A family of random waveform models for audio coding
Matthieu Kowalski, Bruno Torrésani

To cite this version:
Matthieu Kowalski, Bruno Torrésani. A family of random waveform models for audio coding. ICASSP 2006, May 2006, Toulouse, France. pp.11024, 10.1109/ICASSP.2006.1660693 . hal-00347439

HAL Id: hal-00347439
https://hal.science/hal-00347439
Submitted on 15 Dec 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.
A FAMILY OF RANDOM WAVEFORM MODELS FOR AUDIO CODING

M. Kowalski and B. Torrésani

LATP, CMI, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France
email : kowalski@cmi.univ-mrs.fr; Bruno.Torresani@cmi.univ-mrs.fr.

ABSTRACT

We study the behavior of hybrid random waveform models for audio signals, involving sparse random series of waveforms, with random coefficients. Similar approaches have been considered in the recent years. However, these do generally not rely on explicit models, and are of more “algorithmical” nature. The models we propose allow us to analyze mathematical properties of such signals and corresponding estimators, and derive estimation algorithms, which do not rely on complex optimization techniques.

1. INTRODUCTION

Recently, models for audio signal involving decompositions

signal = tonal + transient + residual

have been proposed [1, 2], together with algorithms for estimating the corresponding components, with various applications. Often, the models were not specified explicitly, the proposed methods being more algorithmic in nature (see for example [3, 4, 5, 6]).

We study here more specific models, based on expansions on waveform systems. The main idea is to start with an orthonormal basis (or a frame, see [7] for the definition) of waveforms with respect to which a given component (tonal, or transient) is assumed to admit sparse expansions. Starting from an explicit model for such expansions yields estimates for the behavior of the observed coefficients (i.e. the coefficients computed from the realization of the signal) with respect to these bases (or frames), which we use to derive corresponding estimation algorithms.

We work here in a finite N-dimensional setting, and denote by $I_N$ a corresponding index set. We denote by $U$ and $Ψ$ two orthonormal bases of $C^N$, and consider the dictionary constructed as the union $U ∪ Ψ$, on which sparse signal expansions are to be sought. The models under study are based on the following ingredients :

• Significance maps : two (small) subsets $Λ$ and $∆$ of the index set $I_N$, modelled as random sets.

• Coefficients : to each $δ ∈ ∆$ (resp. $λ ∈ Λ$) is associated a random variable $β_δ$ (resp. $α_λ$).

The random waveform model associated with these ingredients takes the form of a linear combination of the (random) elements of $Ψ$ and $U$ labelled by the significance maps $Λ$ and $∆$ respectively, with (random) coefficients $β_δ$ and $α_λ$. Given such a signal model, we address the following main problems : from one (or several) realization(s) of the signal, assuming sparsity (i.e. the significance maps are small sets) and dictionary incoherence (i.e. the two bases are “significantly different”),

• Estimate the parameters of the model (variances of coefficients, distribution of the significance maps,...)

• Estimate the significance maps $Λ$ and $∆$.

• Estimate the two layers : tonal $x_Λ = ∑_{δ ∈ Λ} β_δ u_δ$ and transient $x_Ψ = ∑_{λ ∈ Λ} α_λ ψ_λ$.

Several approaches may be developed to solve such problems, including Bayesian MCMC type algorithms [8]. We stick here to simpler approaches, based on the behavior of observed coefficients, which present the advantage of lower computational load. The distribution of observed coefficients turns out to be understandable in a fairly simple way, which justifies the algorithmic approach we propose. In a few words, when the layers are sparse enough in the considered bases, and if the latter are different enough from each others, the pdf of the coefficients of the signal with respect to the $U$ and $Ψ$ bases may be approximated as a mixture of two (or more) gaussian pdfs. Then, standard procedures may be developed for the estimation of parameters and significance maps, and therefore for the layers. Our approach differs from that of [1, 2] in that it permits the simultaneous estimation of both layers, rather than a sequential estimation.

The random waveform model and the corresponding estimation problem are the main aspects that we discuss here. The random waveform model is presented in Section 2, where the behavior of observed coefficients is also studied. The estimation algorithms and corresponding numerical results are given in Section 3, and Section 4 is devoted to an outline of possible extensions of the model, and applications.

