A STUDY OF BERNOULLI AND STRUCTURED RANDOM WAVEFORM MODELS FOR AUDIO SIGNALS

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

The empirical pdf of wavelet or MDCT coefficients of audio signal generally feature a sharp peak at the origin, together with heavy tails. We show that such features may be reproduced if audio signals are modelled as sparse series of waveforms, randomly taken from a union of two significantly different orthonormal bases. In this context we obtain estimates for the behavior of "observed" coefficients, and numerical results on audio signals. Unlike more classical approaches involving optimization algorithms, our approach approaches thus relies on an explicit model. These allow to analyze mathematical properties of such signals and corresponding estimators, and derive simple estimation algorithms.

1. PROBLEM STATEMENT

Recently, audio signal involving decompositions of the form

signal = tonal + transient + residual

have received some attention (see for example [1, 2, 3, 4]), and algorithms for estimating the corresponding components have been proposed. Such techniques possess a wide range of applications, including among others audio signal coding and compression (the hope being that the improvement in the coding of the different components can compensate the fact that several components have to be encoded), denoising (the rationale being that a good representation for all components will concentrate its energy in a small amount of data, which would not be true for the noise), transcription,.... Such signal representations can often be obtained as by-products of some sparse coding algorithms, in which the models are not specified explicitly (see for example [5, 6, 7]).

We study here more specific models, based on expansions on elementary waveform systems. The main idea is to start with an orthonormal basis (or a frame) of waveforms from which a given component (tonal, or transient) is supposed to admit sparse expansions. Such assumptions are supported by the following "experimental fact" : both wavelet and MDCT coefficients of audio signals feature a significant peak near the origin, together with heavy tails. These pdfs could be modelled, in first approximation, as mixtures of two pdfs with significantly different variances : small and large coefficients. The latter may be interpreted in terms of *layers* of different nature in the signal. An illustration of this fact can be found in **Fig. 1**, in which the pdfs of wavelet and MDCT coefficients of two audio signals (organ and castanet) are displayed. These pdfs indeed exhibit the above mentioned behavior, and suggest developing signal models that would indeed match the behaviors observed in **Fig. 1**.
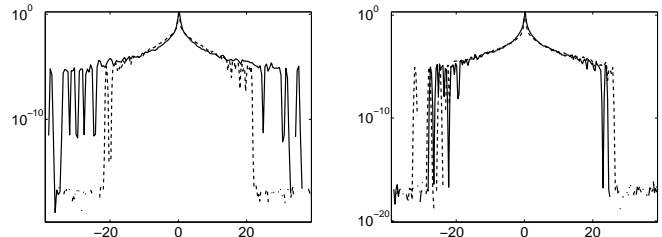
Fig. 1. pdf of various representations of two sample signals : castanet (solid line) and organ (dotted line) : wavelet (left) and MDCT (right) coefficients.

The models we propose aim at reproducing such a behavior, and are based on the following ingredients. We work in a finite dimensional setting, and denote by $I_N = \{1, \ldots N\}$ a corresponding index set.

- **Waveforms :** we consider orthonormal bases $\mathbf{U} = \{u_1, ..u_n\}$ and $\mathbf{\Psi} = \{\psi_1, ...\psi_n\}$ of $\mathbb{C}^N$, and the *dictionary* constructed as the union $\mathbf{U} \cup \mathbf{\Psi}$.

- **Significance maps :** let Λ and Δ be random subsets of the index set I_N, and for a given realization of Δ (resp. Λ), let X_n^Δ (resp. X_n^Λ) denote the corresponding (random) indicator : $X_n^\Delta = 1$ if $n \in \Delta$ and 0 otherwise (similarly for Λ).

- **Coefficients :** to each $\delta \in \Delta$ (resp. $\lambda \in \Lambda$) is associated a random variable β_δ (resp. α_λ). These random variables will be assumed i.i.d. $\mathcal{N}(0, \tilde{\sigma}^2)$ (resp. $\mathcal{N}(0, \sigma^2)$), and the α and β coefficients are also assumed independent.

The random waveform model associated with these data takes the form

$$ x = \sum_{n=1}^N X_n^\Delta \beta_n u_n + \sum_{m=1}^N X_m^\Lambda \alpha_m \psi_m + r \, , \qquad (1) $$

where r is some *residual* signal, modelled as a second order, wide sense (cyclically) stationary random signal. Here, we shall limit ourselves to the simple case of a Gaussian white noise.

