

Random models for sparse signals expansion on unions of bases with application to audio signals

Matthieu Kowalski, Bruno Torrésani

▶ **To cite this version:**

Matthieu Kowalski, Bruno Torrésani. Random models for sparse signals expansion on unions of bases with application to audio signals. IEEE Transactions on Signal Processing, 2008, 56 (8), pp.3468-3481. 10.1109/TSP.2008.920144 . hal-00142088v2

HAL Id: hal-00142088
https://hal.science/hal-00142088v2

Submitted on 11 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Random Models for Sparse Signals Expansion on Unions of Bases with Application to Audio Signals

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. 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 insignificant 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 purposes. The proposed approach is illustrated by numerical experiments on audio signals, using MDCT bases. However, it is general enough to be applied without much modifications in different contexts, for example in image processing.

Index Terms

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

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

I. INTRODUCTION

The goal of this paper is to present a new approach to the problem of sparse signal regression in union of bases, and corresponding algorithms. 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 *dictionary* of elementary waveforms is chosen, with respect to which signal expansions are sought. The dictionary has to be complete, and can be either a basis of the underlying signal space, or a redundant system. In the latter case, the expansion of the signal is not unique, and sparsity is then used as a selection criterion. Unions of bases represent a popular choice for redundant systems, as they are often associated with efficient decomposition and synthesis algorithms.

Sparsity may be implemented in various ways. Let us quote among others greedy algorithms (see e.g. [1]), variational approaches (see for instance [2] and references therein), or Bayesian formulations (see [3], [4]), these approaches being often closely connected.

Greedy algorithms provide efficient approaches, in which few iterations are generally sufficient to yield satisfactory approximations. In such approaches it may be shown that if the dictionary is sufficiently *incoherent*, the expansion may be recovered if the signal indeed admits a sparse enough expansion.

Variational approaches often lead to thresholding or iterative thresholding strategies. We point out that these generally treat all coefficients equally, i.e. the same thresholds are applied to all of them.

Bayesian approaches have the advantage of relying on explicit signal and observation models, from which MAP (*Maximum A Posteriori*) or MMSE (*Minimum Mean Squared Error*) 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, Markov chain Monte Carlo (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 [4]. It was shown there that these are indeed extremely promising approaches, that unfortunately still require high computational power.

The approach we propose in this paper attempts to combine the advantage of probabilistic modeling and thresholding strategies. It relies on 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 (based upon adaptive thresholding) for the identification.

More precisely, we focus on a signal model of the form

$$x(t) = \sum_{i \in I} X_i c_i u_i(t) \ ,$$

where I is some generic index set, the waveforms u_i form the dictionary labeled by I, the X_i are Boolean random variables controlling the sparsity of the expansion, and the coefficients c_i 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 be identified using appropriate estimation algorithms. The conditions (further specified in Remarks 2 and 4) are mainly 1) the sparsity of the signal model (controlled by the distribution of the Boolean variables X_i), and 2) the incoherence of the dictionary (controlled by a series of weights, to be introduced in section II below).

When these conditions are fulfilled, the estimation algorithms yield good estimates for the significance maps (i.e. the subset of the index set I within which $X_i = 1$), and a corresponding subset of the dictionary. In a few words, the significance maps are obtained by adaptive thresholding of the analysis coefficients. By adaptive, we mean that the thresholds are coefficient dependent, and are estimated from the signal. In a simpler version, coefficient independent thresholds are obtained by suitable averages (which we call *mean-field* estimates).

Finally, we show that the corresponding synthesis coefficients may be estimated by regression (either standard L^2 regression, or sparse regression), involving the estimated sub-dictionary.

The main part of the analysis is performed for the case of Bernoulli significance map model, i.e. model assuming that indices of significant coefficients are iid. Specializing to more specific bases, namely MDCT bases, we also introduce a slightly more complex, *structured* model for significance maps: a Hierarchical-Bernoulli model, which aims at enforcing frequency persistence of significant coefficients with respect to a time localized MDCT basis. New significance maps estimators are introduced, that lead to adaptive thresholding strategies on fixed time groups of coefficients. Such simple structured models should be considered as a first step towards more realistic signal models, implementing more complex modeling in the coefficient domain.

The theoretical analysis is illustrated in the context of audio signal processing, using dictionaries constructed as unions of two MDCT bases. 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 allows one to encode separately the signal's components that are coherent with the different bases. 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*. As a by-product, the proposed approach also yields a decomposition of the signal into different layers, for example the tonal, transient and residual layers in [5], [6]. Our results show that the model above is generally adequate for describing audio signals, provided they do not 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.

The paper is organized as follows. The theoretical analysis of this paper is provided in Section II, where the models are specified, and the behavior of the analysis coefficients and various quantities of interest are analyzed in details. Special attention is paid to mean-field estimates, i.e. averages with respect to significance maps, which play a key role in the estimation, either by themselves, or as initializations of iterative algorithms. Corresponding algorithms are described in Section III, and in the appendix. The models and algorithms are illustrated by a number of numerical results on audio signal denoising and coding (Section IV).

II. Hybrid waveform signal models

We start with a description of the random waveform models, expressing signals as sparse random sums of basis functions. The goal of this section is to provide an analysis of the behavior of the analysis coefficients of such signals, and derive corresponding estimators for the parameters.

A. Generalities

Let $\mathcal{H}$ denote a (finite or infinite dimensional) separable real Hilbert space, and let $\mathcal{V} = \{v_n, n \in I\}$ and $\mathcal{U} = \{u_m, m \in I\}$ 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

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

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 $\alpha_n, \beta_m \in \mathbb{R}$ 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 (or 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 the numerical applications, we limit ourselves to a specific pair of orthonormal bases: $\mathcal{U}$ 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 $\mathcal{V}$ 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 [5], [6]), 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 are denoted by $\mathbb{P}_0$ and $\mathbb{E}_0$. The global probability measure and expectation will be denoted by $\mathbb{P}$ and $\mathbb{E}$ respectively.

