Random models for audio signals expansion on hybrid MDCT dictionaries

Matthieu Kowalski, Bruno Torrésani

Random models for audio signals expansion on hybrid MDCT dictionaries

Matthieu Kowalski and Bruno Torrésani*

Abstract

A new approach for signal expansion with respect to hybrid dictionaries, based upon probabilistic modeling is proposed and studied, with emphasis on audio signal processing applications. The signal is modeled as a sparse linear combination of waveforms, taken from the union of two orthonormal bases, with random coefficients. The behavior of the analysis coefficients, namely inner products of the signal with all basis functions, is studied in details, which shows that these coefficients may generally be classified in two categories: significant coefficients versus unsignificant coefficients. Conditions ensuring the feasibility of such a classification are given. When the classification is possible, it leads to efficient estimation algorithms, that may in turn be used for de-noising or coding purpose. The proposed approach is illustrated by numerical experiments on audio signals, using MDCT bases.

Index Terms

Sparse Representations, Non-linear signal approximation, Time-frequency decompositions, Denoising.

I. Introduction

Sparse decomposition and approximation methods have recently exhibited a great potential for several signal processing tasks, such as denoising, coding and compression, or source separation. Often, a

Matthieu Kowalski and Bruno Torrésani are with Laboratoire d'Analyse, Topologie et Probabilités, Université de Provence, 39 rue F. Joliot-Curie, 13453 Marseille cedex 13, France. E-mail: kowalski@cmi.univ-mrs.fr, torresan@cmi.univ-mrs.fr

This work received support from the European Commission funded Research Training Network HASSIP (HPRN-CT-2002-00285).

Matthieu Kowalski is supported by a joint grant of the French Centre National de la Recherche Scientifique (CNRS), and of the Région Provence Alpes Côte d'Azur (PACA).

EDICS: SSP-SSAN; SPC-CODC; MAL-BAYL

dictionary of elementary waveforms is chosen, which respect to which signal expansions are seeked. The dictionary has to be complete, but is often redundant, so that the expansion of the signal is not unique, and sparsity is then used as a selection criterion.

In the context of audio signal processing, dictionaries constructed as unions of two (or more) orthonormal bases of the signal space have been studied in more details (see for example [1]), and it has been shown that such dictionaries may be used efficiently for signal coding for example. The rationale for such a choice is the fact that audio signal generally feature significantly different components (termed layers), which can hardly be accounted for by the same bases. Using two (or more) bases allow one to encode separately the signal's components that are coherent with the different bases. It also yields a decomposition of the signal into different layers, for example the tonal, transient and residual layers in [1], [2]). Standard choices for the bases are MDCT (modified discrete cosine transform) bases, and/or wavelet bases. This is the way we shall follow in this paper. Notice that similar ideas have also found their way in other applications, such as image processing (see for instance [3] and references therein).

Sparsity may be implemented in various ways. Let us quote among others variational approaches, greedy algorithms (see e.g. [4]), or Bayesian formulations (see [5], [6]), these approaches being often closely connected. In these situations, it may be shown that if the dictionary is sufficiently *incoherent* (in our case, if the two bases are different enough, in a sense to be specified), the expansion may be recovered if the signal indeed admits a sparse enough expansion.

Bayesian approaches have the advantage of relying on explicit signal and observation models, from which MAP or MMSE estimates can be derived. The actual computation of these estimates then relies either on classical descent techniques (when an equivalent variational formulation can be found), or on more complex optimization schemes (for example, MCMC schemes) when the functional to be optimized is more complex. Such approaches have been developed for audio signal regression and denoising using a pair of MDCT bases [6]. It was shown there that these are indeed extremely promising approaches, that unfortunately still require high computational power.

The approach we describe here is based upon simpler considerations, and the study of *observed coefficients*, or *analysis coefficients*, i.e. inner products of the signal with the elements of the dictionary (when the dictionary is redundant, these do not necessarily correspond to the coefficients yielding sparse expansions, termed *synthesis coefficients*). When a signal model is specified sufficiently precisely, the behavior of these coefficients may be suitably characterized, which leads to simple algorithms for the identification.

We focus here on a signal model of the form

$$x(t) = \sum_{\gamma \in \Gamma} X_\gamma c_\gamma u_\gamma(t) \ ,$$

where Γ is some generic index set, the waveforms u_γ form the dictionary labeled by Γ, the X_γ are Boolean random variables controlling the sparsity of the expansion, and the coefficients c_γ are independent normal random variables. We provide a thorough analysis of the behavior of the analysis coefficients, and we show that in such a context, under appropriate conditions, they may be (with good accuracy) modeled via Gaussian mixture models. The latter may in turn be identified using appropriate estimation algorithms. The conditions are mainly 1) the sparsity of the signal model (controlled by the distribution of the Boolean variables X_γ), and 2) the incoherence of the dictionary (controlled by a series of weights, to be introduced below).

When these conditions are fulfilled, the estimation algorithms yield good estimates for the significance maps (i.e. the subset of the index set Γ within which $X_\gamma = 1$), and a corresponding subset of the dictionary. We then show that the corresponding synthesis coefficients may be estimated by regression (either standard L^2 regression, or sparse regression), involving the estimated sub-dictionary.

We limit ourselves to simple random models for the significance map Γ: Bernoulli model and hierarchical Bernoulli model. The latter allows us to introduce *structures* in the coefficient domain, i.e. explicit dependencies between neighboring (in the time-frequency domain) coefficients.

The theoretical analysis of this paper (provided in Section II) and the corresponding algorithms (described in Section III, and in the appendix for more specific points) are illustrated by a number of numerical results on audio signal denoising and coding (Section IV). We limit ourselves to a dictionary built as the union of two MDCT bases with different time-frequency resolutions. The narrow band basis (i.e. with long window) is used to estimate a *tonal layer* in the signal, and the wide band basis (i.e. small window) is used to estimate the *transient layer*.

Our results show that the model above is generally adequate for describing audio signals, provided they don't contain random-like components (as do wind instruments for example). It provides results whose quality is comparable with concurrent approaches, but generally requires much lower computational power.

II. HYBRID WAVEFORM SIGNAL MODELS

A. Generalities

Let $\mathcal{H}$ denote a (finite or infinite dimensional) separable Hilbert space, and let $\mathbf{V} = \{v_n, n \in I\}$ and $\mathbf{U} = \{u_n, n \in I\}$ be two orthonormal bases of $\mathcal{H}$. Here, I denotes a generic index set (in the finite dimensional situation, we denote $I = \{1, \dots N\}$). We denote by

$$\mathcal{D} = \mathbf{V} \cup \mathbf{U}$$

the dictionary made as the union of these two bases. $\mathcal{D}$ is clearly (over)complete in $\mathcal{H}$, and any $x \in \mathcal{H}$ admits infinitely many expansions in the form

$$x = \sum_{n \in I} \alpha_n v_n + \sum_{m \in I} \beta_m u_m \ ,$$

where $\alpha_n, \beta_m \in \mathbb{C}$ are the *synthesis coefficients*. We are interested in *sparse signals*, i.e. signals $x \in \mathcal{H}$ that may be written as

$$x = \sum_{\lambda \in \Lambda} \alpha_\lambda v_\lambda + \sum_{\delta \in \Delta} \beta_\delta u_\delta + r \ , \tag{1}$$

where Λ, Δ are small subsets of the index set I, termed *significance maps* and $r \in \mathcal{H}$ is a small (possibly vanishing) residual.

Given such a sparse signal, the non-uniqueness of its expansion with respect to the dictionary makes it difficult to identify unambiguously the model (1). The approach we propose uses the *analysis coefficients*

$$a_n = \langle x, v_n \rangle \ , \ b_m = \langle x, u_m \rangle \ , \tag{2}$$

and develops a strategy to estimate the relevant such coefficients, from which a sparse expansion may be identified.

In this work, we limit ourselves to a specific pair of orthonormal bases: $\mathbf{U}$ is a local trigonometric (i.e. an MDCT basis, see for example [7]) basis (tuned in such a way to achieve good frequency resolution), and $\mathbf{V}$ is a local trigonometric basis with good time resolution. The index sets are then two-dimensional (a time index and a frequency index), and we write them as such when necessary. Other choices for the bases are possible (for example a combination of MDCT and wavelet bases, as in [1], [2]), as well as extensions to frames (that would however require significant modifications).

B. Random hybrid models

Let us now introduce an explicit *model* for the sparse signal in (1). The ingredients of such models are essentially twofold: a model for the *significance maps* Λ and Δ, and, given the significance maps Λ and Δ, a model for the coefficients $\{\alpha_\lambda, \lambda \in \Lambda\}$ and $\{\beta_\delta, \delta \in \Delta\}$.

Definition 1: Given two orthonormal bases in $\mathcal{H}$ as above, a corresponding *random hybrid model* is defined by

i. A discrete probability model for the significance maps. The corresponding probability measures for the (random) maps Λ and Δ will be denoted by $\mathbb{P}_\Lambda$ and $\mathbb{P}_\Delta$, and the expectations by $\mathbb{E}_\Lambda$ and $\mathbb{E}_\Delta$.

ii. A probability model for the synthesis coefficients $\{\alpha_\lambda, \lambda \in \Lambda\}$ and $\{\beta_\delta, \delta \in \Delta\}$, conditional to the significance maps. The corresponding probability measure and expectation will be denoted by $\mathbb{P}_0$ and $\mathbb{E}_0$.

We shall denote by X_n and $\tilde{X}_n$ the indicator random variables, corresponding to the maps Λ and Δ, i.e.

$$X_n = \begin{cases} 1 & \text{if } n \in \Lambda \\ 0 & \text{otherwise} \end{cases}, \quad \tilde{X}_n = \begin{cases} 1 & \text{if } n \in \Delta \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

and by p_n and $\tilde{p}_n$ the membership probabilities

$$p_n = \mathbb{P}_\Lambda \{X_n = 1\} \ , \ \tilde{p}_n = \mathbb{P}_\Delta \left\{\tilde{X}_n = 1\right\} \ . \tag{4}$$

The corresponding signal model therefore takes the form

$$x = \sum_{n \in I} X_n \alpha_n v_n + \sum_{m \in I} \tilde{X}_m \beta_m u_m + r \ . \tag{5}$$

The simplest possible model for the significance maps is the *Bernoulli* model: given a fixed *membership probability* $p_n = p \ \forall n$, the index values $n \in I$ are iid, and belong to Λ with probability p and to $\overline{\Lambda}$ (the complementary set) with probability $1 - p$. The membership probability for Δ will be denoted by $\tilde{p}_n = \tilde{p} \ \forall n$. More sophisticated models for the significance maps, which we term *structured models*, can involve correlations between elements of the significance maps.