Numerical experiments show that such models do succeed at reproducing the behavior of the pdfs displayed in **Fig. 1** and **Fig. 2** show the pdfs of MDCT and wavelet coefficients of signals generated according to the above model.

Given such a signal model, the main problems are the following : 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"),

1. Parameter estimation : estimate the parameters of the model (variances of coefficients, distribution of the significance maps,...)
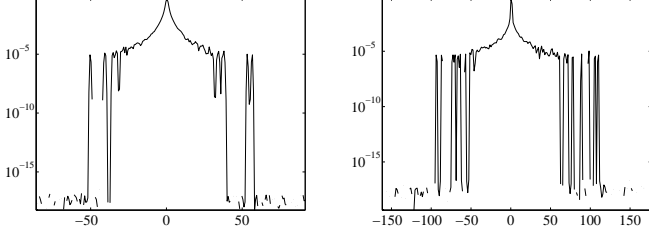
Fig. 2. pdfs of wavelet (left) and MDCT (right) coefficients of a synthetic signal generated following the Random Waveform Model.

2. Estimate the significance maps Λ and Δ.

3. Estimate the two *layers* : the $\mathbf{U}$ and Ψ parts of the signal

$$x_{\mathbf{U}} = \sum_{n=1}^{N} X_n^{\Delta} \beta_n u_n \,, \quad x_{\Psi} = \sum_{m=1}^{N} X_m^{\Lambda} \alpha_m \psi_m \quad (2)$$

To identify the so-obtained layers with tonal and transient components of ausio signals, the orthonormal bases $\mathbf{U}$ and Ψ have to be chosen adequately. Following, among others [4] we limit ourselves to the case of a wavelet basis for the transient part, and an MDCT basis (with sufficiently large window, say about 25 or 50 msec long) for the tonal part. Other choices are possible, among which MDCT bases with different window sizes, or more generally frames.

2. THEORETICAL STUDY

2.1. Parseval weights, and behavior of observed coefficients

The main ingredients of the study will be the *observed* coefficients of the signal with respect to the two bases :

$$b_n = \langle x, u_n \rangle \,, \qquad a_m = \langle x, \psi_m \rangle \,. \quad (3)$$

Conditionally to the significance maps Λ and Δ these are (correlated) zero mean Gaussian random variables. Their covariance structure is governed by the Parseval weights

$$p_n(\Lambda) = \sum_{\lambda \in \Lambda} |\langle u_n, \psi_\lambda \rangle|^2 \,, \quad (4)$$

$$\tilde{p}_m(\Delta) = \sum_{\delta \in \Delta} |\langle u_\delta, \psi_m \rangle|^2 \,, \quad (5)$$

whose distribution depends on the coherence of the dictionary, and the distribution of the significance maps. More precisely :

- The sparser the significance maps, the smaller the Parseval weights.
- The more "different" the two bases, the smaller the Parseval weights.

The uncorrelatedness of the coefficients α and β in fact yield the following Gaussian mixture model for the observed coefficients. For wavelet coefficients (a similar result holds true for MDCT coefficients) we can state

Theorem 1 *Under the above assumptions, with r a white noise with variance σ_0^2,*

1. *Conditional to the significance maps, the observed wavelet coefficients are zero-mean Gaussian random variables, with variance*

$$Var\{a_m\} = \sigma^2 X_m^{\Lambda} + \tilde{\sigma}^2 p_m(\Delta) + \sigma_0^2 \quad (6)$$

2. *For every eigenvalue z of the covariance matrix, there is an index $k \in I_N$ such that*

$$|z - Var\{a_k\}| \leq \tilde{\sigma}^2 \sqrt{p_k(\Delta)} \sum_{\ell \neq k} \sqrt{p_\ell(\Delta)} \quad (7)$$

<u>Sketch of the proof</u> : The computation of the covariance matrices of the observed coefficients may be carried out explicitly, and yields the following result

$$\mathbb{E}_0\{a_k \overline{a}_\ell\} = (\sigma^2 X_k^{\Lambda} + \sigma_0^2)\delta_{k\ell} + \tilde{\sigma}^2 \sum_{m=1}^{N} X_m^{\Delta} \langle \psi_k, u_m \rangle \langle u_m, \psi_\ell \rangle \,,$$

The estimate in the second part of the theorem then follows from Gershgorin's disk theorem. The case of MDCT coefficients is handled similarly. □

Taking the randomness of the significance maps into account, we obtain a mixture of two Gaussian mixtures, whose characteristics depend upon the distribution of the Parseval weights. We denote by p and $\tilde{p}$ the *membership probabilities* of the maps (points of the maps are assumed to be identically distributed)

$$\tilde{p} = \mathbb{P}\{n \in \Delta\} \,, \quad p = \mathbb{P}\{n \in \Lambda\} \,, \quad n \in I_N \,. \quad (8)$$

Numerical simulations (not shown here) actually show that if the significance maps are sparse enough (i.e. if p and $\tilde{p}$ are small enough), and if the two bases are sufficiently different, then the Parseval weights tend to be small, and the distribution of observed coefficients reproduces fairly well the "experimental" shapes displayed in **Fig. 1**.