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

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

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

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

The corresponding signal model therefore takes the form

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

The simplest possible model for the significance maps is the *Bernoulli* model: given a fixed *membership probability* $p_n = p$, $n \in I$, 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 $\tilde{p}_m = \tilde{p}$, $m \in I$. More sophisticated models for the significance maps, termed *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) read

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

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. For convenience, we introduce the reduced variances

$$\sigma_n = \sigma f_n \ , \ \tilde{\sigma}_n = \tilde{\sigma} \tilde{f}_n \ , \tag{6}$$

where $\sigma = \max_n(\sigma_n)$ and $\tilde{\sigma} = \max_n(\tilde{\sigma}_n)$, so that $f_n \leq 1$, $\tilde{f}_n \leq 1$ Later on, when specializing to time-frequency bases $\mathcal{U}$ and $\mathcal{V}$, f and $\tilde{f}$ will be fixed *frequency profiles*, that model "typical" decay of the coefficients with respect to frequency.

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, combined with MCMC optimization algorithms, 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, we 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

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

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

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_n and b_n coefficients are zero-mean normal random variables, with covariance matrices $\mathcal{C}_{mn} = \mathbb{E}_0 \{a_m a_n\}$, $\tilde{\mathcal{C}}_{mn} = \mathbb{E}_0 \{b_m b_n\}$, given by

$$\begin{aligned} \mathcal{C}_{mn} &= (\sigma_m^2 X_m + s^2)\delta_{mn} + \sum_{i \in I} \widetilde{X}_i \tilde{\sigma}_i^2 \langle v_m, u_i \rangle \langle u_i, v_n \rangle \ , \\ \tilde{\mathcal{C}}_{mn} &= (\tilde{\sigma}_m^2 \widetilde{X}_m + s^2)\delta_{mn} + \sum_{i \in I} X_i \sigma_i^2 \langle u_m, v_i \rangle \langle v_i, u_n \rangle \ . \end{aligned}$$

In particular, the diagonal terms read

$$\mathbb{E}_0 \left\{ a_n^2 \right\} = \sigma_n^2 X_n + \sum_{i \in I} \widetilde{X}_i \tilde{\sigma}_i^2 \langle v_n, u_i \rangle^2 + s^2 \ . \tag{9}$$

If the randomness of the significance maps is taken into account, 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 the random variables defined by

$$\tilde{\gamma}_n(\Delta) = \sum_{\delta \in \Delta} \tilde{f}_\delta^2 \langle v_n, u_\delta \rangle^2 \ , \ \gamma_n(\Lambda) = \sum_{\lambda \in \Lambda} f_\lambda^2 \langle u_n, v_\lambda \rangle^2 \ . \tag{10}$$

where f_λ and $\tilde{f}_\delta$ are defined in equation (6).

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

$$\tilde{p}_n(\Delta) = \sum_{\delta \in \Delta} \langle v_n, u_\delta \rangle^2 \, , \ p_n(\Lambda) = \sum_{\lambda \in \Lambda} \langle u_n, v_\lambda \rangle^2 \ .$$

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 $\tilde{p}_n(\Delta)$ 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 $\mathbf{M}$ (resp. $\widetilde{\mathbf{M}}$) the operator defined by a diagonal matrix in the $\mathcal{V}$ (resp. $\mathcal{U}$) basis

$$\mathbf{M}v_n = f_n v_n \ , \ \widetilde{\mathbf{M}}u_n = \tilde{f}_n u_n \ . \tag{11}$$

Then $\gamma_n(\Lambda)$ is the squared norm of the orthogonal projection of $\mathbf{M}u_n$ onto the linear span of the functions $\{v_\lambda, \lambda \in \Lambda\}$:

$$\gamma_n(\Lambda) = \sum_{\lambda \in \Lambda} \langle u_n, \mathbf{M}v_\lambda \rangle^2 = \sum_{\lambda \in \Lambda} \langle \mathbf{M}u_n, v_\lambda \rangle^2 \ ,$$

It follows from Parseval's formula that for all n and Λ,

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

(the last inequality results from $f_n \leq 1 \, \forall n$), and a similar expression for $\tilde{\gamma}_n(\Delta)$.

$\mathbf{M}$ is a well-defined operator in the finite dimensional case. In infinite-dimension situations, i.e. for continuous time signals, additional assumptions on the normalized variances f_n^2, $\tilde{f}_n^2$ are needed to ensure the boundedness of $\mathbf{M}$.

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

$$\mathbb{E}_0 \left\{ a_n^2 \right\} = \begin{cases} \sigma_n^2 + \tilde{\gamma}_n(\Delta)\tilde{\sigma}^2 + s^2 & \text{if } n \in \Lambda \\ \tilde{\gamma}_n(\Delta)\tilde{\sigma}^2 + s^2 & \text{if } n \notin \Lambda \end{cases} . \tag{12}$$

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_n have a significantly different behavior depending on whether X_n 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 for the γ weights may be obtained. By mean-field estimate, we mean estimates involving expectation with respect to only one of the two significance maps. For example, a_n coefficients (related to the significance map Λ) are studied, using averages with respect to Δ. These will be used for instance for initializing iterative algorithms for which an initial estimate of the significance map is needed.

For example, the first moment of the γ weights reads

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

We give below the mean-field estimates for the a coefficients (i.e. expectation with respect to $\mathbb{P}_\Delta$, which yields a Λ dependent result), similar expressions may be derived for the b coefficients.