The simplest instance for coefficient models, to which we shall stick here, assumes that significant coefficients are independent $\mathcal{N}(0, \sigma_n^2)$ random variables, in other words their pdf (conditional to Λ and Δ) reads

$$\begin{aligned} \rho_{\alpha_n}(z|\Lambda) &= (1 - X_n)\delta_0(z) + X_n \mathcal{N}(0, \sigma_n^2) \ , \\ \rho_{\beta_n}(z|\Delta) &= (1 - \tilde{X}_n)\delta_0(z) + \tilde{X}_n \mathcal{N}(0, \tilde{\sigma}_n^2) \ . \end{aligned}$$

The residual is modeled here as a Gaussian white noise, with variance s^2.

The variances σ_n^2 and $\tilde{\sigma}_n^2$ are coefficient dependent, as follows. We now specialize to the case where the atoms u_n and v_n are time-frequency atoms, i.e. n is actually a time-frequency index, namely a couple $n = (k_n, \nu_n)$ of integers. In this situation, we shall assume that the variances σ_n and $\tilde{\sigma}_n$ only depend on the frequency index

$$\sigma_n = \sigma_{\nu_n} = \sigma f(\nu_n) \ , \ \tilde{\sigma}_n = \tilde{\sigma}_{\nu_n} = \tilde{\sigma}\tilde{f}(\nu_n) \ , \tag{6}$$

f and $\tilde{f}$ being fixed *frequency profiles*, that model "typical" decay of the coefficients with respect to frequency, and σ, $\tilde{\sigma}$ being normalized so that $f(\nu_n) \leq 1$, $\tilde{f}(\nu_n) \leq 1$.

C. Behavior of the analysis coefficients

Given this hybrid waveform model, and a realization x of a corresponding signal, the parameters and the significance maps may be estimated in a purely Bayesian framework by considering their posterior probability distribution, conditional to the observation. This approach has proven efficient for audio signal denoising, at the price of high computational costs [8].

In this work, we have chosen to stick to a simpler approach, based on the study of the *analysis coefficients*, defined in (2), from which a sparse expansion of x with respect to the dictionary is estimated.

As a first step, let us start by studying the distribution of these analysis coefficients, *conditional to the significance maps*. Setting $\rho_n = \langle r, v_n \rangle$ and $\tilde{\rho}_n = \langle r, u_n \rangle$, one easily sees that

$$a_n = \langle x, v_n \rangle \ = \alpha_n X_n + \sum_{m \in I} \beta_m \tilde{X}_m \langle u_m, v_n \rangle + \rho_n \tag{7}$$

$$b_n = \langle x, u_n \rangle \ = \beta_n \tilde{X}_n + \sum_{m \in I} \alpha_m X_m \langle v_m, u_n \rangle + \tilde{\rho}_n \ , \tag{8}$$

i.e. that the analysis coefficients may be expressed as sums of independent Gaussian random variables. Thus one can state

Proposition 1: Conditional to the significance maps, the a_k and b_k coefficients are zero-mean normal random variables, with covariance matrices $\mathcal{C}_{k\ell} = \mathbb{E}_0 \{a_k \overline{a}_\ell\}$, $\tilde{\mathcal{C}}_{k\ell} = \mathbb{E}_0 \left\{b_k \overline{b}_\ell\right\}$, given by

$$\begin{aligned} \mathcal{C}_{k\ell} &= (\sigma_k^2 X_k + s^2)\delta_{k\ell} + \sum_{m \in I} \tilde{X}_m \tilde{\sigma}_m^2 \langle v_k, u_m \rangle \langle u_m, v_\ell \rangle \ , \\ \tilde{\mathcal{C}}_{k\ell} &= (\tilde{\sigma}_k^2 \tilde{X}_k + s^2)\delta_{k\ell} + \sum_{m \in I} X_m \sigma_m^2 \langle u_k, v_m \rangle \langle v_m, u_\ell \rangle \ . \end{aligned}$$

In particular, the diagonal terms read

$$\mathbb{E}_0 \left\{|a_k|^2\right\} = \sigma_k^2 X_k + \sum_{m \in I} \tilde{X}_m \tilde{\sigma}_m^2 |\langle v_k, u_m \rangle|^2 + s^2 \ . \tag{9}$$

Hence, the a (resp. b) coefficients are distributed according to a (random) mixture of (several) normally distributed zero-mean random variables. The distribution of these is governed by the cross term in the right hand side of the covariance coefficients in Proposition 1. Focusing on the diagonal terms of the covariance matrix, let us introduce the following quantities

Definition 2: Let Δ and Λ be two subsets of the index set I. For $n \in I$, the weighted projection weights, or γ weights, are defined by

$$\tilde{\gamma}_n(\Delta) = \sum_{m \in I} \tilde{X}_m \tilde{f}(\nu_m)^2 |\langle v_n, u_m \rangle|^2 \gamma_n(\Lambda) = \sum_{m \in I} X_m f(\nu_m)^2 |\langle u_n, v_m \rangle|^2 \ . \tag{10}$$

with ν_m the frequency component of the time-frequency index $m = (k_m, \nu_m)$.

Notice that the γ weights are random variables. Their distributions will play a key role in the upcoming analysis.

Remark 1: The γ weights are reminiscent of the Parseval weights

$$\tilde{p}_n(\Delta) = \sum_{m \in I} \tilde{X}_m |\langle v_n, u_m \rangle|^2 \ , \ p_n(\Lambda) = \sum_{m \in I} X_m |\langle u_n, v_m \rangle|^2 \ .$$

introduced in [9], [10]. Indeed, in the simple case of constant variances $\sigma_n = \sigma \ \forall n$, one has $\gamma_n(\Lambda) = p_n(\Lambda)$, and a similar expression for the $\tilde{\gamma}$ weights. The Parseval weights have a simple geometric interpretation, namely $p_n(\Delta)$ is the norm of the orthogonal projection of v_n onto the linear span of $\{u_\delta, \delta \in \Delta\}$. The γ weights may be given a similar interpretation. Let us denote by $\mathbb{M}$ (resp. $\tilde{\mathbb{M}}$) the operator defined by a diagonal matrix in the $\mathbf{V}$ (resp. $\mathbf{U}$) basis

$$\mathbb{M} v_n = f(\nu_n) v_n \ , \ \tilde{\mathbb{M}} u_n = \tilde{f}(\nu_n) u_n \ . \tag{11}$$

Then

$$\gamma_n(\Lambda) = \sum_{m \in I} X_m |\langle u_n, \mathbb{M} v_m \rangle|^2 = \sum_{m \in I} X_m |\langle \mathbb{M} u_n, v_m \rangle|^2 \ ,$$

(The last inequality being a consequence of the assumption $f(\nu) \leq 1$), and $\gamma_n(\Lambda)$ is then the squared norm of the orthogonal projection of $\mathbb{M} u_n$ onto the linear span of the basis functions $\{v_m, m \in \Lambda\}$. In addition it follows from Parseval's formula that for all n and Λ,

$$\gamma_n(\Lambda) \leq \|\mathbb{M} u_n\|^2 \leq 1 \ ,$$

and a similar expression for $\tilde{\gamma}_n(\Delta)$.

$\mathbb{M}$ is a well-defined operator in the finite dimensional case. In infinite-dimension situations, *i.e.* for continuous time signals, additional assumptions on the frequency profiles f, $\tilde{f}$ are needed to ensure the boundedness of $\mathbb{M}$.

Remark 2: The diagonal terms (9) of the covariance matrix take the following form:

$$\mathbb{E}_0\left\{|a_k|^2\right\} = \begin{cases} \sigma_k^2 + \tilde{\gamma}_k(\Delta)\tilde{\sigma}^2 + s^2 & \text{if } k \in \Lambda \\ \tilde{\gamma}_k(\Delta)\tilde{\sigma}^2 + s^2 & \text{if } k \notin \Lambda \end{cases} . \tag{12}$$

Taking into account the γ weights leads to the following simple consideration on the behavior of observed coefficients: if the distribution of the γ weights is peaked near a small value, then the coefficients a_k have a significantly different behavior depending on whether X_k vanishes or not. In addition, the smaller the variance of the weights, the easier the discrimination between the two behaviors.

Characterizing the distribution of the γ weights is not an easy task (moment estimates in the case of the Bernoulli model are provided below). Nevertheless, if we assume that the elements of the significance map Λ (resp. Δ) are identically distributed, with $\mathbb{P}\{n \in \Lambda\} = p$ (resp. $\mathbb{P}\{n \in \Delta\} = \tilde{p}$), *mean-field* type estimates, *i.e.* estimates for Λ or Δ averages of the γ weights, may be obtained. For example, the first moment of the γ weights reads

$$\mathbb{E}_\Lambda\left\{\gamma_n(\Lambda)\right\} = p\,\|\mathbb{M}u_n\|^2 \;;\; \mathbb{E}_\Delta\left\{\tilde{\gamma}_n(\Delta)\right\} = \tilde{p}\,\|\tilde{\mathbb{M}}v_n\|^2 \;. \tag{13}$$

We give below the mean field estimates for the a coefficients, similar estimates may be derived for the b coefficients.

Proposition 2: Assume that the elements of the significance map Λ are identically distributed, with $\mathbb{P}\{n \in \Delta\} = \tilde{p}$. Then we have the mean field estimate

$$\mathbb{E}_\Lambda\left\{\mathbb{E}_0\left\{a_k^2\right\}\right\} = \sigma_k^2 X_k + \tilde{p}\|\mathbb{M}u_k\|^2\tilde{\sigma}^2 + s^2 \;.$$

Our goal will be to estimate the significance map Δ from the analysis coefficients. In this respect, it is convenient to normalize the analysis coefficients by the frequency profiles. In such a way, the variances are stabilized, in the sense that the leading term below has constant variance σ^2:

$$\mathbb{E}_\Lambda\left\{\mathbb{E}_0\left\{\frac{a_k^2}{f(\nu_k)^2}\right\}\right\} = \sigma^2 X_k + \tilde{p}\frac{\|\mathbb{M}u_k\|^2}{f(\nu_k)^2}\tilde{\sigma}^2 + \frac{s^2}{f(\nu_k)^2} \tag{14}$$

$$= \sigma^2 X_k + \tilde{p}\;\tilde{\sigma}^2 \sum_{m\in I}\frac{\tilde{f}(\nu_m)^2}{f(\nu_k)^2}|\langle v_m, u_k\rangle|^2 + \frac{s^2}{f(\nu_k)^2} \;. \tag{15}$$

We notice that the distribution of the renormalized coefficients is governed by the quantity $\sum_{m\in I}\frac{\tilde{f}(\nu_m)^2}{f(\nu_k)^2}|\langle v_m, u_k\rangle|^2$.

Remark 3: This renormalization ensures that the leading term $\sigma^2 X_k$ in (14) has a constant variance σ^2. The variance of the second term varies as a function of k. However, for suitable choices of the bases $\mathbf{U}$ and $\mathbf{V}$, and the frequency profiles f, $\tilde{f}$, this dependence is weak enough to allow one to approximate the mean field distribution of renormalized coefficients using a mixture of two or three Gaussian distributions.

