HAL open science

Sparse signal decomposition on hybrid dictionaries using musical priors

Hélène Papadopoulos, Matthieu Kowalski

To cite this version:

Hélène Papadopoulos, Matthieu Kowalski. Sparse signal decomposition on hybrid dictionaries using musical priors. ISMIR 2011, Oct 2011, Miami, United States. p.687-692. hal-00621048

HAL Id: hal-00621048

https://hal.science/hal-00621048v1

Submitted on 24 Nov 2011

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.

SPARSE SIGNAL DECOMPOSITION ON HYBRID DICTIONARIES USING MUSICAL PRIORS

Hélène Papadopoulos and Matthieu Kowalski
Laboratoire des Signaux et Systèmes
UMR 8506, CNRS-SUPELEC-Univ Paris-Sud
91172 Gif-sur-Yvette Cedex
helene.papadopoulos@lss.supelec.fr
matthieu.kowalski@lss.supelec.fr

ABSTRACT

This paper investigates the use of musical priors for sparse expansion of audio signals of music on overcomplete dictionaries taken from the union of two orthonormal bases. More specifically, chord information is used to build structured model that take into account dependencies between coefficients of the decomposition. Evaluation on various music signals shows that our approach provides results whose quality measured by the signal-to-noise ratio corresponds to state-of-the-art approaches, and shows that our model is relevant to represent audio signals of Western tonal music and opens new perspectives.

1. INTRODUCTION

We propose in this paper a new approach for structured *sparse* decomposition of a music signal in an overcomplete time-frequency dictionary. Starting from existing methods that are based on physical signal properties, we propose to incorporate musical priors in order to built signal representations that are more suitable to music. For this, we take advantage of the recent works that have been done on chord estimation in the context of music content processing.

The problem of representing an audio signal using a time-frequency dictionary has been given a lot of attention these last few years. The specificity of music audio signals is that, very often, several types of components are superimposed, as for instance tonal components (the partials of the notes) and transients (the attacks of the notes). These various components may have significantly different behaviors. For instance fast varying transient require short analysis window whereas low varying tonals require long windows. Thus, they cannot be represented within the same basis. This is why *hybrid* models allowing a simultaneous representation of different components have been proposed [4, 12, 17, 22].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2011 International Society for Music Information Retrieval.

Among the various existing transforms, the modified discrete cosine transform (MDCT) [15] is a standard choice for the bases [6, 14]. Following these approaches, we consider in this work a dictionary built as the union of two MDCT bases with different time-frequency resolutions. The narrow band basis - with long time resolution - is used to estimate the tonal parts of the signal, and the wide band basis - with short time resolution - is used to estimate the transient parts. Such a dictionary is chosen overcomplete, and thus the expansion of the signal with respect to the dictionary is not unique. *Sparsity* may be used as a selection criterion for finding the expansion coefficients, in the sense that only a few coefficients of the decomposition of the signal on the bases are significantly nonzero. The signal can thus be well approximated by a limited number of coefficients. This problem is often referred to as *sparse regression*.

A common approach to find a sparse expansion of signals in overcomplete dictionaries consist of minimizing the ℓ_1 norm of the expansion, and is known as *basis pursuit* [1], or LASSO [21]. Various methods have been also proposed: they include variational approaches [13], probabilistic approaches [14], greedy methods, such as matching pursuit algorithms [2, 16], or Bayesian formulations as for instance EM-based algorithms [9]. In the framework of Bayesian variable selection, MCMC (Markov chain Monte Carlo) type approaches that consider a dictionary constructed as the union of two orthonormal bases have been proposed [5, 7]. One of the main advantages of MCMC techniques is their robustness because they scan the whole of the posterior distribution and thus are unlikely to fall into local minima. However, this is done at the expanse of high computational cost.