"Mean field" type approximations yield fairly simple approximations of the distribution of observed coefficients as mixture of <u>two</u> Gaussians. Denote by $\mathbb{E}_\Delta$ the expectation with respect to the significance map Δ. Then one has for example

Corollary 1 *Assume that the elements of the index set I_N are identically distributed. Then on average with respect to Δ,*

$$\mathbb{E}_\Delta\{\mathbb{E}\{a_m a_n\}\} = \delta_{mn} \left(\sigma^2 X_m^{\Lambda} + \tilde{p}\tilde{\sigma}^2 + \sigma_0^2\right) \,. \quad (9)$$

This results kind of supports a model of mixture of two gaussians. However, it is worth stressing that the latter equation does not involve the two bases, and only exploits the sparsity of the significance maps. In other words, it does not depend on whether the two bases $\mathbf{U}$ and Ψ are very different or not (it would hold true as well if $\mathbf{U} = \Psi$).

Interestingly enough, the estimates given in Theorem 1 and Corollary 1 do not involve second order moments of the significance maps. Informations regarding the dependence between elements in the index set show up in higher order moments of the observed coefficients. Different models, including Bernoulli models, or structured models such as Markov or Ising models, yield different estimates.

2.2. Estimating parameters, significance maps and coefficients in the case of the Bernoulli model

We now limit ourselves to the case of significance maps distributed according a *Bernoulli model* : starting from fixed membership probabilities p (resp. $\tilde{p}$), the index values $n \in I_N$ are <u>iid</u>, and belong to Λ (resp. Δ) with probability p (resp. $\tilde{p}$) and to $\overline{\Lambda}$ (resp. $\overline{\Delta}$) with probability $1 - p$ (resp. $1 - \tilde{p}$).

Given observed coefficients, the simplest strategy for estimating parameters of the model is to rely on Theorem 1, and fit a Gaussian

mixture model to the empirical distribution of observed coefficients. In such a way, one ends up with estimates for membership probabilities p and $\tilde{p}$, and variances σ^2, $\tilde{\sigma}^2$ and σ_0^2. Given this empirical estimates, observed coefficients may be classified accordingly, which yields estimates $\widehat{\Delta}$ and $\widehat{\Lambda}$ for the significance maps (see Section 3 for more details).

However, it is worth mentioning that the Gaussian mixture fit here is not a simple problem, as according to Theorem 1, the observed coefficients are distributed according to a mixture of two (hopefully significantly different) Gaussian mixtures. In our approach, the strategy is to fit it with a mixture of a small number of Gaussians, most generally larger than two (which would be the choice suggested by Corollary 1).

Finally, assuming that the parameters and significance maps have been suitably estimated, the estimation of the layers may be carried out in a simple way. Depending on the situation, two different approaches for that problem can be exploited.

1. Assume that the $\boldsymbol{\Psi}$ and $\mathbf{U}$ components of the signal are sparse enough. Then the estimates $\widehat{\Delta}$ and $\widehat{\Lambda}$ of the significance maps generate a *sub-dictionary* $\widehat{\mathcal{D}} = \{\psi_\lambda, \lambda \in \widehat{\Lambda}\} \cup \{u_\delta, \delta \in \widehat{\Delta}\}$ of the complete waveform dictionary $\mathbf{U} \cup \boldsymbol{\Psi}$. Therefore, the orthogonal projection of the signal x onto the linear span of $\mathcal{D}$ directly yields the desired decomposition.

2. If the signal is not sparse enough, i.e. if the estimated significance maps are large sets, the above technique (which involves the inversion of a matrix as large as the dictionary) may yield high computational load. In such a case, instead of an orthogonal projection (that minimizes the distance between x and the linear span of the dictionary), a Wiener-type method may be used (which amounts to minimize the distance between x and the linear span of the dictionary, on average with respect to $\mathbb{P}_0$). More precisely, an estimate of the form
$$\begin{cases} \hat{x}_{\boldsymbol{\Psi}} &= \sum_{\lambda \in \Lambda} t_\lambda a_\lambda \psi_\lambda \,, \\ \hat{x}_{\mathbf{U}} &= \sum_{\delta \in \Delta} \tilde{t}_\delta b_\delta u_\delta \end{cases} \tag{10}$$
is seeked, where the weights are chosen so as to minimize the mean squared error.