D. Significance maps estimation in the case of the Bernoulli model

Assume that the points of the index set are iid. Then the probability distribution of the significance map is given by

$$\mathbb{P}\{\Delta\} = \tilde{p}^{|\Delta|}(1-\tilde{p})^{N-|\Delta|} \ , \ \mathbb{P}\{\Lambda\} = p^{|\Lambda|}(1-p)^{N-|\Lambda|} \ ,$$

and the marginal distribution of the analysis coefficients takes the simple form

$$\rho_{a_n}(\xi) \ = \ (1-p)\sum_{\Delta}\mathbb{P}\{\Delta\}\mathcal{N}(0,\tilde{\gamma}_n(\Delta)\tilde{\sigma}^2+s^2) + p\sum_{\Delta}\mathbb{P}\{\Delta\}\mathcal{N}(0,\sigma^2+\tilde{\gamma}_n(\Delta)\tilde{\sigma}^2+s^2) \ . \quad (16)$$

The distribution of the coefficients is thus a mixture of two Gaussian mixtures, whose behavior is governed by the γ weights. Assume for the sake of simplicity that the distribution of the random variables $\tilde{\gamma}_n(\Delta)$ is sharply concentrated near a small value, say the membership probability $\tilde{p}$ (see (13)). Then the two Gaussian mixtures are zero-mean, and possess significantly different variances. In such situations, one may attempt to separate them, in order to estimate those index values n that belong to the significance map. The separation will be based on the amplitude of the coefficients: large coefficients will be assigned to the significance map. We describe below how the corresponding threshold values are estimated.

As mentioned before, in the Bernoulli model, moment estimates of the distribution of the γ weights may be obtained, in addition to the first given in (13). For the second moment one has

$$\begin{aligned}\mathbb{E}_\Lambda\left\{\gamma_n(\Lambda)^2\right\} &= p^2\sum_{m\neq m'}\sigma_m^2\sigma_{m'}^2|\langle u_n,v_m\rangle|^2|\langle u_n,v_{m'}\rangle|^2 + p\sum_m\sigma_m^4|\langle u_n,v_m\rangle|^4 \\ &= \left(\mathbb{E}_\Lambda\left\{p_n(\Lambda)\right\}\right)^2 + p(1-p)\sum_m\sigma_m^4|\langle u_n,v_m\rangle|^4 \ ,\end{aligned}$$

hence

$$\begin{aligned}\mathrm{Var}\{\gamma_n(\Lambda)\} &= p(1-p)\sum_m f(\nu_m)^4|\langle u_n,v_m\rangle|^4 \\ &= p(1-p)\sum_m|\langle u_n,\mathbb{M}v_m\rangle|^4 \ . \end{aligned} \quad (17)$$

The third moment can also be calculated, and yields the skewness

$$\mathrm{S}\{\gamma_n(\Lambda)\} = \frac{\mathbb{E}\left\{\gamma_n(\Lambda)^3\right\}}{\mathbb{E}\left\{\gamma_n(\Lambda)^2\right\}^{3/2}} = \frac{1-2p}{\sqrt{p(1-p)}}\frac{\sum_{k=1}^N|\langle v_k,\mathbb{M}u_n\rangle|^6}{\left(\sum_{k=1}^N|\langle v_k,\mathbb{M}u_n\rangle|^4\right)^{3/2}} \ . \quad (18)$$

Remark 4: As stressed in Remark 2 above, discriminating between the two types of analysis coefficients is easier when the first and second order moments of the γ weights are small.

1) The first moment is essentially controlled by the sparsity of the expansion, represented here by the membership probability p. The sparser the significance maps, the smaller the γ weights.
2) The variance is controlled by the membership probability p and the incoherence of the dictionary. Indeed, introducing $B_4 = \sup_n \sum_m |\langle u_n, \mathbb{M} v_m \rangle|^4$, which may be seen as a generalization of the $4 - Babel$ function [11], we obtain $\text{Var}\{\gamma_n(\Lambda)\} \leq p(1-p)B_4$.
3) The skewness is controlled by the position of p relative to $1/2$.

The separation of Gaussian mixtures problem can be formulated as follows. Denote by Y_n the MAP estimate for X_n:

$$Y_n = \begin{cases} 1 & \text{if } \mathbb{P}\{X_n = 1 | a_n, \Delta\} \geq \mathbb{P}\{X_n = 0 | a_n, \Delta\} \\ 0 & \text{otherwise} \end{cases} .$$

This MAP estimate for X_n will give a threshold adapted for *each* analysis coefficients, which correspond to the intersection of the two Gaussian curves of the mixture. More precisely, we have:

$$\begin{aligned} \mathbb{P}\{X_n = q | a_n, \Delta\} &\propto \mathbb{P}\{X_n = q\} \mathbb{P}\{a_n | X_n, \Delta\} \\ &\propto \begin{cases} p\, \mathcal{N}(0, \sigma_n^2 + \tilde{\gamma}_n(\Delta)\tilde{\sigma}_n^2 + s^2) \text{ if } q = 1 \\ (1-p)\, \mathcal{N}(0, \tilde{\gamma}_n(\Delta)\tilde{\sigma}_n^2 + s^2) \text{ if } q = 0 \end{cases} . \end{aligned} \tag{19}$$

Set for simplicity

$$w_{n;0} = \tilde{\gamma}_n(\Delta)\tilde{\sigma}^2 + s^2 \ , \ w_{n;1} = w_{n;0} + \sigma_n^2 \ ,$$

then one can state

Proposition 3: 1) Assume the elements of the significance maps Λ (resp. Δ) are iid, with $\mathbb{P}\{n \in \Lambda\} = p$ (resp. $\mathbb{P}\{n \in \Delta\} = \tilde{p}$). Assume the synthesis coefficients α_n (resp. β_n) are independent $\mathcal{N}(0, \sigma_n^2)$ (resp. $\mathcal{N}(0, \tilde{\sigma}_n^2)$) random variables. Then the MAP estimate Y_n for X_n is given by:

$$Y_n = \begin{cases} 1 & \text{if } |a_n| \geq \tau_n \\ 0 & \text{otherwise} \end{cases} ,$$

with $\tau_n = \sqrt{\frac{2\ w_{n;1} w_{n;0}}{w_{n;1} - w_{n;2}} \ln\left[\left(\frac{1-p}{p}\right)\left(\frac{w_{n;1}}{w_{n;2}}\right)\right]}$.

2) The type one and type two error probabilities read as follows

$$\mathbb{P}\{Y_k = 0 | X_k = 1\} = p \text{ erf}\left(\sqrt{\frac{w_{k;0}}{\sigma_k^2} \ln\left[\frac{p}{1-p}\left(1 + \frac{\sigma_k^2}{w_{k;0}}\right)\right]}\right) ,$$

$$\mathbb{P}\{Y_k = 1 | X_k = 0\} = (1-p) \text{ erfc}\left(\sqrt{\left(1 + \frac{w_{k;0}}{\sigma_k^2}\right) \ln\left[\frac{1-p}{p}\left(1 + \frac{\sigma_k^2}{w_{k;0}}\right)\right]}\right) .$$

Proof: The intersection point τ_n of the two Gaussians $\mathcal{N}(0, w_{n;0})$ and $\mathcal{N}(0, w_{n;1})$, is given by

$$\tau_n^2 \;=\; \frac{2\; w_{n;1} w_{n;0}}{w_{n;1} - w_{n;2}} \ln \left[\left(\frac{1-p}{p} \right) \left(\frac{w_{n;1}}{w_{n;2}} \right) \right] \;,$$

which proves the first part. For the second part, we simply observe that

$$\mathbb{P}\{Y_k = 0 | X_k = 1\} = p \text{ erf} \left(\sqrt{\frac{\ln \left[\left(\frac{1-p}{p} \right) \left(\frac{w_{k;1}}{w_{k;0}} \right) \right]}{w_{k;1}/w_{k;0} - 1}} \right)$$

$$\mathbb{P}\{Y_k = 1 | X_k = 0\} = (1-p) \text{ erfc} \left(\sqrt{\frac{\ln \left[\left(\frac{1-p}{p} \right) \left(\frac{w_{k;1}}{w_{k;0}} \right) \right]}{1 - w_{k;0}/w_{k;1}}} \right)$$

where erfc is the complementary error function [12]. ■

Remark 5: The error probabilities for the significance map Λ are controlled by $w_{n;0}/\sigma_k^2$. Again, we notice that the γ weights play a crucial role: the smaller $\tilde{\gamma}_k(\Delta)$ (and the noise variance), the lower the error probabilities.

E. Structured models

1) Generalities: Unlike the Bernoulli model, structured significance maps models involve correlations between significance map elements. For example, assuming $\mathbf{U}$ is an orthonormal basis of time-frequency waveforms, correlations may be introduced between consecutive time indices, to model time persistence properties of the corresponding (tonal) layer. Similarly, correlations between frequencies may be introduced to model signal components with short duration, such as transients (frequency persistence). In such situations, the marginal pdf of observed coefficients is still given by (16), but the probabilities are not as simple as before.

Interestingly enough, due to the decorrelation of the α and β coefficients, the correlations in significance maps do not show up in the second order moments of the observed coefficients a and b, i.e. in matrices $\mathcal{C}$ and $\tilde{\mathcal{C}}$. For instance, neither $\mathbb{E}_\Lambda \left\{ \mathcal{C}_{k\ell}^\Lambda \right\}$ nor $\mathbb{E}_\Delta \left\{ \mathcal{C}_{k\ell}^\Lambda \right\}$ involve the correlation functions of the significance maps $\Gamma_{k\ell} := \mathbb{E}_\Lambda \left\{ X_k X_\ell \right\}$ or $\tilde{\Gamma}_{k\ell} := \mathbb{E}_\Lambda \left\{ \tilde{X}_k \tilde{X}_\ell \right\}$.

2) Hierarchical Bernoulli model: Dependencies between neighboring coefficients may be introduced in a simple way by replacing the above Bernoulli model with a hierarchical Bernoulli one. We present the model in the framework of the transient layer modeling. The idea is to account for time values at which no transient coefficient exist, and segment the time indices into transient and non transient ones. Notice that a similar model could also be developed for the tonal significance map Δ.