2. HYBRID WAVEFORM MODELS

2.1. Generalities

Let $ℋ$ denote a finite dimensional Hilbert space of dimension $N$, let $I = {1, ..., N}$ and let $Ψ = \{ψ_n, n ∈ I\}$ and $U = \{u_n, n ∈ I\}$ be two orthonormal bases of $ℋ$. We denote by $D = Ψ ∪ U$ the dictionary made as the union of these two bases. $D$ is clearly (over)complete in $ℋ$, and any $x ∈ ℋ$ admits an infinite number of expansions in the form $x = ∑_{n=1}^N α_n ψ_n + ∑_{m=1}^N β_m u_m$, where $α_n, β_m ∈ C$. We are interested in sparse signals, i.e. signals $x ∈ ℋ$ that may be written as

$$x = ∑_{λ ∈ Λ} α_λ ψ_λ + ∑_{δ ∈ ∆} β_δ u_δ + r,$$

where $Λ, ∆$ are small subsets of the index set $\{1, 2, ..., N\}$, and $r ∈ ℋ$ 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 expansion (1), and one must rely to additional arguments to impose sparsity constraints. Greedy algorithms provide a way to
enforce such constraints. In such situations, it may be shown that if the two bases $\Psi$ and $U$ are sufficiently incoherent, the expansion may be recovered if the sets $\Lambda$ and $\Delta$ are small enough.

In the numerical examples presented in this work, we limit ourselves to two different pairs of orthonormal bases : $U$ will be a local trigonometric basis (tuned in such a way to achieve good frequency resolution), and $\Psi$ will be either a wavelet basis, or a local cosine basis with good time resolution. In both cases, the index sets will actually be two-dimensional (a time index and a frequency index), and we shall write them as such when necessary. Other choices for the bases are possible, as well as extensions to frames (not considered here).

2.2. Random hybrid models

We now introduce an explicit model for the sparse signal in (1). The ingredients of such models are essentially twofold : a model for the sets $\Lambda$ and $\Delta$, hereafter termed significance maps, and, given these significance maps $\Lambda$ and $\Delta$, a model for the coefficients $\alpha, \lambda \in \Lambda$ and $\beta, \delta \in \Delta$.

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

i. A discrete probability model for the significance maps. The corresponding probability measures for the (random) maps $\Lambda$ and $\Delta$ will be denoted by $P_{\Lambda}$ and $P_{\Delta}$, and the expectations by $E_{\Lambda}$ and $E_{\Delta}$.

ii. A probabilistic model for the coefficients $\alpha, \lambda \in \Lambda$ and $\beta, \delta \in \Delta$, conditional to the significance maps. The corresponding probability measure and expectation will be denoted by $P_{\alpha}$ and $E_{\alpha}$.

The simplest possible model for the significance maps is the Bernoulli model : given a fixed membership probability $p$, the index values $n \in \{1, \ldots, N\}$ are iid. $\mathcal{N}(0, \sigma^2)$ random variables, in other words their pdf (conditional to $\Lambda$) reads

$$P_{\alpha_n}(z|\Lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2\sigma^2}.$$  

The residual will be modeled here as a Gaussian white noise, with variance $\sigma^2$.

**Remark 1** Since $n$ is actually a time-frequency index, an obvious (and fairly realistic) modification consists in taking frequency dependent values for the variances of $\alpha$ and $\beta$ coefficients, and coloured stationary noise for the residual.

2.3. Behavior of the observed coefficients

In practice, the only available observation is the signal $x$. We choose to exploit the observed coefficients

$$a_n = \langle x, \psi_n \rangle, \quad b_n = \langle x, u_n \rangle$$  

for estimating a sparse expansion. As a first step, let us start by studying the distribution of these observed coefficients, conditional to the significance maps. For this, we introduce the indicator random variables $X_n^\Lambda = 1$ if $n \in \Lambda$ and 0 otherwise, and similarly for $X_n^\Delta$. Then,

$$a_n = \langle x, \psi_n \rangle = \alpha_n X_n^\Lambda + \sum_{m=1}^{N} \beta_m X_m^\Delta \langle u_m, \psi_n \rangle$$  

and one can state

**Theorem 1** Conditional to the significance maps, the $a$ and $b$ coefficients are zero mean normal random variables, with covariance $C_{k\ell}^\Lambda = E_0 \{a_k \sigma_l \}$, $C_{k\ell}^\Delta = E_0 \{b_k b_l \}$.