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

$$\mathbb{E}_\Delta \left\{ \mathbb{E}_0 \left\{ a_n^2 \right\} \right\} = \sigma_n^2 X_n + \tilde{p} \|\widetilde{\mathbf{M}} v_n\|^2 \tilde{\sigma}^2 + s^2 \ .$$

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:

$$\mathbb{E}_\Delta \left\{ \mathbb{E}_0 \left\{ \frac{a_n^2}{f_n^2} \right\} \right\} = \sigma^2 X_n + \tilde{p} \frac{\|\widetilde{\mathbf{M}} v_n\|^2}{f_n^2} \tilde{\sigma}^2 + \frac{s^2}{f_n^2} \tag{14}$$

$$= \sigma^2 X_n + \tilde{p} \ \tilde{\sigma}^2 \sum_{i \in I} \frac{\tilde{f}_i^2}{f_n^2} \langle v_i, u_n \rangle^2 + \frac{s^2}{f_n^2} \ . \tag{15}$$

We notice that the distribution of the renormalized coefficients is governed by $\sum_{i \in I} \frac{\tilde{f}_i^2}{f_n^2} \langle v_i, u_n \rangle^2$.

Remark 3: This renormalization ensures that the leading term $\sigma^2 X_n$ in (14) has a constant variance σ^2. The variance of the second term varies as a function of n. In order that such a dependence be weak, constraints on the bases and normalized variances will have to be imposed, as we shall see below in section II-E.

D. Significance maps estimation for 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

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

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

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

hence

$$\text{Var}\{\gamma_n(\Lambda)\} = p(1-p)\sum_m \langle u_n, \mathbf{M}v_m\rangle^4 \ . \tag{17}$$

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

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

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. Indeed, in such situations, the distribution of analysis coefficients is close to a mixture of two Gaussians, with significantly different variances.

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 $B_4 = \sup_n \sum_m \langle u_n, \mathbf{M}v_m\rangle^4$, 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

$$w_{n;0}^2 = \tilde{\gamma}_n(\Delta)\tilde{\sigma}^2 + s^2 \ , \ w_{n;1}^2 = w_{n;0}^2 + \sigma_n^2 \ . \tag{20}$$

Then the intersection point τ_n of the two Gaussians $\mathcal{N}(0, w_{n;0}^2)$ and $\mathcal{N}(0, w_{n;1}^2)$, is given by

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

and one can immediately obtain

Proposition 3: 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:

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

In addition, the simplicity of the model allows one to compute the error probabilities, which read as follows

$$\mathbb{P}\{Y_n = 0 | X_n = 1\} = p \ \text{erf} \left(\sqrt{\frac{w_{n;0}^2}{\sigma_n^2} \ln \left[\frac{p}{1-p} \left(1 + \frac{\sigma_n^2}{w_{n;0}^2} \right) \right]} \right) \ ,$$

$$\mathbb{P}\{Y_n = 1 | X_n = 0\} = (1-p) \ \text{erfc} \left(\sqrt{\left(1 + \frac{w_{n;0}^2}{\sigma_n^2} \right) \ln \left[\frac{1-p}{p} \left(1 + \frac{\sigma_n^2}{w_{n;0}^2} \right) \right]} \right) \ ,$$

where erfc is the complementary error function [12].

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

Remark 6: The same analysis may be done starting from mean-field estimates. In such a case, we have seen in Remark 4 that the distribution of analysis coefficients may be approximated by a mixture of a small number of Gaussians. This yields a global threshold (with the γ coefficients replaced by their average) instead of coefficient-dependent thresholds.

E. Time-frequency bases for audio signals

The above discussion applies to arbitrary bases of $\mathcal{H}$. Let us now briefly describe the situation we shall consider below. The two bases $\mathcal{U}$ and $\mathcal{V}$ will be MDCT bases, and the index set I (as well as the significance maps) will be twofold: time and frequency indices. The first MDCT basis $\mathcal{U}$ is constructed from a wide window function, i.e. consists of frequency localized atoms, while the basis $\mathcal{V}$, built from a

narrow window, consists of time localized atoms. In such a context, the partial reconstruction $\sum_{\delta\in\Delta} b_\delta u_\delta$ (resp. $\sum_{\lambda\in\Lambda} a_\lambda v_\lambda$) will be called the *tonal layer* (resp. the *transient layer*) of the signal.

In the numerical applications to audio signals below, we shall limit ourselves to situations where the normalized variances $f^2_{k,\nu}$ and $\tilde{f}^2_{k,\nu}$ depend on the frequency part ν only, and model "typical" frequency decay of coefficients.

Coming back to remark 3, we assume that the basis functions u and v under consideration are sufficiently well localized in the frequency domain, so that only a few terms of the matrix of scalar products $\langle u_n, v_m\rangle$ are non-negligible. If in addition the frequency profiles f_n and $\tilde{f}_n$ vary slowly as functions of the frequency index, the dependence with respect to n in (15) may be considered weak in first approximation.

F. Structured models

In this section, we again limit ourselves to time-frequency bases, as described above.

1) Generalities: Unlike the Bernoulli model, structured significance maps models involve correlations between significance map elements. For example, assuming $\mathcal{U}$ 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 $\mathbb{E}_\Lambda\{\mathcal{C}_{mn}\}$ nor $\mathbb{E}_\Delta\left\{\tilde{\mathcal{C}}_{mn}\right\}$ involve the correlation functions of the significance maps $\mathbb{E}_\Lambda\{X_m X_n\}$ or $\mathbb{E}_\Delta\left\{\widetilde{X}_m\widetilde{X}_n\right\}$.

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 $n = (k_n, \nu_n) \in \Lambda = \Lambda_t \times \Lambda_f$ 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

$$\widetilde{X}_n \sim \mathcal{B}(\tilde{p}) \ ; \ X_n \sim T_{k_n}\mathcal{B}(p_2) + (1 - T_{k_n})\delta_0 \ , \text{ with } \ T_{k_n} \sim \mathcal{B}(p_1) \ . \tag{21}$$

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 (with $w_{k,\nu;0}$ and $w_{k,\nu;1}$ as defined in (20)).