Let $n = (k, \nu) \in \Lambda = \Lambda_t \times \Lambda_f$ a time-frequency index, with Λ_t (resp. Λ_f) the time (resp. frequency) index set. Let X_n denote the corresponding indicator random variables and T_{k_n} the time indicator random variables. The random variables X_n are distributed following a Bernoulli law $\mathcal{B}(p_1)$ conditionally to the time indicator variables T_{k_n} which are distributed following a Bernoulli law $\mathcal{B}(p_2)$. That can be written as

$$\tilde{X}_n \sim \mathcal{B}(\tilde{p}) \ ; \ X_k \sim T_k \mathcal{B}(p_2) + (1 - T_k)\delta_0 \ , \ \text{with} \ \ T_k \sim \mathcal{B}(p_1) \ . \tag{20}$$

To estimate the significance map, we first focus on the submap Λ_t. Instead of using as before the analysis coefficients one by one, all coefficients on the same time index are collected into the following quantity

$$\begin{aligned} c_k &= \sum_{\nu=1}^{|\Lambda_f|} \frac{a_{k,\nu}^2}{w_{k,\nu;0}^2} \\ &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;0}} X_{k,\nu} + \frac{1}{w_{k,\nu;0}} \sum_{\delta \in \Delta} \beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu} \right]^2 \ . \end{aligned} \tag{21}$$

Let us set for simplicity $\tilde{\pi}_{k,\nu} = \frac{1}{w_{k\nu;0}} \sum\limits_{\delta \in \Delta} \beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu}$. We rewrite (21) to separate the coefficients c_k with $k \in \Lambda_t$ from the others

$$c_k = \begin{cases} \sum\limits_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;0}} X_{k,\nu} + \tilde{\pi}_{k,\nu}(\Delta) \right] & \text{if } k \in \Lambda_t \\ \sum\limits_{\nu=1}^{|\Lambda_f|} \tilde{\pi}_{k,\nu}(\Delta)^2 & \text{if } k \notin \Lambda_f \end{cases} \ . \tag{22}$$

The coefficients β_δ are distributed according to a normal law $\mathcal{N}(0, \tilde{\sigma}_\delta^2)$, and the coefficients $\rho_{k,\nu}$ according to $\mathcal{N}(0, s^2)$. Then, the coefficients $\tilde{\pi}_{k,\nu}(\Delta)$ are normally distributed according to

$$\tilde{\pi}_{k,\nu}(\Delta) \sim \mathcal{N} \left(0, \frac{\sum\limits_{\delta \in \Delta} \tilde{\sigma}_\delta^2 |\langle u_\delta, v_{k,\nu} \rangle|^2 + s^2}{w_{k,\nu;0}} \right) \sim \mathcal{N}(0, 1) \ .$$

In case $k \notin \Lambda_f$, the coefficients c_k are distributed according to a χ^2 with $|\Lambda_f|$ degrees of freedom. Coefficients $c_{k,\nu}$, $k \in \Lambda_t$, are expected to take large values and appear as outliers for the above mentioned χ^2 distribution.

The main shortcoming of such an approach is that the significance map Δ has to be known in order to normalize the coefficients, and then to obtain the coefficients c_k. To avoid this, we limit ourselves to

approximations of the c_k coefficients, and introduce new coefficients c'_k:

$$\begin{aligned} c'_k &= \sum_{\nu=1}^{|\Lambda_f|} \frac{a_{k,\nu}^2}{f(\nu)^2} \\ &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{f(\nu)} X_{k,\nu} + \sum_{\delta\in\Delta} \frac{\beta_\delta}{f(\nu)} \langle u_\delta, v_{k,\nu}\rangle + \frac{\rho_{k,\nu}}{f(\nu)} \right]^2 \\ &= \begin{cases} \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{f(\nu)} X_{k,\nu} + \tilde{\pi}'_{k,\nu}(\Delta) \right]^2 & \text{if } k \in \Lambda_t \\ \sum_{\nu=1}^{|\Lambda_f|} \tilde{\pi}'_{k,\nu}(\Delta)^2 & \text{if } k \notin \Lambda_f \end{cases} , \end{aligned} \tag{23}$$

with $\tilde{\pi}'_{k,\nu} = \sum_{\delta\in\Delta} \frac{\beta_\delta \langle u_\delta, v_{k,\nu}\rangle + \rho_{k,\nu}}{f(\nu)} \sim \mathcal{N}\left(0, \frac{w_{k,\nu;0}}{f(\nu)}\right)$.

Although the variances of the $\pi_{k,\nu}$ are different, the distribution of the $\{c'_{k,\nu},\ k \notin \Lambda_t\}$ may be approximated with a good accuracy by a two parameters χ^2 law, and the coefficients $\{c'_{k,\nu},\ k \in \Lambda_t\}$, may be seeked as outliers for that χ^2 law.

After the pre-selection of analysis coefficients $a_{k,\nu}$, $k \in \Lambda_t$, the Bernoulli model is used to complete the selection, and to obtain an estimate of the significance map.

Remark 6: An another point of view would amount is to normalize all the coefficients c_k according to their membership to the submaps Λ_t:

$$\begin{aligned} d_k &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;1}} X_{k,\nu} + \frac{1}{w_{k,\nu;0}} \sum_{\delta\in\Delta} \beta_\delta \langle u_\delta, v_{k,\nu}\rangle + \rho_{k,\nu} \right]^2 \text{ if } k \in \Lambda_t \ , \\ d'_k &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{1}{w_{k,\nu;0}} \sum_{\delta\in\Delta} \beta_\delta \langle u_\delta, v_{k,\nu}\rangle + \rho_{k,\nu} \right]^2 \text{ if } k \notin \Lambda_t \ . \end{aligned} \tag{24}$$

In this case, we will obtain $p_1|\Lambda_t|$ coefficients d_k distributed according to a χ^2 with $|\Lambda_f|$ degrees of freedom, and $(1-p_1)|\Lambda_t|$ coefficients d'_k distributed according to the same χ^2 as previously. A MAP estimation, denoted by Z_k for the random variables T_k can be formulated as follow

$$Z_k = \begin{cases} 1 & \text{if } p\chi^2(d_k) > (1-p)\chi^2(d'_k) \\ 0 & \text{otherwise} \end{cases} . \tag{25}$$

F. Variances estimation

If estimates are available for the significance maps, the γ weights can be estimated too. The next proposition then gives powerful estimators for the parameters σ and $\tilde{\sigma}$.

Proposition 4: Let p and $\tilde{p}$ denote the membership probabilities. Let a_n (resp. b_n) be the analysis coefficients and $f(\nu_n)$ (resp. $\tilde{f}(\nu_n)$) the corresponding frequency profiles. Let

$$v_1 = \frac{1}{|\Lambda|}\sum_{n\in\Lambda}\frac{a_n^2}{f(\nu_n)^2}\ ,\ v_2 = \frac{1}{|\Delta|}\sum_{n\in\Delta}\frac{b_n^2}{\tilde{f}(\nu_n)^2}\ .$$

Then, the estimates defined by

$$\hat{\sigma}^2 = \frac{v_1 - \epsilon_1 v_2}{1-\epsilon_1\epsilon_2}\ ,\ \hat{\tilde{\sigma}}^2 = \frac{v_2 - \epsilon_2 v_1}{1-\epsilon_1\epsilon_2} \tag{26}$$

with $\epsilon_1 = \frac{1}{|\Lambda|}\sum_{n\in\Lambda}\frac{\tilde{\gamma}_n(\Delta)}{f(\nu_n)^2}$ and $\epsilon_2 = \frac{1}{|\Delta|}\sum_{n\in\Delta}\frac{\gamma_n(\Lambda)}{\tilde{f}(\nu_n)^2}$, are convergent and unbiased.

Proof: First, the estimate for σ and $\tilde{\sigma}$ are unbiased: in matrix form, we have

$$\mathbb{E}\left\{\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right\} = \begin{pmatrix} 1 & \epsilon_1 \\ \epsilon_2 & 1 \end{pmatrix}\begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}\ .$$

Solving the linear system shows that the above estimators $\hat{\sigma}^2$ and $\hat{\tilde{\sigma}}^2$ are unbiased.

To prove the convergence of the estimator, we just have to prove that the variance vanishes as the number of observations goes to infinity

$$\text{var}\{\hat{\sigma}\} = \frac{\text{var}\{v_1\} + \epsilon_1^2\text{var}\{v_2\} - 2\epsilon_1\text{cov}\{v_1, v_2\}}{(1-\epsilon_1\epsilon_2)^2}\ .$$

v_1 (resp. v_2) is the classical estimator for the expectation of $\frac{a_n^2}{f(\nu_n)^2},\ n\in\Lambda$ (resp. $\frac{b_m^2}{\tilde{f}(\nu_m)^2},\ m\in\Delta$). We have then

$$\text{var}\{v_1\} \underset{N\to\infty}{\to} 0\ ,\ \text{var}\{v_2\} \underset{N\to\infty}{\to} 0\ .$$

We just have to prove that the covariance $\text{cov}\{v_1, v_2\} \underset{N\to\infty}{\to} 0$. For this, we need to compute for $n\in\Lambda$ and $m\in\Delta$

$$\mathbb{E}\left\{\frac{a_n^2}{f(\nu_n)^2}\frac{b_m^2}{\tilde{f}(\nu_m)^2}\right\} = \sigma^2\tilde{\sigma}^2 + \sigma^2\tilde{\sigma}^2\frac{\gamma_m(\Delta)}{\tilde{f}(\nu_m)^2}\frac{\tilde{\gamma}_n(\Lambda)}{f(\nu_n)^2} + \sigma^4\frac{\langle v_n, u_m\rangle^2}{\tilde{f}(\nu_m)^2} + \tilde{\sigma}^4\frac{\langle u_m, v_n\rangle^2}{f(\nu_n)^2} + 4\sigma^2\tilde{\sigma}^2\langle u_m, v_n\rangle^2\ .$$

Then,

$$\begin{aligned}
\text{cov}\{v_1, v_2\} &= \mathbb{E}\{v_1 v_2\} - \mathbb{E}\{v_1\}\mathbb{E}\{v_2\} \\
&= \sigma^2\tilde{\sigma}^2 + \sigma^2\tilde{\sigma}^2\epsilon_1\epsilon_2 + \sigma^4\epsilon_2 + \tilde{\sigma}^4\epsilon_1 + \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{n\in\Lambda}\sum_{m\in\Delta}\langle u_m, v_n\rangle^2 - (\sigma^2 + \epsilon_1\tilde{\sigma}^2)(\tilde{\sigma}^2 + \epsilon_2\sigma^2) \\
&= \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{n\in\Lambda}\sum_{m\in\Delta}\langle u_m, v_n\rangle^2 \le \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{n\in\Lambda}\sum_{m=1}^{N}\langle u_m, v_n\rangle^2 = \frac{4\sigma^2\tilde{\sigma}^2}{|\Delta|} \underset{N\to\infty}{\to} 0\ ,
\end{aligned} \tag{27}$$

which concludes the proof. ■

G. Coefficients estimation

Estimation of the coefficients can be done after the estimation of the significance maps. This can be done by regression, or "wiener type" algorithm. We postpone the discussion of this operation to the end of section III.

III. ALGORITHMS

In this section we describe in some details all the algorithms deduced from the analysis of the model in section II. We first develop mean-fields algorithms which will be used as initialization for iterative algorithms described just after to estimate the significance maps. After this estimation, two methods are presented to estimate the synthesis coefficients. Some more specific details on implementation are provided at the end of the section.

In these practical applications, $\mathcal{H}$ is a finite dimensional space. We denote by $N = \dim(\mathcal{H})$ the length of the signal.