In particular, the diagonal terms read

$$E_0 \{|a_k|^2\} = \sigma^2 X_k^\Lambda + \delta^2 \sum_{\delta \in \Delta} |\langle \psi_k, u_\delta \rangle|^2 + \sigma_0^2 .$$  

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 Theorem 1. Focusing on the diagonal terms of the covariance matrix, let us introduce the following quantities

**Definition 2** Let $\Delta$ and $\Lambda$ be two subsets of the index set $I_N$. For $n \in I_N$, the Parseval weights are defined by

$$p_n(\Delta) = \sum_{\delta \in \Delta} |\langle \psi_n, u_\delta \rangle|^2 , \quad p_n(\Lambda) = \sum_{\lambda \in \Lambda} |\langle u_n, \psi_\lambda \rangle|^2 .$$  

It follows from Parseval’s formula that for all $n$ and $\Delta$ (resp. $\Lambda$),

$$0 \leq p_n(\Delta) \leq 1$$  

The following elementary result provides an estimate for the order of magnitude of the Parseval weights :

$$E_{\Lambda} \{p_n(\Lambda)\} = p ; \quad E_{\Delta} \{p_n(\Delta)\} = \tilde{p} .$$  

Taking into account the distribution of $\Delta$ leads to this simple consideration on the behavior of observed coefficients : if the distribution of the Parseval weights is peaked near the origin, then the coefficients $a_k$ will be accurately described as a mixture of two Gaussians. Otherwise, more Gaussians will have to be taken into account.

**Remark 2** The distribution of the Parseval weights is however difficult to study practically, as it depends on both the significance maps and the incoherence of the dictionary $\mathcal{D}$. In a few words : the sparser the significance maps, the smaller the Parseval weights; and the less coherent the dictionary, the smaller the Parseval weights.

The spectrum of the covariance matrix is of interest too. The following estimate results from a direct application of Gershgorin’s disk theorem.
Theorem 2 Let \( z \) be an eigenvalue of the covariance matrix \( C^\Lambda \). Then there exists \( k \in I_N \) such that
\[
|z - (\sigma^2 X_k^\Lambda + \tilde{\sigma}^2 p_k(\Delta) + \sigma_0^2)| \leq \tilde{\sigma}^2 \epsilon_k(\Delta),
\]where
\[
\epsilon_k(\Delta) = \sum_{\ell \neq k} \sum_{\delta \in \Delta} |\langle \psi_{\ell}, u_k \rangle |^2 |\langle u_k, \psi_\delta \rangle |.
\]
Hence, if for all \( k \) one has that \( p_k(\Delta) > \epsilon_k(\Delta) \), the covariance matrix is non-degenerate. However, this is not the general situation (in particular for small noise variance \( \sigma_0 \)), and the vector of \( a_k \) coefficients is sometimes a degenerate Gaussian random vector, whose distribution may nevertheless be characterized by its characteristic function.

2.4. 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
\[
\mathbb{P}\{\Delta\} = \tilde{p}^{|\Delta|}(1 - \tilde{p})^{N - |\Delta|}, \quad \mathbb{P}\{\Lambda\} = p^{|\Lambda|}(1 - p)^{N - |\Lambda|},
\]the coefficients marginal distribution taking the simple form
\[
\rho_{a_a}(\xi) = (1 - p) \sum_\Delta \mathbb{P}\{\Delta\} \exp\left(-\xi^2/(2(\sigma^2 p_{\Delta}(\Delta) + \sigma_0^2))\right) + p \sum_\Delta \mathbb{P}\{\Delta\} \exp\left(-\xi^2/(2(\tilde{\sigma}^2 p_{\Delta}(\Delta) + \sigma_0^2))\right)
\]
\[
(10)
\]
The distribution of the coefficients is thus a mixture of two Gaussian mixtures, whose behavior is again governed by the Parseval function.