$$c_k = \sum_{\nu=1}^{|\Lambda_f|} \frac{a_{k,\nu}^2}{w_{k,\nu;0}^2} = \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;0}} X_{k,\nu} + \tilde{\pi}_{k,\nu} \right]^2 \ , \tag{22}$$

where we have set for simplicity $\tilde{\pi}_{k,\nu} = \frac{1}{w_{k\nu;0}} \sum_{\delta \in \Delta} \beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu}$. We rewrite (22) to separate the coefficients c_k with $k \in \Lambda_t$ from the others

$$c_k = \begin{cases} \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;0}} X_{k,\nu} + \tilde{\pi}_{k,\nu}(\Delta) \right]^2 & \text{if } k \in \Lambda_t \\ \sum_{\nu=1}^{|\Lambda_f|} \tilde{\pi}_{k,\nu}(\Delta)^2 & \text{if } k \notin \Lambda_t \end{cases} . \tag{23}$$

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

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

In case $k \notin \Lambda_t$, 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:

$$\begin{aligned} c'_k &= \sum_{\nu=1}^{|\Lambda_f|} \frac{a_{k,\nu}^2}{f_{k,\nu}^2} \\ &= \begin{cases} \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{f_{k,\nu}} X_{k,\nu} + \tilde{\pi}'_{k,\nu}(\Delta) \right]^2 & \text{if } k \in \Lambda_t \\ \sum_{\nu=1}^{|\Lambda_f|} \tilde{\pi}'_{k,\nu}(\Delta)^2 & \text{if } k \notin \Lambda_t \end{cases} , \end{aligned} \tag{24}$$

with $\tilde{\pi}'_{k,\nu} = \sum_{\delta \in \Delta} \frac{\beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu}}{f_{k,\nu}} \sim \mathcal{N}\left(0, \frac{w_{k,\nu;0}^2}{f_{k,\nu}^2}\right)$.

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 sought 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 7: Alternatively, we may also normalize all the coefficients c_k according to their membership to the submaps Λ_t:

$$\begin{aligned} d_k &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{\alpha_{k,\nu}}{w_{k,\nu;1}} X_{k,\nu} + \frac{1}{w_{k,\nu;0}} \sum_{\delta \in \Delta} \beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu} \right]^2 \text{ if } k \in \Lambda_t \ , \\ d'_k &= \sum_{\nu=1}^{|\Lambda_f|} \left[\frac{1}{w_{k,\nu;0}} \sum_{\delta \in \Delta} \beta_\delta \langle u_\delta, v_{k,\nu} \rangle + \rho_{k,\nu} \right]^2 \text{ if } k \notin \Lambda_t \ . \end{aligned} \tag{25}$$

In this case, we 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

$$Z_k = \begin{cases} 1 & \text{if } p\chi^2(d_k) > (1-p)\chi^2(d'_k) \\ 0 & \text{otherwise} \end{cases} \ . \tag{26}$$

G. 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_n (resp. $\tilde{f}_n$) the corresponding frequency profiles. Let

$$\theta_1 = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \frac{a_\lambda^2}{f_\lambda^2} \ , \ \theta_2 = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \frac{b_\delta^2}{\tilde{f}_\delta^2} \ .$$

Then, the estimates defined by

$$\hat{\sigma}^2 = \frac{\theta_1 - \epsilon_1 \theta_2}{1 - \epsilon_1 \epsilon_2} \ , \ \hat{\tilde{\sigma}}^2 = \frac{\theta_2 - \epsilon_2 \theta_1}{1 - \epsilon_1 \epsilon_2} \tag{27}$$

with $\epsilon_1 = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \frac{\tilde{\gamma}_\lambda(\Delta)}{f_\lambda^2}$ and $\epsilon_2 = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \frac{\gamma_\delta(\Lambda)}{\tilde{f}_\delta^2}$, are convergent and unbiased.

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

$$\mathbb{E}\left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \right\} = \begin{pmatrix} 1 & \epsilon_1 \\ \epsilon_2 & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} \ .$$

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

$$\mathrm{var}\{\hat{\sigma}\} = \frac{\mathrm{var}\{\theta_1\} + \epsilon_1^2 \mathrm{var}\{\theta_2\} - 2\epsilon_1 \mathrm{cov}\{\theta_1, \theta_2\}}{(1-\epsilon_1\epsilon_2)^2} \ .$$

θ_1 (resp. θ_2) is the classical estimator for the expectation of $\frac{a_\lambda^2}{f_\lambda^2},\ \lambda \in \Lambda$ (resp. $\frac{b_\delta^2}{\tilde{f}_\delta^2},\ \delta \in \Delta$). We have then

$$\mathrm{var}\{\theta_1\} \underset{N\to\infty}{\to} 0 \ , \ \mathrm{var}\{\theta_2\} \underset{N\to\infty}{\to} 0 \ .$$

We just have to prove that the covariance $\mathrm{cov}\{\theta_1, \theta_2\} \underset{N\to\infty}{\to} 0$. For this, we need to compute for $\lambda \in \Lambda$ and $\delta \in \Delta$

$$\mathbb{E}\left\{\frac{a_\lambda^2}{f_\lambda^2}\frac{b_\delta^2}{\tilde{f}_\delta^2}\right\} = \sigma^2\tilde{\sigma}^2 + \sigma^2\tilde{\sigma}^2\frac{\gamma_\lambda(\Delta)}{f_\lambda^2}\frac{\tilde{\gamma}_\delta(\Lambda)}{\tilde{f}_\delta^2} + \sigma^4\frac{\langle u_\delta, v_\lambda\rangle^2}{\tilde{f}_\delta^2} + \tilde{\sigma}^4\frac{\langle u_\delta, v_\lambda\rangle^2}{f_\lambda^2} + 4\sigma^2\tilde{\sigma}^2\langle u_\delta, v_\lambda\rangle^2 \ .$$