A. Mean field algorithms

Initially, no information about the significance map is available. The mean field approximations naturally yields estimates for the significance maps, that may used either directly (see the denoising applications in section IV-B), or as initialization for a more precise approach. The two significance maps are estimated independently of each other.

1) Bernoulli model: It has been shown in section II-C that the distribution of the coefficients a_n and b_n can be approximated by a mixture of a small number of Gaussian distributions. The corresponding estimation may be performed by a suitable `EM` type algorithm. A classification of the analysis coefficients according to the estimated Gaussian gives the estimate of the significance map.

The `EM` algorithm we propose is slightly modified to be able to process in parallel on the renormalized analysis coefficients and the original ones. This is necessary, in order to take the noise into account. Indeed, as may be seen from equation (14), in the case $k \notin \Lambda$ and $s^2 >> \tilde{\sigma}^2\tilde{\gamma}(\Delta)^2$, the Gaussian distribution corresponding to the noise is deformed by the renormalization: it is therefore necessary to work on the original analysis coefficients to estimate the parameters of this Gaussian distribution. Theoretical aspects and technical details on this modified `EM` algorithm are provided in Appendix A.

According to Proposition 3, the classification of analysis coefficients is equivalent to an adaptive thresholding: the largest coefficients are considered to correspond to the significance map, but the threshold depends on the coefficients.

The choice of the number of terms (two or three) to estimate in the Gaussian mixture with the EM algorithm depends of the target application. When the distribution of coefficients is fitted using three Gaussians, the third one corresponds to modeling of the noise, and the corresponding coefficients are not taken into account. Therefore, this produces sparser significance maps.

1) Separation of a mixture of two Gaussian distributions:
   a) A first large variance Gaussian function which corresponds to the analysis coefficients belonging to the significance map, and a second small variance one for the other ones. We use on the renormalized analysis coefficients $\frac{a_k}{f(\nu_k)}$ (see equation (14))
   b) If the noise is expected to have a large variance, the Gaussian with large variance is estimated on the renormalized analysis coefficients, and the Gaussian with small variance on the original analysis coefficients (this Gaussian correspond to the noise).
2) Separation of a mixture of three Gaussian distributions. Compared to the first algorithm, a third Gaussian distribution is added, with a very small variance. This Gaussian correspond to the noise and will be estimated on the original analysis coefficients. This choice yields sparser maps.

A good practical strategy is to first attempt to separate three Gaussians distributions and, if the estimated map does not contain enough coefficients to describe the signal accurately enough (for example, less than 0.1% of the size of the signal), turn to the "two Gaussians" model instead.

The Bernoulli estimation for the significance map yields quite satisfactory results for the tonal map. For the transient significance map, the hierarchical Bernoulli estimate described below turns out to give better estimations.

2) Hierarchical Bernoulli model: The coefficients c'_k defined in equation (23) are used to obtain an estimate of the transient submap Λ_t, through a statistical test. We showed in section II-E.2 that the distribution of the c'_k coefficients which do not correspond to the transient submap (this will be our null hypothesis) can be approximated by a two parameters χ^2 distribution. We use a goodness of fit test for outliers detection, and thus to detect the coefficients c'_k corresponding to the transient lines. The two hypotheses of the test are

$$H_0 : \{c'_1, \ldots, c'_K\} \text{ follow a } \chi^2 \text{ law}$$

$$H_1 : \{c'_1, \ldots, c'_K\} \text{ do not follow a } \chi^2 \text{ law}$$

This test is used in a classification algorithm which proceed as follow. While the goodness of fit test on the set of the c'_k coefficients is rejected (the set does not follow a χ^2 law), the largest coefficient is rejected. The rejected coefficients are the ones corresponding to the transient lines.

The goodness of fit test we choose is the Kuiper test. The latter is more suitable than a Kolmogorov test, since it gives more importance to the tail of the distribution, where the interesting c'_k coefficients lie. A description and the significance level of different tests can be found in [13]. The statistical test is done at the 1% significance level. Three cases have to be taken into account

- The test is accepted at the beginning. No coefficients correspond to a transient line. An empty set is returned and there is no transient.
- The test is always rejected. All coefficients correspond to a transient line.
- The test is rejected during I iterations. It is the general situation, where I coefficients correspond to a transient line.

Once selection of the transient lines is done, the selection in frequency can be done like in the previous section.

3) Iterative mean-field algorithms: The mean field algorithms can be iterated on the residual obtained after one pass of the algorithm. This can improve the estimate of the different layers.

B. Iterative adaptive thresholding Algorithms

The algorithms described in section III-A above rest on a mean field approximation of the γ weights, which is used to compute coefficients thresholds. We now outline iterative algorithms that use estimates from the previous iteration rather than mean field estimates. Mean field estimates are used as initializations.

1) Bernoulli model: A first estimate for the significance maps gives an estimate for the γ weights. After that, one can estimate all the parameters of the model, thanks to proposition 4 in section II-F.

These estimates can be exploited in a `CEM` algorithm which uses the MAP estimate for X_n and $\tilde{X}_n$ described in proposition 3 in section II-D. The algorithm can be summarized as follow. After an initialization for the significance maps Λ and Δ and the parameters p, $\tilde{p}$, σ and $\tilde{\sigma}$, the following fours stages are iterated:

1) The γ weights are computed.
2) The maps are re-estimated using the estimators given in Proposition 3.
3) The parameters σ and $\tilde{\sigma}$ are re-estimated according to Proposition 4.
4) The parameters p and $\tilde{p}$ are re-estimated with $p = \frac{|\Lambda|}{N}$ and $\tilde{p} = \frac{|\Delta|}{N}$.

It is worth noticing that we do not have any estimate available for the noise variance s, which then has to be known in advance. Alternatively, s may be used as a tuning parameters for the algorithm, which controls sparsity for the maps. For the initialization, we used the estimates given by the algorithms described in section III-A.

2) Hierarchical Bernoulli model: An algorithm similar to the previous one was developed. The only change is the estimation of the significance map Λ which uses first the MAP estimation formulated in remark 6, section II-E.2. This first performs a classification of the analysis coefficients in time, and second exploits the Bernoulli model to conclude the classification.

The parameter p_1 is a tuning parameter of the algorithm: the smaller p_1, the sparser the significance maps. p_2 is estimated by the `EM` algorithm used to conclude the classification.

C. Coefficients estimation

After the significance maps have been estimated, the corresponding significant coefficients may be estimated, which amounts to a regression problem. We assume that the significance maps have been suitably estimated. The estimation of the coefficients can be done using two different approaches:

- A mean-field approach, in which the coefficients are estimated by a minimization of the mean squared error. This approach does not necessarily generate sparse expansions.
- By linear regression, which could improve sparsity if desired.

1) Regression approaches:

a) L^2 regression: Estimation of the significance maps actually amounts to a dimension reduction. Let x a signal, and $\hat{\Lambda}$ and $\hat{\delta}$ be estimates for the significance maps. These estimates generate a subdictionary $\hat{\mathcal{D}} = \{u_\delta, \delta \in \hat{\delta}\} \cup \{v_\lambda, \lambda \in \hat{\Lambda}\}$ of the complete waveform dictionary $\mathbf{U} \cup \mathbf{V}$, and we denote by $\mathcal{H}_{\hat{\mathcal{D}}}$ the subspace of $\mathcal{H}$ spanned by $\hat{\mathcal{D}}$.

The easiest way to estimate the two layers $x_{\mathbf{U}}$ and $x_{\mathbf{V}}$, is to compute an orthogonal projection of the signal x onto $\mathcal{H}_{\hat{\mathcal{D}}}$

$$\hat{x} = \underset{y \in \mathcal{H}_{\hat{\mathcal{D}}}}{\operatorname{argmin}} ||x - y||^2 \ . \tag{28}$$

One can write

$$\hat{x} = \hat{x}_{\mathbf{V}} + \hat{x}_{\mathbf{U}} \ , \tag{29}$$

with

$$\hat{x}_{\mathbf{V}} = \sum_{\lambda \in \hat{\Lambda}} \hat{\alpha}_\lambda v_\lambda \ , \ \hat{x}_{\mathbf{U}} = \sum_{\delta \in \hat{\Delta}} \hat{\beta}_\delta u_\delta \ . \tag{30}$$

The estimates $\hat{\alpha}_\lambda$ and $\hat{\beta}_\delta$ for the coefficients are obtained by solving the linear system

$$\boldsymbol{G} \left(\hat{\alpha}_1, ..., \hat{\alpha}_{|\hat{\Lambda}|}, \hat{\beta}_1, ..., \hat{\beta}_{|\hat{\Delta}|} \right)^T = \left(a_1, ..., a_{|\hat{\Lambda}|}, b_1, ..., b_{|\hat{\Delta}|} \right)^T \ , \tag{31}$$

where $\boldsymbol{G}$ is the gram matrix of the dictionary $\hat{\mathcal{D}}$. The Gram matrix $\mathbf{G}$ is left invertible if and only if the selected atoms form a frame their linear span $\hat{\mathcal{H}} = \text{span}\{\hat{\mathcal{D}}\}$, which is always true here.

b) Sparse regression: To improve sparsity, the orthogonal projection may be replaced with a sparse regression, for example performing

$$\hat{x} = \underset{y \in \hat{\mathcal{H}}}{\operatorname{argmin}} ||x - y||_2^2 + \lambda ||y||_1 \ , \tag{32}$$

where λ is tuning parameter which acts on sparsity. Following Chen and Donoho in [14] for basis pursuit denoising, we choose the default value $\lambda = s\sqrt{2\log(\#\hat{\mathcal{D}})}$.

This sparse regression problem can be solved efficiently using appropriate fixed point algorithms, such as the FOCUSS algorithm [15].

2) Wiener type algorithm: When the signal is not sparse enough, and so the significance maps are too large, inverting the Gram matrix becomes computationally expensive. In such a case, a valuable alternative is provided by a Wiener-type or mean-field method which minimizes the mean squared error (conditional to the significance maps)

$$\hat{x} = \underset{y}{\operatorname{argmin}} \, \mathbb{E}_0 \left\{ ||x - y||^2 \right\} \ , \tag{33}$$

where the estimator $y = \sum_{\lambda \in \Lambda} \hat{\alpha}_\lambda v_\lambda + \sum_{\delta \in \Delta} \hat{\beta}_\delta u_\delta$ is seeked in the special form:

$$\hat{\alpha} = t_\lambda \alpha \ , \ \hat{\beta} = t_\delta \beta \ . \tag{34}$$

The minimization may be performed explicitly, and yields estimates for α_λ and β_δ that take the form of suitably weighted analysis coefficients

$$\hat{\alpha}_\lambda = \frac{\sigma^2}{\sigma^2 + \gamma_\lambda(\Delta)\tilde{\sigma}^2 + s^2} \, a_\lambda \ , \ \hat{\beta}_\delta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \gamma_\delta(\Lambda)\sigma^2 + s^2} \, b_\delta \ . \tag{35}$$

Since the coefficient estimation is posterior to the significance map estimation, estimates for Δ and Λ are available, that can be used in this scheme. These estimates turn out to be poorer estimates, but easier to compute.