In order to fully exploit the dual nature of audio music signals mentioned above, some approaches consider dependencies between significant coefficients. In the time-frequency plane, partials of the note will generate horizontal lines localized in frequency, whereas the attacks of the notes and the percussive sounds will generate vertical lines localized in time. Ideally, this structure should be reflected in the signal decomposition. This is why we are interested in finding a signal approximation that is not only sparse, but also structured. Previous approaches that use unstructured priors, such as Bernoulli models have shown

that they generate isolated coefficients with high amplitude in both bases [7, 14]. These components do not have any musical meaning and are usually perceived as "musical artifacts" or "musical noise" in the reconstructed signal. Considering dependencies between atoms coefficients and using structured priors allows reducing the number of such undesirable components. Various approaches have been proposed for introducing dependencies between coefficients in the time-frequency domain. Structures can be modeled directly in the coefficients themselves, such as in [13]. However, dependencies are often introduced in the time-frequency indices, rather than directly in the coefficients themselves. Among existing approaches, frequency persistency properties of the transient layer can be modeled using structured Bernouilli models [14]; persistency along the frequency axis is favored using Markov models [17]; in [8], structural constraints on the coefficients that rely on physical properties of the signal are imposed for both layers, using two types of Markov chains. It results in a "horizontal structure" for the tonal layer and a "vertical structure" for the transient layer. Up to now, additional structure constraints that have been added rely on physical properties of the signal. The originality of our work is that we propose to incorporate priors that are based on musical information. Relying on the model presented in [8] within a Bayesian framework, we build a structured model for sparse signal decomposition that incorporates musical priors for tonal layer modeling. Our model is particularly well adapted to the tonal structure of signals and fits the intrinsic nature of Western tonal music.

Sparse representation of signals have recently proved to be useful for a wide range of applications in signal processing, such as denoising [6], coding and compression [3, 20] or source separation [7]. Here, we focus on the task of denoising an excerpt of musical audio. Our approach provides results whose quality in term of signal-to-noise ratio (SNR) corresponds to state-of-the-art approaches, while better reflecting the nature of music audio signal.

The structure of the paper is as follows. First, in Section 2, we present our model for sparse signal decomposition on hybrid dictionaries that incorporates musical priors; our main contribution is described in part 2.3. We briefly address the problem of parameters estimation in Section 3. In Section 4, we present and discuss the results of our model. Conclusions and perspectives for future works are given in Section 5.

2. SIGNAL MODEL

This section introduces first the mathematical model used to represent the audio signal, and then defines the priors chosen in a Bayesian context. Particularly, the new musical prior based on the *chromagram* is exposed in section 2.3.

2.1 Model

In this part, we describe our model for signal decomposition with sparse constraint on a *hybrid* dictionary of elementary waveforms or *atoms*. The dictionary is constructed as the union of two orthonormal bases with different time-frequency resolution that account respectively for the tonal and the transient parts of the signal. We rely on the model proposed in [8] and we consider a tree-layer signal model of the form: $signal = tonals + transients + residual$.

Let $V = \{v_n, n = 1, \ldots, N\}$ and $U = \{u_n, n = 1, \ldots, N\}$ be two MDCT bases of $\mathbb{R}^N$ with respectively long frame ℓ_{ton} to achieve good frequency resolution for tonals and short frame ℓ_{tran} to achieve good time resolution for transients. The MDCT is a bijective linear transform and we note $n_{ton} = \frac{N}{\ell_{ton}}$ and $n_{tran} = \frac{N}{\ell_{tran}}$ the number of frames for each basis (see Fig. 2). Here, n is a time-frequency index and will be denoted in the following $n = (q, \nu) \in [1, \ell_{ton}] \times [1, n_{ton}]$ or $n = (q, \nu) \in [1, \ell_{tran}] \times [1, n_{tran}]$.

We denote $D = V \cup U$ the dictionary made as the union of these two bases. D is overcomplete in $\mathbb{R}^N$, and any $x \in \mathbb{R}^N$ admits infinitely many expansions in the form:

$$x = \sum_{n \in I} \alpha_n v_n + \sum_{m \in I} \beta_m u_m + r \tag{1}$$

where $I = \{1, \ldots, N\}$, α_n and β_n are the expansion coefficients and r represents the noise term.

We are interested in sparse signals, i.e. signals that may be written as:

$$x = \sum_{\lambda \in \Lambda} \alpha_\lambda v_\lambda + \sum_{\delta \in \Delta} \beta_\delta u_\delta + r \tag{2}$$

where Λ and Δ are small subsets of the index set $I = \{1, \ldots, N\}$ that account for the significant coefficients. In what follows, they will be referred to as *significance maps*.

