. J. Wolfe, S. J. Godsill, and W.-J. 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.

[13] F. Li and N. R. Zhang, "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics," *Journal of the American statistical association*, vol. 105, no. 491, pp. 1202–1214, 2010.

[14] C. Févotte, B. Torrésani, L. Daudet, and S. 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.

[15] J. Griffin, K. Latuszynski, and M. Steel, "In search of lost (mixing) time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p," *arXiv preprint arXiv:1708.05678*, 2017.

[16] G. Zanella and G. Roberts, "Scalable importance tempering and Bayesian variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

[17] T. Hayashi and H. Iwata, "EM algorithm for Bayesian estimation of genomic breeding values," *BMC genetics*, vol. 11, no. 1, p. 3, 2010.

[18] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on signal processing*, vol. 46, no. 4, pp. 886–902, 1998.

[19] V. Ročková and E. I. George, "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 828–846, 2014.

[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[22] A. Chambolle and C. Dossal, "On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"," *Journal of Optimization theory and Applications*, vol. 166, no. 3, pp. 968–982, 2015.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[24] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[25] A. E. Gelfand and A. F. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, vol. 85, no. 410, pp. 398–409, 1990.

[26] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[27] J. J. Benedetto, C. Heil, and D. F. Walnut, "Differentiation and the Balian-Low theorem," *Journal of Fourier Analysis and Applications*, vol. 1, no. 4, pp. 355–402, 1994.

[28] T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor analysis and algorithms*. Springer, 1998, pp. 267–294.

[29] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–517.

[30] J. Murphy and S. 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.

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

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

[33] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, "An introduction to total variation for image analysis," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, no. 263-340, p. 227, 2010.

[34] C. Févotte and M. Kowalski, "Estimation with low-rank time–frequency synthesis models," *IEEE Transactions on Signal Processing*, vol. 66, no. 15, pp. 4121–4132, 2018.

[35] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

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

[37] S. Godsill, P. Rayner, and O. Cappé, "Digital audio restoration," in *Applications of digital signal processing to audio and acoustics*. Springer, 2002, pp. 133–194.

[38] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[39] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[40] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[41] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, p. 910167, 2003.

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