D. Miscellaneous details on implementation

We give here additional details on a few practical aspects of the algorithm.

1) Structure of the Gram matrix, and inversion: A potential bottleneck is the computation of the γ weights, which enter the computation of thresholds at each iteration. This requires an efficient computation of the scalar products, which can be done as described below.

Remark 7: In the case of the union of two MDCT bases, the Gram matrix has a very specific structure. Let $\{a_k\}_{k \in \mathbb{N}}$ a sequence of reals and $\{w_n\}_{n \in \mathbb{N}}$ a window of size $\delta \in \mathbb{R}$ such as the sequence $\{u_n\}_{n \in \mathbb{N}}$

$$u_n(t) = w_n(t) \cos\left[\frac{\pi}{\delta}(\nu + \frac{1}{2})(t - a_{m+r}) \right] \ , \ \nu \in \mathbb{N}$$

is the first MDCT basis. We denote by $\{v_n\}_{n\in\mathbb{N}}$ the second MDCT basis. One can state

$$\langle u_m, v_n\rangle = \langle u_{m+r}, v_{n+kr}\rangle \ . \tag{36}$$

This remark allows us to compute the Gram matrix in reasonable time. Furthermore, the fact that the $\mathbf{U}$ and $\mathbf{V}$ basis functions are compactly supported yields extremely sparse matrices, which can be stored with low memory requirement and access time.

Regarding the inversion of the Gram matrix, we notice that the Gram matrix is positive definite. This ensures that its inversion may be done efficiently, using for example a conjugate gradient algorithm (see for example [16]).

2) Segmentation of the signal: In practical situations, the algorithms cannot be applied to a long signal as a whole, mainly for two reasons:

- The statistical properties of audio signals are generally (slowly) varying with time. The parameters of the model cannot remain constant throughout signal, which therefore has to be segmented into blocks so that the model can be considering stable within a given block.
- Most algorithms involved in our approach (like many audio signal processing algorithms) have complexity $O(N \log N)$ or higher. Hence, a large signal cannot be processed as a whole.

Following standard practice, we therefore process long signals by first segmenting them into blocks (a typical length of a block is one of two fifths of a second), and run the algorithm within each block. Therefore, we obtain a set of parameters (variances, membership probabilities,...) for each block.

IV. Numerical results

The model and corresponding algorithms were tested on different "classical" applications: signal separation into two layers for single captor blind sources separation, denoising and audio coding. In each case, we only report here on the results obtained with the best adapted version of the algorithm.

The parameters used for the algorithms are as follows. The sampling rate is 44100 Hz. Unless otherwise specified, we used two MDCT bases, one with a 128 samples long window for modeling transients[1], and one with a 4096 samples window for modeling tonals. We segment of the signal to estimate the transient layer on about 186 ms of signal. The tonal layer is usually estimated on all the signal. The frequency

[1]Even though the corresponding time length, about 3 μs, does not make sense from the perceptual point of view, we obtained significantly better results with such a very short window, the reason being probably that the two windows have to be significantly different to be able to discriminate between different layers.

profiles used are the same for the two bases, and have the following form: $f(\nu) = \frac{1}{1+\frac{\nu}{\nu_0}}$. We chose $\nu_0 = 500$.

The algorithms are coded in MATLAB and performed on a 2x3GHz Linux PC with 2Go RAM.

All sound files corresponding to the examples of this section are available at

`http://www.latp.univ-mrs.fr/~kowalski/IEEE07.html`.

A. Application to separation into two layers: transient + tonal

One of the first applications of signal expansions on unions of bases to audio signals was the $transient+tonal+noise$ separation [1]. This problem may be seen as a single sensor blind source separation problem, the sources being the three layers: transient, tonal and noise.

We follow here these lines, and apply our approach to the problem of separating a "tonal" signal and a more impulsive one from a single mixture. The separation of the two sources is done by the separation in two layers of the signal: the tonal instrument will be recovered in the tonal layer, and the transient instrument will be recovered in the transient layer.

This is illustrated on the following example: an instantaneous mixture of a trumpet signal (the tonal one) and a castanets signal (the transient one). The mixture and the separation are provided in figure 1, sound files are available at the web site given in the introduction of the section IV. We used The third algorithm of the section III-A.1 with three iterations. The size of the window for the tonal MDCT basis is 8192 samples (about 186 ms), and the size of the window for the transient MDCT basis is 128 samples (about 3 ms). The mixture signal has 2^{17} samples (about 3 sec of sound).

The main features of the estimation may be seen in the plots of the estimated layers. As may be seen (and heard from the sound files), the separation is fairly satisfactory. Nevertheless, it clearly appears that the estimated castanets signal lost its "tonal" part, which has been "captured" by the estimated trumpet signal and sounds like artifacts. This is not surprising and correspond to the model.

An objective performance measure (which does unfortunately not make much sense for audio signals) is provided by the signal to noise ratio (SNR). The SNR for the trumpet signal is 10.8 dB and the castanets SNR equals 5.7 dB. The very low SNR obtained for the castanets signal is explained by the loss of the tonal layer. Nevertheless, the interesting information for this signal lies in the transient layer which is very well estimated.

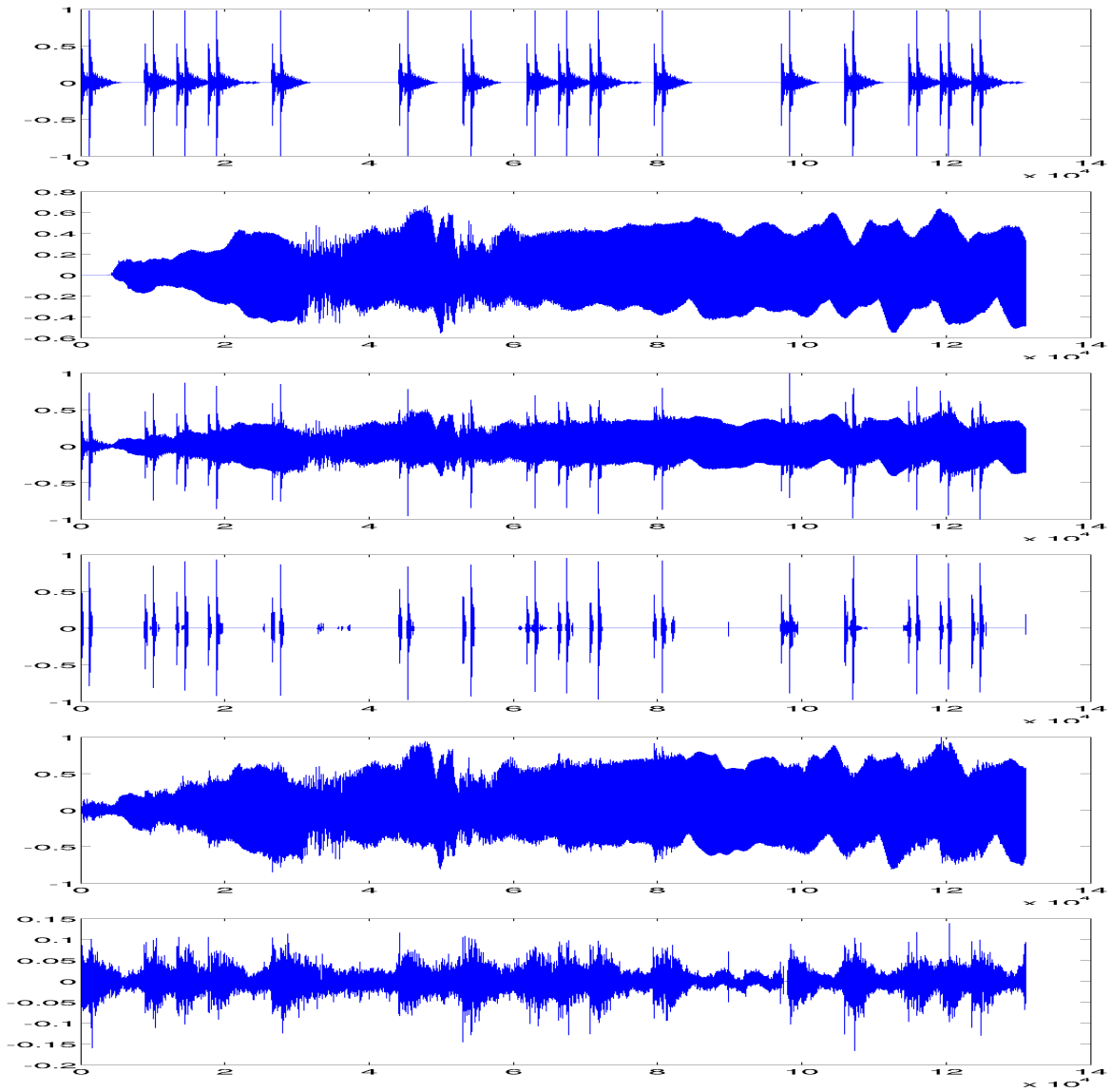

Fig. 1. Blind source separation of trumpet + castanets mixture. From top to bottom: original castanets signal, original trumpet signal, mixture, estimate of the castanets signal, estimate of the trumpet signal, residual.

B. Application to denoising

Denoising is a natural application for this type of decomposition. Additive Gaussian white noise is a standard benchmark, even though it is quite an idealistic situation. In such a case, the noise is not sparse with respect to any basis, and is expected to be recovered in the residual r of equation (1).

The third algorithm of the section III-A.1 mixed with the second was tested on such artificial noisy signal. Gaussian white noise was added to different types of signals, so as to obtain a 6 dB SNR. All

signals have 2^{18} samples (about 6 sec of sound). The results (output SNR) are summarized in tables I for monophonic signals, and II for polyphonic signals (our algorithm is called *Hybrid algorithm*). SNRs are always provided for the reconstruction $tonals + transients$. The restored signals are quite satisfactory. The panpipe signal is especially interesting because of the blow. The blow can be seen like a non-sparse residual in the model (1). As expected, the restored signal provide the panpipe without the blow which has been captured by the residual.

To assert the quality of the results, we compared then to the results obtained with MCMC algorithms provided in [8]. Gaussian white noise was added to the piano signal, so as to yield 10 dB input SNR.

The comparison is provided in table III. In terms of SNR, our results are of lower quality than those obtained by the various MCMC algorithms. This is not surprising, since those approaches are supposed to exploit the complete posterior distribution of synthesis coefficients, while ours relies on approximations. However, let us recall that SNR is not a completely relevant measure of distortion for audio signals, and that complementary evaluations have to be done by listening the signals. From the sound files, it clearly appears that MCMC3 outperforms all other methods, at the price of high computational cost. Our approach provides restored signals that are more pleasant to listen than the MCMC1 and MCMC2. MCMC1 and MCMC2 yield a lot of artifacts and "musical" noise, but our algorithm looses a little bit more high frequencies.