We introduce two indicator random variables $\gamma_{ton,n}$ and $\gamma_{tran,m}$ corresponding to the significance maps Λ and Δ:

$$\gamma_{ton,n} = \begin{cases} 1 & \text{if } n \in \Lambda \\ 0 & \text{otherwise} \end{cases} \quad \gamma_{tran,m} = \begin{cases} 1 & \text{if } m \in \Delta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We can therefore rewrite Eq. (2) as:

$$x = \sum_{n \in I} \gamma_{ton,n} \alpha_n v_n + \sum_{m \in I} \gamma_{tran,m} \beta_m u_m + r \tag{4}$$

2.2 Coefficient Priors

We assume that, conditional upon the significance maps Λ and Δ , the coefficients α_n and β_n are independent zero-mean normal random variables:

$$\begin{aligned} p(\alpha_n | \gamma_{ton,n}, \sigma_{ton,n}) &= (1 - \gamma_{ton,n}) \delta_0(\alpha_n) + \\ &\quad \gamma_{ton,n} \mathcal{N}(\alpha_n | 0, \sigma^2_{ton,n}) \\ p(\beta_n | \gamma_{tran,m}, \sigma_{tran,m}) &= (1 - \gamma_{tran,m}) \delta_0(\beta_m) + \\ &\quad \gamma_{tran,m} \mathcal{N}(\beta_m | 0, \sigma^2_{tran,m}) \end{aligned} \tag{5}$$

where δ_0 is the Dirac delta function and the variances $\sigma_{ton,n}$ and $\sigma_{tran,m}$ are given a conjugate inverted-Gamma prior. Sparsity is enforced when $\gamma_n = 0$ (resp. $\gamma_m = 0$). In this case, the coefficients α_n (resp. β_n) are set to 0.

2.3 Indicator Variable Priors

The significance maps Λ and Δ are given structured priors. The one corresponding to the tonal basis encodes musical information while the one corresponding to the transient basis is based on physical properties of the signal. Both of them are "vertical" structures.

2.3.1 Model for Tonals

For the significance map corresponding to the tonals, we propose to model dependencies between indicator variables using musical information. Let us assume that we know the score corresponding to the musical excerpt and that, for each frame $q \in \{1, \ldots, n_{ton}\}$, we know which notes the signal is composed of.

Here, we want to work directly on audio. However, the symbolic transcription (the score) of a piece of music is not always available, especially in music such as jazz music where there is a large part devoted to improvisation. In addition, algorithms that extract a transcription from an audio signal, such as multi-f0 estimation algorithms [24], are still limited and costly. However, numbers of recent works have shown that it is possible to accurately extract robust mid-level representation of the music, such as the chord progression [18].

We propose to give musical prior to the indicator variables using musical information obtained from the chord progression. The output of a chord estimation algorithm consists in a progression of chords chosen among a given chord lexicon. Each chord may be characterized by the semitone pitch classes or chroma that corresponds to the notes it is composed of. Since their introduction in 1999, *Pitch Class Profile* [10] or *chroma*-based representations [23] have become a common feature for estimating chords. They are traditionally 12-dimensional vectors, with each dimension corresponding to the intensity associated with one of the 12 semitone pitch classes (chroma) of the Western tonal music scale, regardless of octave. The succession of chroma vectors over time is known as *chromagram*.

In general, the chord lexicon does not distinguish between any possible combination of simultaneous notes but is typically reduced to a set of chords of 3 or 4 notes. The number of notes composing the chords will be denoted by N_c in the following. Here, we limit our chord lexicon to the 24 major and minor triads ($N_c = 3$). The method we propose could be extended to larger dictionaries.

The chord progression does not provide an exact transcription of the music. For instance, passing notes are in general ignored, missing notes in the harmony may be added. Moreover, the chords are estimated regardless of octave. However, experiments show that the provided musical information is sufficient enough to build musically meaningful priors.

Given a fixed frame index q, let $\{p_k^c\}_{k=1,\ldots,N_c}$ denote the semitone pitch-classes (chroma) corresponding to the estimated chord c_q. Let also $\{p_\nu^{MDCT}\}_{\nu=1,\ldots,\ell_{ton}}$ denote the semitone pitch classes corresponding to each MDCT bin.