2.5. Structured models

Unlike the Bernoulli model, structured significance maps models involve correlations between significance maps elements. For example, assuming \( U \) is an orthonormal basis of time-frequency waveforms, correlations may be introduced between successive 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 as in (10), but the probabilities are not as simple as before.

Interestingly enough, due to the decorrelation of the \( \alpha \) and \( \beta \) coefficients, the correlations in significance maps do not show up in the second order moments of the observed coefficients \( a \) and \( \beta \), i.e. in matrices \( C^\Lambda \) and \( C^\Delta \). For instance, neither \( \mathbb{E}_\Lambda \{ C^\Lambda_{kl} \} \) nor \( \mathbb{E}_\Delta \{ C^\Delta_{kl} \} \) involve the correlation functions of the significance maps \( \Gamma_{kl}^\Lambda := \mathbb{E}_\Lambda \{ X_{k}^\Lambda X_{l}^\Lambda \} \) or \( \Gamma_{kl}^\Delta := \mathbb{E}_\Delta \{ X_{k}^\Delta X_{l}^\Delta \} \). Significance maps correlations manifest themselves in higher order moments, for example 4th order moment. Indeed, according to Isserlis’ formula
\[
\mathbb{E}_0 \{ a_{\alpha} a_{\alpha} a_{\alpha} a_{\alpha} \} = C_{\alpha}^0 C_{\alpha}^\Lambda + C_{\alpha}^0 C_{\alpha}^\Lambda + C_{\alpha}^0 C_{\alpha}^\Lambda + C_{\alpha}^0 C_{\alpha}^\Lambda,
\]
and the expectations \( E_\Lambda \) and \( E_\Delta \) of the latter quantities do involve the significance maps correlations \( \Gamma_{kl}^\Lambda \) and \( \Gamma_{kl}^\Delta \). However, these expressions can hardly be exploited in estimation algorithms.

3. HYBRID MODEL ESTIMATION

From now on, we limit ourselves to the simpler, Bernoulli model. We propose a new family of “three-steps” procedures for the estimation and separation of the two layers above, in the framework of the Bernoulli model. In the numerical illustrations shown below, we shall be using a pair of orthonormal bases.

3.1. Parameter and significance map estimation

In a first step, the parameters have to be estimated. For that, the empirical distribution of the \( a \) and \( \beta \) coefficients is compared to a Gaussian mixture model, whose parameters are estimated using EM algorithm. When the signal admits a sparse enough expansion, i.e. when the Parseval weights are small enough, a mixture of two Gaussians may be recovered in that way. Otherwise, an iterative EM algorithm may be used to recover a more complex Gaussian mixture, further identified with a mixture of two Gaussian mixtures. In any case, EM provides estimates for the variances \( \sigma^2 + \tilde{\sigma}^2 p_{\Delta}(\Delta) \) and \( \tilde{\sigma}^2 p_{\Delta}(\Delta) \), and similar quantities for the \( b_{\beta} \) coefficients, as well as membership probabilities \( p \) and \( \tilde{p} \).

Once the parameters have been estimated, it may be shown that maximum likelihood estimation of the significance maps \( \Lambda \) and \( \Delta \) amounts to simple thresholding strategies, the threshold being set to the value for which the two pdfs intersect.

3.2. Coefficient estimation and multilayer signal decomposition

The significance maps estimation actually amounts to a dimension reduction. Given a signal \( x \), and the estimates \( \hat{\Lambda} \) and \( \tilde{\Lambda} \) for \( \Delta \), the subspace of \( \mathcal{H}_\Phi \) spanned by the subdictionary \( \mathcal{D} = \{ \psi_\alpha, \lambda \in \hat{\Lambda} \} \cup \{ u_\delta, \delta \in \tilde{\Lambda} \} \). The estimates for the \( U \) and \( \Psi \) layers of \( x \) using the reduced dictionary are obtained from its orthogonal projection onto \( \mathcal{H}_\Phi \) : the signal approximation is \( \hat{x} = \hat{x}_U + \hat{x}_\Phi \), with
\[
\hat{x}_U = \sum_{\lambda \in \hat{\Lambda}} \hat{\beta}_\lambda u_\lambda, \quad \hat{x}_\Phi = \sum_{\lambda \in \tilde{\Lambda}} \hat{\alpha}_\lambda \psi_\lambda.
\]
(11)
The coefficients \( \hat{\alpha} \) and \( \hat{\beta} \) are obtained via
\[
\begin{pmatrix} \hat{\alpha}_1, \ldots, \hat{\alpha}_{|\hat{\Lambda}|}, \hat{\beta}_1, \ldots, \hat{\beta}_{|\tilde{\Lambda}|} \end{pmatrix}^T = G^{-1} \begin{pmatrix} a_1, \ldots, a_{|\hat{\Lambda}|}, b_1, \ldots, b_{|\tilde{\Lambda}|} \end{pmatrix}^T, \quad (12)
\]
\( G \) being the Gram matrix of the dictionary. Therefore, the estimation of the two layers simply amounts to a matrix inversion, of size \( |\hat{\Lambda}| + |\tilde{\Lambda}| \), which is a reasonable task when the significance maps are sparse enough.