In terms of computing time, our algorithm outperforms all the MCMC algorithms: less than five minutes are needed to process one second of audio signal compared to several hours for the Bayesian+MCMC approaches.

C. Application to audio coding

It has been shown with success that the decomposition into three layers can be useful for audio coding in [17]. The CEM algorithm of the section III-B.1 is used with different value of σ_0 to tune the sparseness of the significance maps. We report here on numerical results obtained using the hierarchical Bernoulli model, with either L^2 regression or sparse regression, and compare them with two other approaches. Again, we use the SNR as a comparison criterion. The two reference approaches are based on MDCT expansion, followed by thresholding of MDCT coefficients, or thresholding of MDCT coefficients weighted by the frequency profiles $f(\nu)$. The latter version is motivated by the fact that our approach uses frequency profiles, that degrade the output SNR but produce better reconstructions from the audio point of view. Therefore, it is more fair to compare the results of our approach with those obtained from weighted MDCT thresholding.

Signal	SNR	Figure
Panpipe	11.6	2
Glockenspiel	16.4	3
Xylophone	12.3	4

TABLE I

SNR AFTER DENOISING FOR SOME MONOPHONIC SIGNALS.

Signal	SNR	Figure
Jazz	16.2	5
Mamavatu	13.3	6

TABLE II

SNR AFTER DENOISING FOR SOME POLYPHONIC SIGNALS.

Algorithms	SNR
MCMC 1	20.7
MCM 2	21.6
MCMC 3	21.6
Jeffrey's + EM	15.3
Hybrid Algorithm	18.2

TABLE III

COMPARISON OF SNR BETWEEN DIFFERENT ALGORITHMS. LINES 1 TO 4 ARE TAKEN FROM [8].

Figure 7 shows the evolution of the SNR as a function of the percentage of retained coefficients (size of the significance maps). As expected (see above) MDCT thresholding yields a significantly better rate distortion curve, while weighted MDCT thresholding is comparable with the two versions of our approach.

The right hand plot of figure 7 shows a zoom of the low rate part of the full plot. As may be seen, for very low rate (less than 1% retained coefficients), our hybrid decompositions outperform weighted MDCT thresholding.

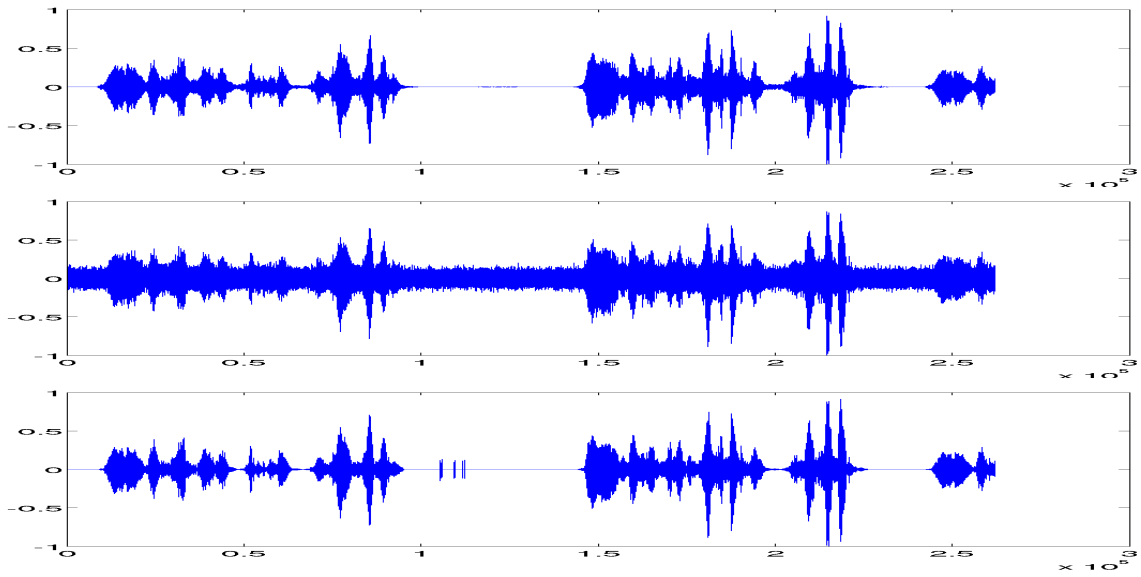

Fig. 2. Panpipe signal. From top to bottom: original signal, noisy signal, restored signal.

V. Conclusion

We have described in this article a family of random waveform models that aim at obtaining sparse representations of audio signals. Compared to other works, the originality of this approach is to start from a mathematical model of the signal able to reproduce the observed statistics of audio signals, like the distribution of the analysis coefficients. We have focused on simple versions of the model, but extensions to more complex situations (in particular more complex significance maps models) are also possible.

The theoretical study of these simple models yields practical algorithms which can be exploited in application like denoising, or decomposition of the signal into layers. The simplicity of the model is reflected by the simplicity of the algorithms themselves, which do not require complicated optimization steps, and therefore need reasonable computing time (without paying special attention to optimization). The mean-field type algorithms are equivalent to adaptive hard-thresholding algorithms, the thresholds being obtained by likelihood maximization of the model. The iterative adaptive thresholding algorithms yield thresholds adapted for *each* coefficient.

The numerical results presented in this paper show the effectiveness of our approach for audio signals. Among the different algorithms presented, the "mean-field" algorithms are the most efficient, in term of expected results, and computation time.

Further work will focus on more realistic models for the significance maps, including more complex

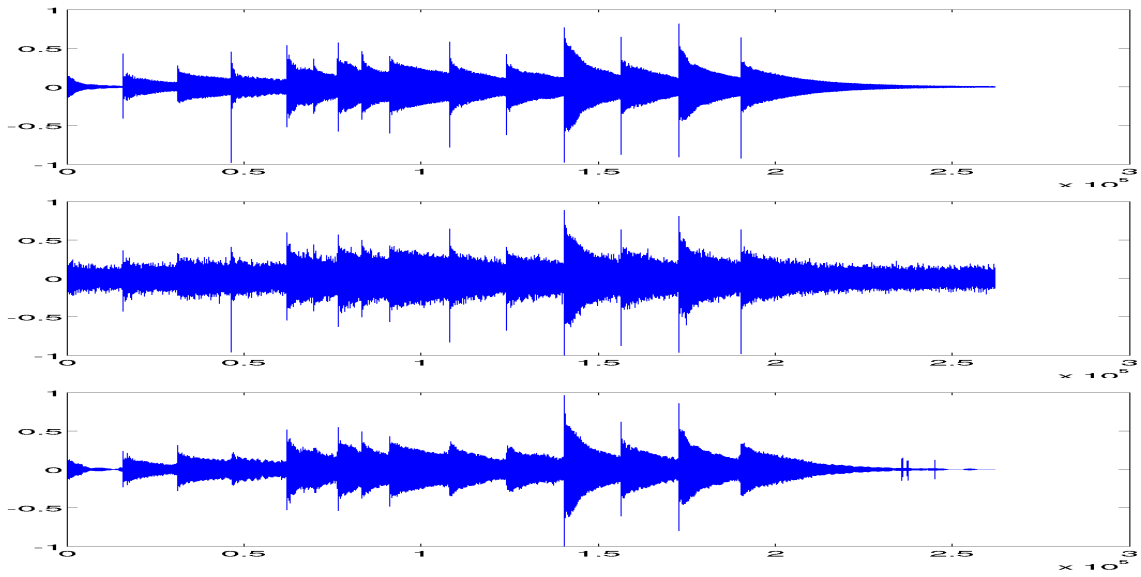

Fig. 3. Glockenspiel signal. From top to bottom: original signal, noisy signal, restored signal.

structures for the maps and between maps. Indeed, although natural from the theoretical and algorithmical points of view, the assumption of independence of the tonal and transient layers is certainly not realistic: a note begins by an attack, so a transient may generally be expected at the beginning of a tonal component. However, extensions of the model in such ways will necessitate much more complex estimation algorithms.

APPENDIX

A. *Expectation Maximization Algorithm (*EM*)*

Let $\{x_1, x_2, \ldots, x_n\}$ denote the observed data, which are independent realizations of a random variable X. The likelihood of the data, conditional to the model with parameter Θ is: $\mathcal{L}(\Theta) = \mathbb{P}(X|\Theta)$.

As the realizations are independent, if $f(.|\Theta)$ denotes the pdf with parameter Θ, one can write:

$$\mathcal{L}(\Theta) = \prod_i^n f(x_i|\Theta) \ . \tag{37}$$

Assume the data follow a *known* mixture model, *after* a transformation ϕ of the coefficients conditionally to their class membership. Denote the classes by $\{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$ and by $\tilde{x} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ the observed data after the transformation, which are the realization of the random variable $\tilde{X}$ defined by:

$$\tilde{x}_i = \phi_k(x_i) \quad \text{if } x_i \in \mathcal{C}_k \ ,$$

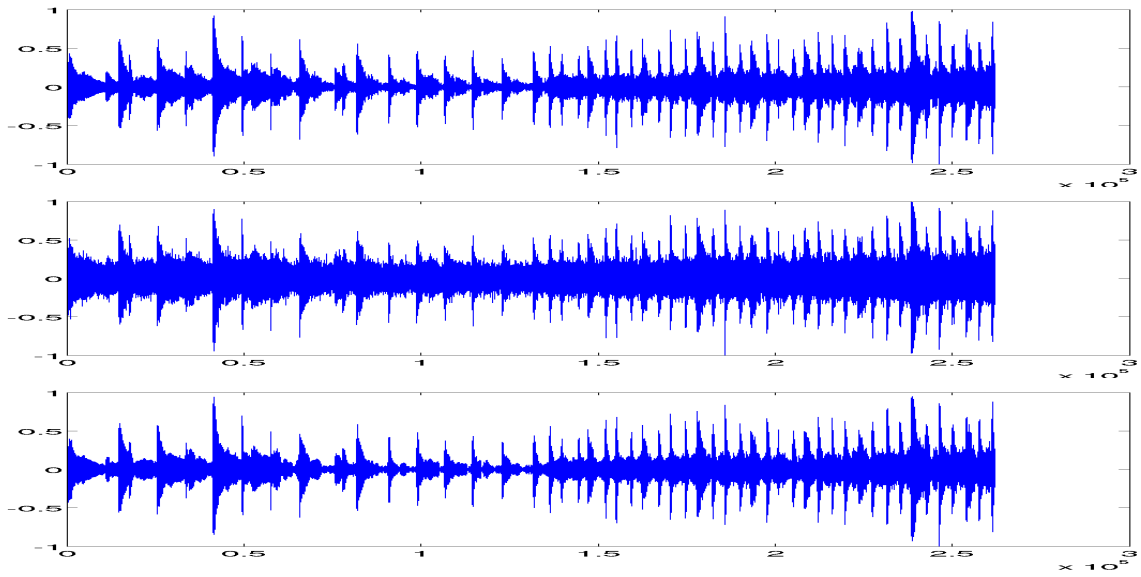

Fig. 4. Xylophone signal. From top to bottom: original signal, noisy signal, restored signal.

with

$$\phi : \ X \in \mathbb{R}^N \longmapsto \tilde{X} = \phi(X) \in \mathbb{R}^N \ .$$

The random variable X is a partial observation. Let Z be a random variable corresponding to the missing hidden data. This random variable show the class of the observation x_i:

$$\begin{cases} z_{i,k} = 1 & \text{if } x_i \in \mathcal{C}_k \\ z_{i,k} = 0 & \text{otherwise} \end{cases} .$$

Denote by $Y = (X, Z)$ the supplemented data and $\tilde{Y} = (\tilde{X}, Z)$ the transformed supplemented data. Let $\pi_k = \mathbb{P}\{Z = k\}$, the complete log-likelihood is:

$$\begin{aligned} \log \mathcal{L}(\tilde{Y}|\Theta) &= \log(P(\tilde{X}, Z|\Theta)) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{c} z_{i,k} \log(\pi_k f(\phi_k(x_i)|\theta_k)) \ . \end{aligned} \tag{38}$$

The $z_{i,k}$, which represent the class of each x_i, allow us to write the log-likelihood, depending of the observed data x_i, the transformations ϕ_k corresponding to the classes $\mathcal{C}_k$, without knowing the partition.