Then,

$$\begin{aligned}
\mathrm{cov}\{\theta_1, \theta_2\} &= \mathbb{E}\{\theta_1\theta_2\} - \mathbb{E}\{\theta_1\}\mathbb{E}\{\theta_2\} \\
&= \sigma^2\tilde{\sigma}^2 + \sigma^2\tilde{\sigma}^2\epsilon_1\epsilon_2 + \sigma^4\epsilon_2 + \tilde{\sigma}^4\epsilon_1 + \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{\lambda\in\Lambda}\sum_{\delta\in\Delta}\langle u_\delta, v_\lambda\rangle^2 - (\sigma^2 + \epsilon_1\tilde{\sigma}^2)(\tilde{\sigma}^2 + \epsilon_2\sigma^2) \\
&= \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{\lambda\in\Lambda}\sum_{\delta\in\Delta}\langle u_\delta, v_\lambda\rangle^2 \leq \frac{4\sigma^2\tilde{\sigma}^2}{|\Lambda||\Delta|}\sum_{\lambda\in\Lambda}\sum_{\delta=1}^{N}\langle u_\delta, v_\lambda\rangle^2 = \frac{4\sigma^2\tilde{\sigma}^2}{|\Delta|} \underset{N\to\infty}{\to} 0 \ ,
\end{aligned} \tag{28}$$

which concludes the proof. ■

H. 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 detail the algorithms deduced from the analysis of the model in section II. We first focus on significance maps estimation, and describe iterative adaptive thresholding algorithms, and mean-field algorithms which will be used as initialization. We then turn to synthesis coefficient estimation, for which again two methods are proposed. 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 yield 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-D, and stressed in remark 4 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 Expectation-Maximization (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 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 $s^2 \gg \tilde{\sigma}^2\tilde{\gamma}(\Delta)^2$, 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 Remark 6, the classification of analysis coefficients is equivalent to an adaptive thresholding: analysis coefficients whose magnitude is larger than a threshold are assigned to the significance map, however the threshold is coefficient dependent.

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 the renormalized analysis coefficients $\frac{a_n}{f_n}$ (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 0.001% 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 (24) are used to obtain an estimate of the transient submap Λ_t, through a statistical test. We showed in section II-F.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

$$H_0 : \{c'_1, \ldots, c'_{|\Lambda_t|}\} \text{ follow a } \chi^2 \text{ law}$$
$$H_1 : \{c'_1, \ldots, c'_{|\Lambda_t|}\} \text{ do not follow a } \chi^2 \text{ law}$$

This test is used in a classification algorithm which proceeds as follows. 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. This 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 as 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 rely 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-G.

These estimates can be exploited in a Classification Expectation-Maximization (`CEM`) [14] algorithm which uses the MAP estimate for X_n and $\widetilde{X}_n$ described in proposition 3 in section II-D. Some generalities on the `CEM` algorithm are given at the end of appendix A. The algorithm can be summarized as follows. After an initialization for the significance maps Λ and Δ and the parameters p, $\tilde{p}$, σ and $\tilde{\sigma}$, the following four 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 use 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 7, section II-F.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 $\mathcal{U} \cup \mathcal{V}$, 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 $x_\mathcal{U}$ and $x_\mathcal{V}$, is to compute an orthogonal projection of the signal x onto $\mathcal{H}_{\hat{\mathcal{D}}}$

$$\hat{x} = \underset{y \in \mathcal{H}_{\hat{\mathcal{D}}}}{\operatorname{argmin}} = \|x - y\|^2 \ . \tag{29}$$

One can write

$$\hat{x} = \hat{x}_\mathcal{V} + \hat{x}_\mathcal{U} \ , \tag{30}$$

with

$$\hat{x}_\mathcal{V} = \sum_{\lambda \in \hat{\Lambda}} \hat{\alpha}_\lambda v_\lambda \ , \ \hat{x}_\mathcal{U} = \sum_{\delta \in \hat{\Delta}} \hat{\beta}_\delta u_\delta \ . \tag{31}$$

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

$$\boldsymbol{G} \left(\hat{\alpha}_1, ..., \hat{\alpha}_{|\hat{\Lambda}|}, \hat{\beta}_1, ..., \hat{\beta}_{|\hat{\Delta}|} \right)^T = \left(a_1, ..., a_{|\hat{\Lambda}|}, b_1, ..., b_{|\hat{\Delta}|} \right)^T \ , \tag{32}$$

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 $\mathcal{H}_{\hat{\mathcal{D}}}$ which is always true here.

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

$$\hat{x} = \underset{y \in \mathcal{H}_{\hat{\mathcal{D}}}}{\operatorname{argmin}} \|x - y\|_2^2 + \lambda \|y\|_1 \ , \tag{33}$$

where λ is tuning parameter which acts on sparsity. Following Chen and Donoho in [15] 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 [16].

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)

$$\hat{x} = \underset{y}{\operatorname{argmin}}\, \mathbb{E}_0 \left\{ \|x - y\|^2 \right\} , \tag{34}$$

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

$$\hat{\alpha} = t_\lambda \alpha \ , \ \hat{\beta} = t_\delta \beta \ . \tag{35}$$

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

$$\hat{\alpha}_\lambda = \frac{\sigma^2}{\sigma^2 + \gamma_\lambda(\Delta)\tilde{\sigma}^2 + s^2}\, a_\lambda \ , \ \hat{\beta}_\delta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \gamma_\delta(\Lambda)\sigma^2 + s^2}\, b_\delta \ . \tag{36}$$

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 and much faster to compute.

IV. Numerical results

The model and corresponding algorithms were tested on various "classical" applications: signal separation into two layers for single sensor blind source separation, denoising and audio coding. In each case, we only report here on the results obtained with the version of the algorithm that is best adapted to the problem under consideration.

Regarding the applications to audio signal, it could be useful to specify some specific terms used here after. An audio signal can be modeled as a superposition of three layers $tonal + transient + noise$ [5], [6], [17]. The transient layer appears to be well defined in time, like the percussive sounds. The tonal layer corresponds to the musical notes, and leaves in the frequency domain. The presented model seems adapted to represent these different features.

Except for the audio coding application, we used the mean-field algorithm corresponding to the Bernoulli-model in order to estimate the tonal layer. The transient layer was estimated by the mean-field algorithm corresponding to the Hierarchical-Bernoulli model. As mentioned in the section III, the two significance maps are estimated independently of each other. Once the significance maps are estimated, we chose to use the L^2 regression to estimate the coefficients. Although this method is slower than

Wiener-type estimate, it ensures better estimates. If one prefer to enforce the speed of the algorithm rather than the quality of the results, the Wiener-type estimate are more suitable.