Remark 3 When the significance maps are too large, so that inverting the Gram matrix becomes computationally expensive, an alternative is provided by Wiener-type estimation [9] : estimates \( \hat{\alpha}_\lambda \) and \( \hat{\beta}_\delta \) are obtained by a suitable weighting of the observed coefficients:
\[
\hat{\alpha}_\lambda = \frac{\sigma^2}{\sigma^2 + \tilde{\sigma}^2 p_{\Delta}(\Delta) + \sigma_0^2} a_\lambda.
\]
The estimates obtained are poorer estimates, but easier to compute.

3.3. Numerical illustrations

The approach proposed in this paper was implemented using the MATLAB scientific computing environment, exploiting an MDCT extracted from the Stanford Wavelab toolbox.

\footnote{1See \url{http://www-stat.stanford.edu/~wavelab}}
We illustrate our approach by an example of multilayered decomposition of xilophone signal (65536 samples at sampling rate $\eta = 44.1 \text{kHz}$, i.e. about 1.5 sec of sound), expanded onto a dictionary constructed as the union of two MDCT bases. The two bases were given windows of respective widths 2048 samples (about 46 msec.) and 128 samples (about 3 msec). The algorithm selected $|\Delta| = 138$ tonal basis functions for the tonal layer, and $|\Lambda| = 976$ transient basis functions for the transient layer, i.e. all together less than 1.7% of the number of samples. The complete decomposition took about 3 mn on a 2.6GHz linux PC with 512 Mo RAM.

We display in Fig. 1 the estimated U component (termed “tonal layer”) and the $\Psi$ component (transient layer). As may be seen, the separation is quite neat, the attacks have been smoothed out in the tonal part, and the transient part is sharply peaked as expected. Even though the result is not perfect, it is nevertheless of very good quality, given the small number of coefficients taken into account for the reconstruction.

![Fig. 1. Xilophone signal. Top : original; bottom : tonal layer and transient layer](image)

4. CONCLUSIONS AND PERSPECTIVES

The random waveform models described in this note provide an interesting and valuable approach for modeling audio signals featuring different components such as tonals, transients,... We have presented here a simple estimation algorithm, that does not rely on complex optimization techniques, and does therefore allow computations in reasonable times. Our approach yields very sparse signal expansions, which may be of interest in various applications, in particular signal coding, following the lines of [1, 2].

Possible modifications and extensions of this approach could involve the following :

- The introduction of frequency dependent variances for $\alpha$ and $\beta$ coefficient in the model (1). Indeed, when $|\Delta|$ and $|\Lambda|$ are taken into account, which raises (among others) the following question : to which extent may the different channels share a common significance maps as a function of desired quality. A scalable version of this approach is currently under study.

- In an audio coding perspective, the proposed scheme is not scalable, i.e. it does not allow the control of the size of the significance maps for audio signals featuring “clean” signal contains essentially tonals and transients (like the xilophone signal studied here), the estimated residual will essentially contain noise. In such case, our approach may be directly used for denoising purpose. However, in some cases, the “clean” signal may contain additional wide band “stochastic like” components, which should be modelled further in order to be distinguishable from the noise.

- Extension to the multichannel situation can also be considered, which raises (among others) the following question : to which extent may the different channels share a common significance maps.

5. REFERENCES