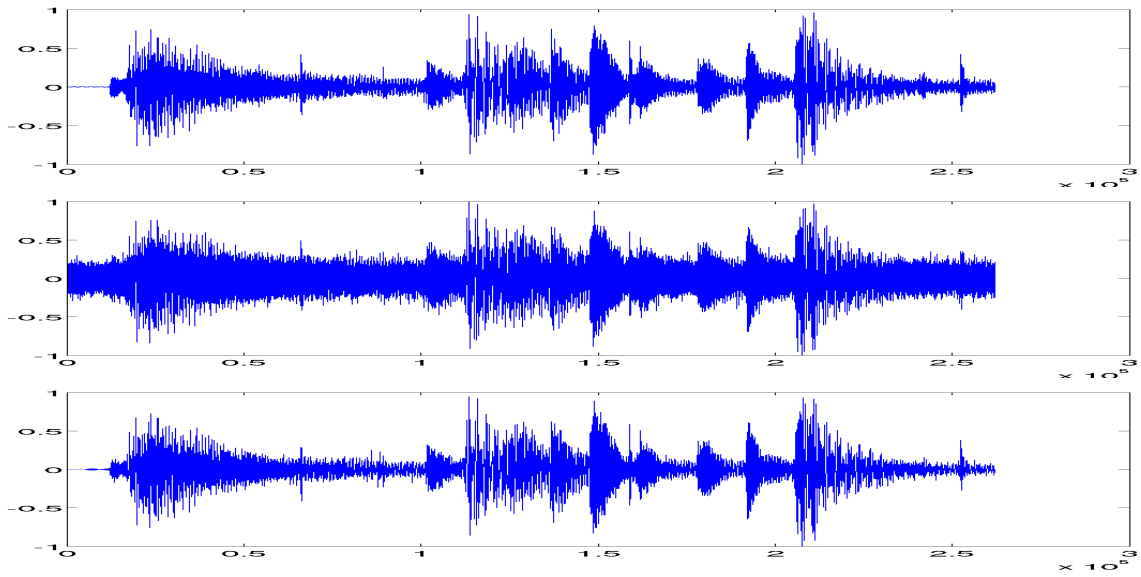

Fig. 5. Excerpt from a Norah Jones song. From top to bottom: original signal, noisy signal, restored signal.

The expectation state at the iteration i is:

$$\begin{aligned} Q(\Theta|\hat{\Theta}^{(t)}) &= \mathbb{E}\{\mathcal{L}(\Theta)|\tilde{X}, \hat{\Theta}^{(t)}\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{c} \mathbb{E}\{z_{i,k}|\tilde{X}, \hat{\Theta}^{(t)}\} \log(\pi_k f(\phi_k(x_i)|\theta_k)) \ , \end{aligned} \tag{39}$$

ie estimate the mean of $z_{i,k}$:

$$\begin{aligned} \hat{z}_{i,k}^{(t)} &= \mathbb{E}\{Z_{i,k}|\tilde{X} = \phi_k(x_i), \hat{\Theta}^{(t)}\} = \mathbb{P}\{Z_{i,k} = 1|\tilde{X} = \phi_k(x_i), \hat{\Theta}^{(t)}\} \\ &= \frac{\pi_k f(\phi_k(x_i), \theta_k)}{\sum_{q=1}^{c} \pi_q f(\phi_q(x_i), \theta_q)} \ . \end{aligned} \tag{40}$$

The maximization state is classically obtained by solving the likelihood equations, depending of the mixture model.

REFERENCES

[1] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002, special issue on Image and Video Coding Beyond Standards. [Online]. Available: http/::www.cmi.univ-mrs.fr/\~{}torresan/papers/SigPro.ps.gz

[2] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Applied and Computational Harmonic Analysis*, vol. 18, no. 2, pp. 137–166, 2005.

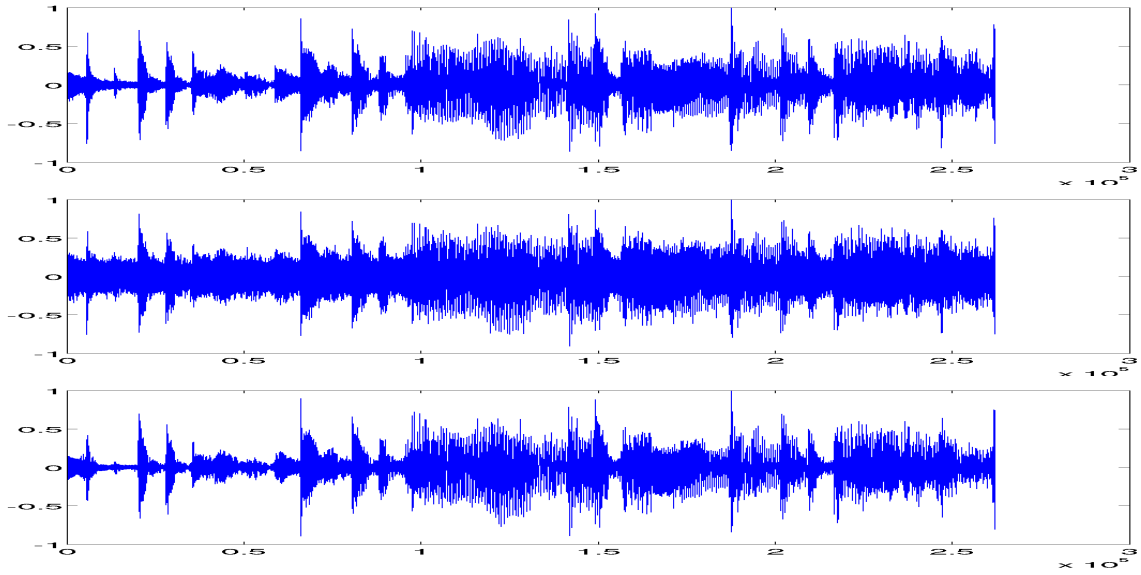

Fig. 6. Excerpt from a Susheela Raman song (Mamavatu). From top to bottom: original signal, noisy signal, restored signal.

[3] G. Teschke, "Multi-frames in thresholding iterations for nonlinear operator equations with mixed sparsity constraints," *Applied and Computational Harmonic Analysis*, vol. 22, no. 1, pp. 43–60, January 2006.

[4] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[5] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1150–1159, Sep. 2003.

[6] 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.

[7] M. Vetterli and J. Kovacević, *Wavelets and Subband Coding*, ser. Signal Processing Series. Englewood Cliffs, NJ: Prentice Hall, 1995.

[8] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani, "Sparse regression with structured priors: Application to audio denoising," in *IEEE International Conference on Acoustics, Speech, and Audio Signal*, Toulouse, France, May 2006.

[9] M. Kowalski and B. Torrésani, "A study of bernoulli and structured random audio models," in *Proceedings of the conference on Signal Processing with Adaptive and Sparse Structured Representations (SPARS'05)*, R. Gribonval, Ed., Rennes, France, November 2005, pp. 59–62.

[10] ——, "A family of random waveform models for audio coding," in *IEEE International Conference on Acoustics, Speech, and Audio Signal*, Toulouse, France, May 2006.

[11] J. A. Tropp, "Greed is good," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[12] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing ed. New York: Dover, 1964. [Online]. Available: http://www.math.sfu.ca/~cbm/aands/

[13] M. A. Stephens, "Tests based on edf statistics," in *Goodness-of-fit techniques*, R. B. D'Agostino and M. A. Stephens, Eds.

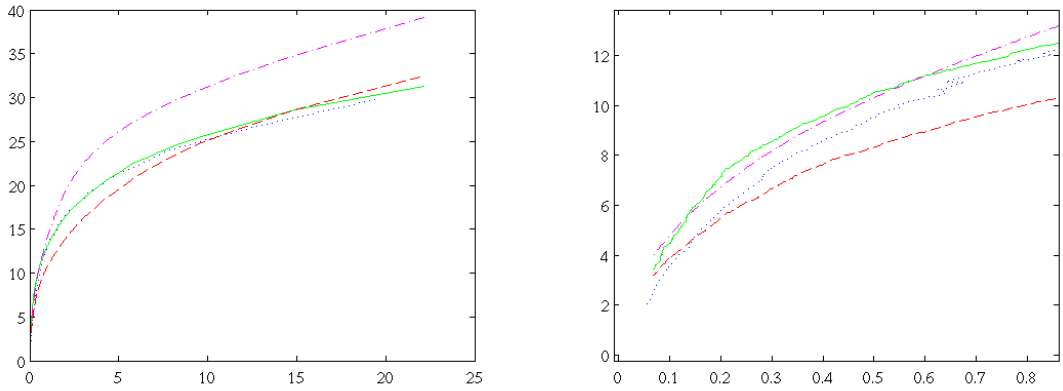

Fig. 7. SNR as a function of the percentage of non-zeros coefficients used to encode the signal. Right: the entire curve, left: zoom. Thresholding in a MDCT basis (dashdotted line), Thresholding after weighting of the coefficients in a MDCT basis (dashed line), hybrid decomposition with L^2 projection (solid line), hybrid decomposition with sparse projection (FOCUSS) (dotted line).

Marcel Dekker, Inc, 1986.

[14] S. S. Chen, D. L. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[15] B. D. Rao, E. Kjersti, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity mesure minimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[16] J. Stoer and R. Burlish, *Introduction to Numerical Analysis*. Springer-Verlag, 1991.

[17] L. Daudet, S. Molla, and B. Torrésani, "Towards a hybrid audio coder," in *International Conference Wavelet analysis and Applications*, J. P. Li, Ed., Chongqing, China, 2004, pp. 13–24.