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

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 the website [18].

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 [5], [6], [17]. This problem may be seen as a single sensor blind source separation problem, the sources being the three layers: transient, tonal and noise. The single sensor blind source separation (BSS) is a very difficult task, and the usual BSS techniques, such as independent component analysis (ICA) [19], require at least two sensors for estimating two sources.

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 website [18].

We chose three Gaussians to approximate the distribution of the tonal layer with the Bernoulli model. For the transient layer, after the selection in time by the Hierarchical-Bernoulli model, the same number of Gaussian was chosen for the selection in the frequency domain. The algorithm is applied three times as suggested in section III-A.3. The size of the window for the tonal MDCT basis is 8192 samples (about

[1]Even though the corresponding time length, about 3 ms, 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.

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 corresponds 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.

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).

Gaussian white noise was added to different types of signals, so as to obtain a 6 dB SNR. All signals have 2^{17} samples (about 3 sec of sound). We follow the strategy proposed in the section III-A.1. We first try to approximate the distribution of the tonal part with three Gaussians and, if the estimated map seems to be insignificant, we used two Gaussians for the approximation. As the noise variance is expected to be large, we work at the same time on the renormalized coefficients, and the original coefficients. The same strategy is used to estimate the transient map after the preselection in the time domain by the Hierarchical-Bernoulli model.

The algorithm was first tested on the xylophone signal which is perfectly designed for: the xylophone has percussive attacks, and present a significant tonal layer. Results are displayed in figure 2. The reconstruction $tonal+transient$ provides a denoised signal which doubles the SNR: 12 dB are obtained. The tonal layer alone provide a 10 dB SNR. Although the transient layer does not contain the main information, the recovery of attacks improves appreciably the final result.

To assess the quality of the results, we compared them with results obtained with three `MCMC` algorithms provided in [8]. Gaussian white noise was added to the piano signal, so as to yield 10 dB input SNR. Numerical results are displayed in figure 3

The three `MCMC` algorithms, correspond to the following models:

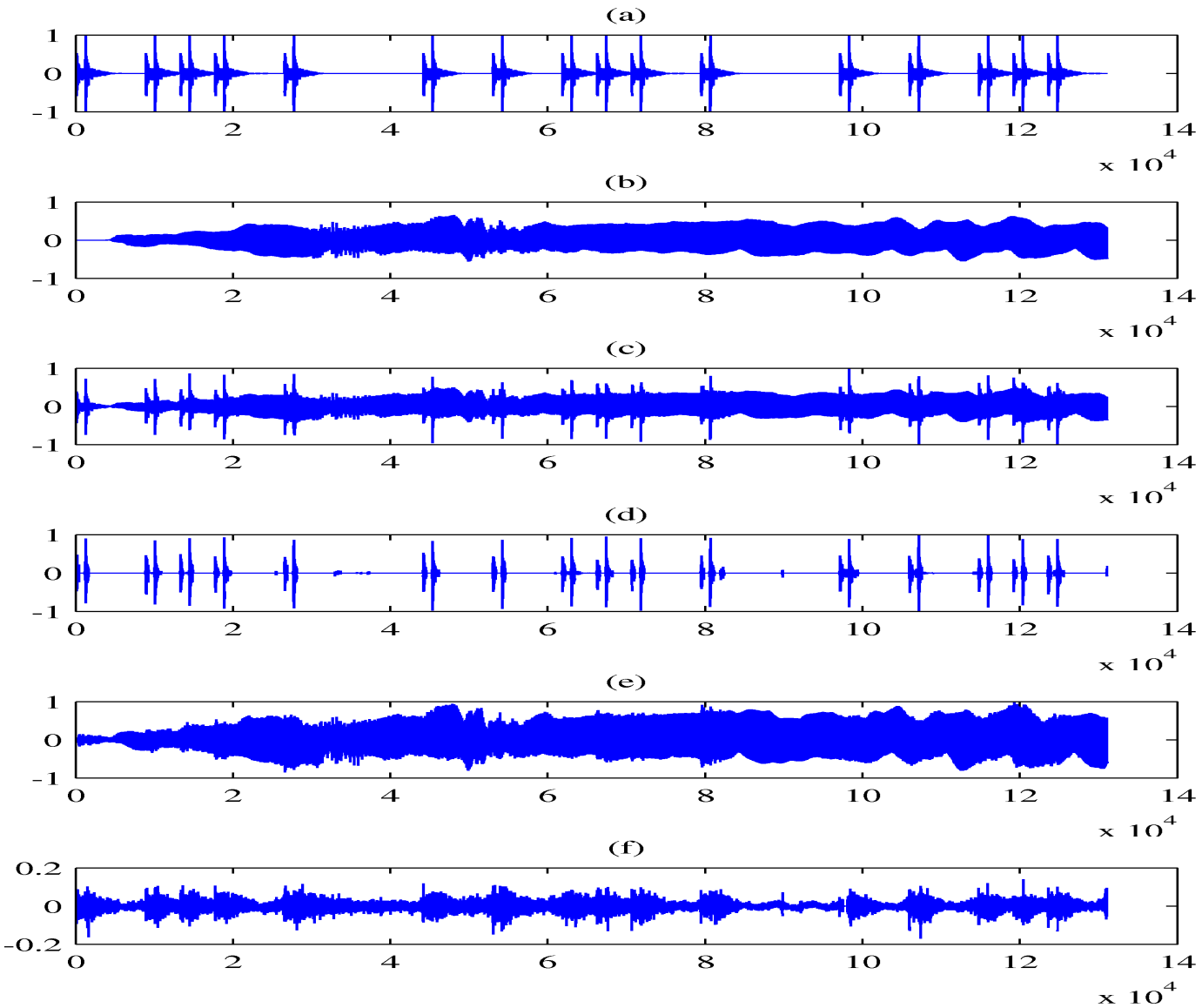


Fig. 1. Blind source separation of trumpet + castanets mixture. From top to bottom: (a) original castanets signal, (b) original trumpet signal, (c) mixture, (d) estimate of the castanets signal, (e) estimate of the trumpet signal, (f) residual.