This may be done thanks to the following

Theorem 2 *Conditionally to the significance map Δ, the optimal weights t_λ and $\tilde{t}_\delta$ for the Wiener-type estimate (10) in the transient layer are given by*
$$\begin{cases} t_\lambda &= \frac{\sigma^2}{\sigma^2 + \tilde{\sigma}^2 p_\lambda(\Delta) + \sigma_0^2} \,, \\ \tilde{t}_\delta &= \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma^2 p_\delta(\Lambda) + \sigma_0^2} \,. \end{cases} \tag{11}$$

3. ALGORITHMS AND NUMERICAL RESULTS

We now illustrate and comment on the results obtained using the "three-steps" procedures we developed for the estimation and separation of the two layers above, in the framework of the Bernoulli model. Using a pair of orthonormal bases (here, wavelet and MDCT bases), we proceed as follows :

- Computation of observed wavelet and MDCT coefficients of the signal, followed by an estimation of the parameters of the models (variances, membership probabilities). For this, we use EM algorithms, which are very well adapted to Gaussian mixtures. However, due to the fact that all Gaussian pdfs that
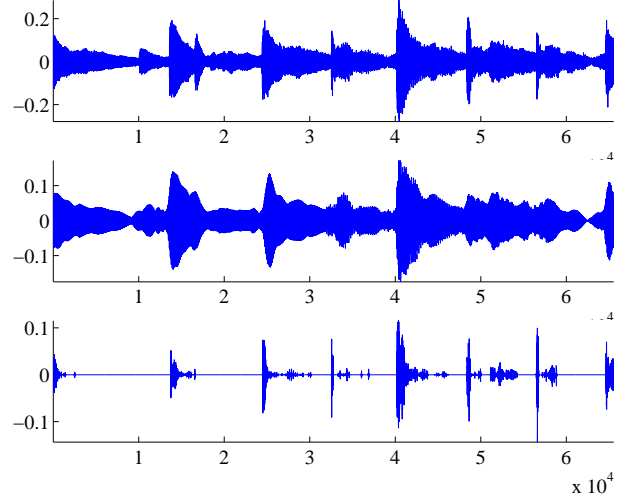

Fig. 3. Xilophone signal, decomposed using an MDCT (window length 2048 samples) and a Daubechies 10 wavelet basis. Top : original; middle : tonal layer; bottom : transient layer

come into play here are zero-mean, and that more than two pdfs are to be estimated, we rely on an approximate approach in which the EM algorithm is ran several times for separating the "large variance" component from several "small variance" ones.

- Estimation of the significance maps : the EM algorithm described above labels the observed coefficients, depending on the pdf they are assigned to. "Large variance" coefficients are assigned to the significance maps Λ and Δ.

- Estimation of the two layers : according to the above discussion, two approaches were tested and compared.

    1. Least square optimization on the subdictionary
    $$\widehat{\mathcal{D}} = \{\psi_\lambda, \lambda \in \widehat{\Lambda}\} \cup \{u_\delta, \delta \in \widehat{\Delta}\} \,.$$
    of $\mathcal{D} = \boldsymbol{\Psi} \cup \mathbf{U}$ induced by the significance maps. This amounts to compute the Gram matrix associated to the family of time-frequency atoms corresponding to $\widehat{\Delta}$ and $\widehat{\Lambda}$, and obtain new coefficients by applying the matrix to the observed coefficients.

    2. Wiener-type filtering of retained coefficients a_λ and b_δ, following the lines of Theorem 2, and equations (10) and (11).

An example of separation between tonal and transient layer based upon the Bernoulli model is displayed in **Fig. 3** and **Fig. 4**. The signal is a short 1.5 sec long piece of xilophone signal, sampled at 44.1 kHz. The two orthonormal bases were an MDCT basis (with maximally smooth, 2048 samples long window), and Daubechies 10 and Daubechies 20 wavelet bases. The two layers were estimated following the lines of the algorithm presented above, using the orthogonal projection method for computing the layers from estimated significance map and coefficients.