`MCMC1` The tonal layer and the transient layer are modeled with a Bernoulli model.

`MCMC2` The tonal layer is modeled with a Markov chain and the transient layer with a Bernoulli model.

`MCMC3` The tonal and transient layers are both modeled with a Markov chain.

The Markov chain models allow one to preserve some "lines" structures as argued in [6].

In addition to these three `MCMC` algorithms, [8] gives the results obtained with an `EM` algorithm with Jeffrey's prior. This last algorithm is closer to ours.

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

TABLE I

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

The comparison is provided in table I (our algorithm is called *Hybrid Algorithm*). 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. The latter produce a lot of artifacts and "musical" noise, but our algorithm looses a little bit more high frequencies. Compared to the "Jeffrey's + EM" algorithm, our results are indisputably better, in term of SNR or by listening the signals.

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 30mn for the Bayesian+MCMC approaches. The inversion of the Gram matrix is by far the most costly operation, Wiener estimates yielding a much faster algorithm.

On the website [18], one can listen the results obtain on different type of noisy signal. One of these, the panpipe signal, is especially interesting. The distinguishing feature of this signal is the presence of a blow, which 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.

The algorithm may seem slow compared to classical denoised algorithm, like soft thresholding [20]. But our algorithm does not require a lot of tuning parameters: only the choice of the two bases is really important, and this choice may be intuitively easy to do, contrary to the choice of the threshold. Furthermore, our algorithm takes into account the different layers present in the signal.

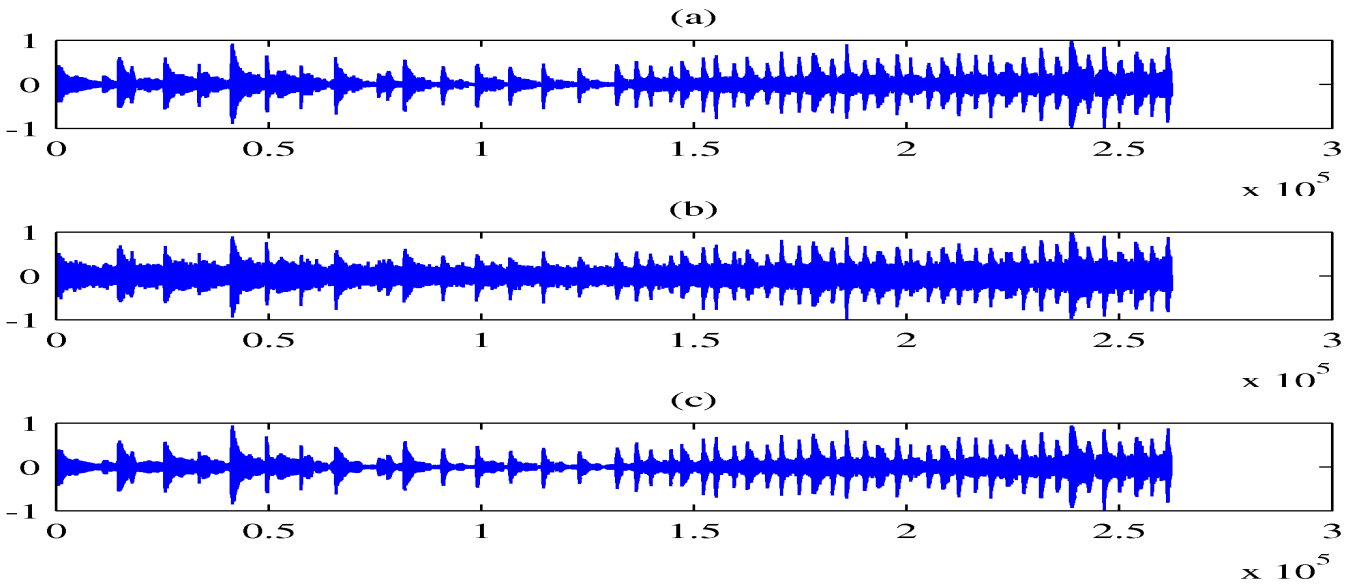


Fig. 2. Xylophone signal. From top to bottom: (a) original signal, (b) noisy signal, (c) restored signal.

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 a different value of the noise variance s 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 normalized variances $f_{k,\nu}$. 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. Figure 4 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 4 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.

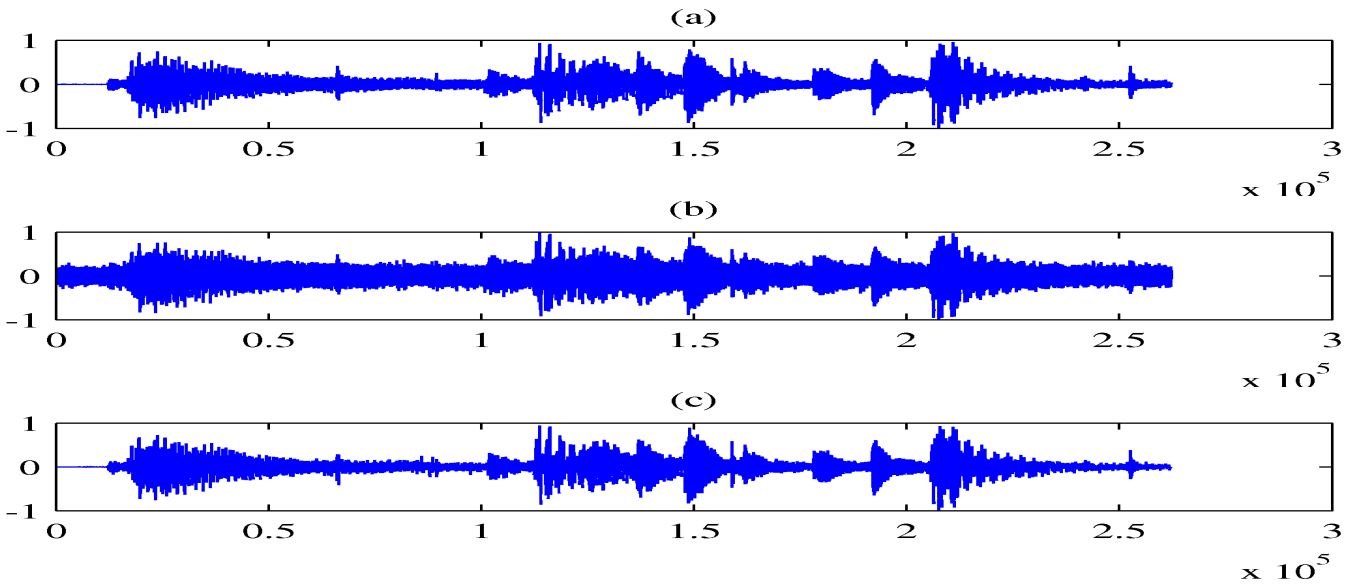


Fig. 3. Excerpt from a Norah Jones song. From top to bottom: (a) original signal, (b) noisy signal, (c) restored signal.

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 to obtain estimates for the significance maps (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, mean-field algorithms are the most efficient, in term of expected results, and computing time.

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

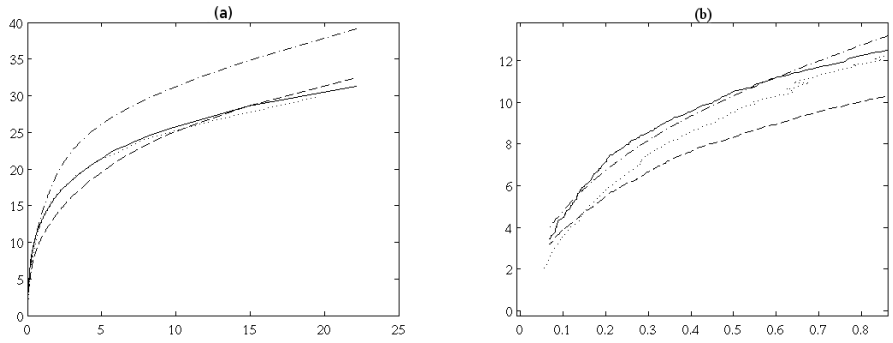


Fig. 4. SNR as a function of the percentage of non-zeros coefficients used to encode the signal. (a) the entire curve, (b) zoom. Thresholding in a MDCT basis (dashdotted line), Thresholding after weighting of the coefficients in a MDCT basis (dashed line), hybrid decomposition with L^2 projection (solid line), hybrid decomposition with sparse projection (FOCUSS) (dotted line).

structures for the maps and between maps, in view of audio applications. A first point will be to extend the theoretical analysis to more complex significance maps models, such as Markov models considered in [4], [6]. A second improvement would amount to implement in the model an explicit dependence between the two significance maps. Indeed, although natural from the theoretical and algorithmical points of view, the assumption of independence of the tonal and transient layers is an important simplification: 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 require much more complex estimation algorithms. The question is, would such added complexity lead to significant improvement in the results? In our opinion, the approaches developed in this paper represent a good compromise between realism for signal models and simplicity for the estimation algorithms. Finally, it should also be pointed out that similar ideas could also be exploited in different contexts, for example for "contour+texture" separation in images, or more generally for multichannel signals (which would require vector valued significance maps).

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 $\mathcal{L}(\Theta) = \mathbb{P}(X|\Theta) = \prod_i^n f(x_i|\Theta)$, where f denotes the pdf.

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: $\tilde{x}_i = \phi_k(x_i) \quad \text{if } x_i \in \mathcal{C}_k$, 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 $\log \mathcal{L}(\tilde{Y}|\Theta) = \log(P(\tilde{X}, Z|\Theta))$ reads

$$\log \mathcal{L}(\tilde{Y}|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{i,k} \log(\pi_k f(\phi_k(x_i)|\theta_k)) \ . \tag{37}$$

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. The expectation state $Q(\Theta|\hat{\Theta}^{(t)}) = \mathbb{E}\{\mathcal{L}(\Theta)|\tilde{X}, \hat{\Theta}^{(t)}\}$ at the iteration i is:

$$Q(\Theta|\hat{\Theta}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{c} \mathbb{E}\{z_{i,k}|\tilde{X}, \hat{\Theta}^{(t)}\} \log(\pi_k f(\phi_k(x_i)|\theta_k)) \ ,$$

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

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

The maximization state is classically obtained by solving the likelihood equations, depending of the mixture model. For the "*Classification*" approach, the $\hat{z}_{i,k}$ are classified by MAP, and the maximization state is done directly by estimation on the classified data.

B. Miscellaneous details on implementation

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 8: 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

$$\langle u_m, v_n\rangle = \langle u_{m+r}, v_{n+kr}\rangle \ . \tag{38}$$

This remark allows us to compute the Gram matrix in reasonable time. Furthermore, the fact that the $\mathcal{U}$ and $\mathcal{V}$ 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 [21]).

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.

REFERENCES

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

[2] G. Teschke, "Multi-frames representations in linear inverse problems with mixed multi-constraints," *Applied and Computational Harmonic Analysis*, vol. 22, no. 1, pp. 43–60, January 2006.

[3] 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.

[4] 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.

[5] 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

[6] 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.

[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. Marcel Dekker, Inc, 1986.

[14] G. Govaert and G. Celeux, "A classification EM algorithm for clustering and two stochastic versions," *Computational Statistics and Data Analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[15] 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.

[16] 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.

[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.

[18] [Online]. Available: http://www.cmi.univ-mrs.fr/~kowalski/IEEE07.html

[19] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994, special issue on Higher-Order Statistics.

[20] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

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