As may be seen, the algorithm did a fairly good job at separating the different components; as could be anticipated, the transients are better resolved when a Daub10 wavelet basis is used than with the
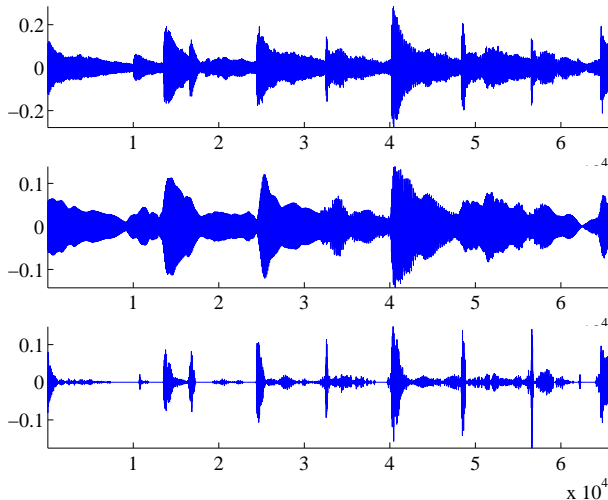
Fig. 4. Xilophone signal, decomposed using an MDCT (window length 2048 samples) and a Daubechies 20 wavelet basis. Top : original; middle : tonal layer; bottom : transient layer

Daub20 basis, which has poorer time localization. In both cases, $|\widehat{\Delta}| + |\widehat{\Lambda}| < 1600$, i.e. about 4% of coefficients (out of 65536) were retained.

Using the Wiener-type approach for computing the transient and tonal estimates (results not shown here) yields similar estimates, of poorer quality though. This could be anticipated, as the Wiener-type estimates come from an *average* error minimization (unlike the Gram matrix based estimation, that exploits the realization of the signal at hand).

4. CONCLUSIONS

We have presented in this note a family of simple random models involving sparse expansions in waveform dictionaries. We have more specially focused on dictionaries constructed as unions of two (significantly different) orthonormal bases Ψ and $\mathbf{U}$ of the underlying signal space. Numerical simulations show that such models turn out to be fairly realistic for describing audio signals, which suggest to exploit them for audio coding.

A more precise study of the behavior of the *observed coefficients* of the signal with respect to the Ψ and $\mathbf{U}$ bases leads to simple strategy for estimating the corresponding layers. This approach turns out to produce very sparse approximations of audio signals. Even though these approximations seem to be of poorer quality than those obtained using more sophisticated approaches (see for example [8]), our approach (which may be seen as a first step towards more elaborate models) is much simpler, and efficient in terms of computational load.

Among the extensions that are currently under study, let us mention two possible ways of improving the models we propose :

- *Frames :* replacing the orthonormal bases with frames of waveforms (see for example [9] for a definition) would allow one to use waveforms possessing better time-frequency localization properties. However, the redundancy of frames happens to make the analysis more difficult, in particular when it comes to estimate the significance maps. Different strategies seem to be needed here.

- *Structures :* As shown in [3, 4], relying on individual coefficients is often not enough to split audio signals into layers in a sensible way. Considering *structured sets* of coefficients instead than individual coefficients turned out to significantly improve the results. The random waveform models we presented here offer ways of implementing structures into the model. However, the resulting estimation algorithms do not seem to remain as simple as the algorithms adapted to the Bernoulli case.

To conclude, let us also mention alternative algorithms, based on variational approaches, which have been considered recently [10]. A systematic comparison with such approaches would probably be extremely instructive.

5. REFERENCES

[1] K. N. Hamdy, A. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representations," in *IEEE International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, 1996, vol. 2, pp. 1045–1048.

[2] T. Verma and T. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[3] 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.

[4] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Applied and Computational Harmonic Analysis*, vol. 18, pp. 137–166, 2005.

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

[6] D.L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decompositions," *IEEE Trans. Inf. Th.*, vol. 47, no. 7, pp. 2845–2862, 2001.

[7] F. Jaillet and B. Torrésani, "Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered Gabor expansions," Tech. Rep., Laboratoire d'Analyse, Topologie et Probabilités, Université de Provence, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France, 2005.

[8] C. Févotte and S. Godsill, "A bayesian approach for blind separation of audio sources," *IEEE Transactions on Speech and Audio Processing*, 2005, to appear.

[9] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, PA, 1992.

[10] G. Teschke, "Multi-frame representations in linear inverse problems with mixed multi-constraints," Tech. Rep. DFG-SPP-1114 preprint 90, Universität Bremen, FB Mathematik und Informatik / ZETEM , Postfach 33 04 40, 28334 Bremen, Germany, 2005.