Assuming a perfect tuning of $A = 440$Hz, a MDCT bin of frequency f_ν is converted to a chroma p_ν^{MDCT} by the following equation:

$$p_\nu^{MDCT} = 12 \log_2 \frac{f_\nu}{440} + 69 \pmod{12}\,^{1} \tag{6}$$

The indicator variables $\{\gamma_{ton,(q,\nu)}\}_{\nu=1,\ldots,\ell_{ton}}$ are given the following membership probabilities:

$$P_\Lambda\{\gamma_{ton,(q,\nu)} = 1\} \tag{7}$$
$$= \begin{cases} p_{ton} & \text{if } \exists k \in [1, N_c] \mid p_\nu^{MDCT} = p_k^c \\ 1 - p_{ton} & \text{otherwise} \end{cases}$$

where $0 \leq p_{ton} \leq 1$. The significance maps corresponding to the tonal layer should reflect the tonal content of the audio signal. In practice, the value p_{ton} will be close to 1 (in our experiments, $p_{ton} = 0.9$) so that atoms corresponding to the notes that are played are given high prior. The significant map for the tonal layer corresponding to the *Glockenspiel* audio signals of our test-set is illustrated in Fig. 1. A set of atoms is selected at each frame according note of the (chord) transcription, regardless of octave. For instance all atoms $\{B1, B2, \ldots\}$ corresponding to the semitone B are selected when the first B note of the *Glockenspiel* is sounded. The significance maps are given structures of "tubes" that have a musical meaning. Note that we provide here a "vertical structure" for tonals.

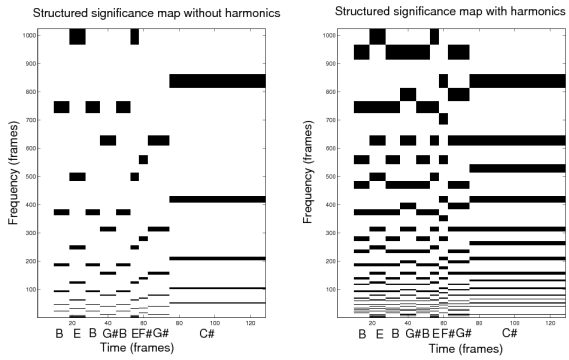

Figure 1. Structured significance map for the *Glockenspiel* using musical information. Left: only notes composing the chord are considered. Right: higher harmonics are considered. The transcription is indicated in the bottom.

Two additional components may be added to improve the model.

• First, the instruments may have been tuned according to a reference pitch different from the standard $A4 = 440$Hz. In this case it is necessary to estimate the tuning of the track and Eq. (8) becomes:

$$p_\nu^{MDCT} = 12 \log_2 \frac{f_\nu}{A_{est}} + 69 \pmod{12} \tag{8}$$

where A_{est} denotes the estimated tuning, here using the method proposed in [19].

[1] $a \pmod{b}$ denotes the mathematical operator *modulo*, the remainder when a is divided by b

• Secondly, higher harmonics may be considered in the model. Each note produces a set of harmonics that results in a mixture of non-zero values in the chroma vector corresponding to the chord. For instance a C note will produce the set of harmonics $\{C-C-G-C-E-G-\ldots\}$. They can be considered in the significance maps, as illustrated in the right part of Fig. 1. Here we take into account the first 6 harmonics of the notes [2].

2.3.2 Model for Transients

Following [8], persistency in frequency of time-frequency coefficients corresponding to transient layer is modeled giving a vertical prior structure to the indicator variables in the second basis. Given a frame index q, the sequence $\{\gamma_{tran,(q,\nu)}\}_{\nu=1,\ldots,\ell_{tran}}$ is modeled by a two-state first-order Markov chain with probabilities $P_{tran,00}$ and $P_{tran,11}$, assumed equal for all frames, and with learned initial probability π_{tran}. The model is illustrated in Fig. 2.

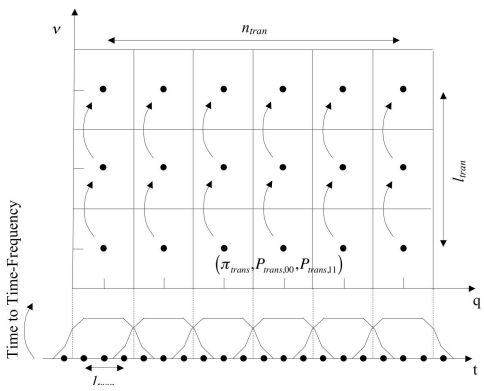

Figure 2. Vertical model for transients. Adapted from [8].

2.4 Residual

The residual signal r is modeled as a Gaussian white noise, with variance σ^2, which is given an inverted-Gamma conjugate prior.

3. MCMC INFERENCE

Following [8], the posterior distribution of the set of parameters and hyperparameters of the model, denoted by θ, is sampled from using a Gibbs sampler [11], which is a standard Markov Chain Monte Carlo (MCMC) technique that simply requires to iteratively sample from the posterior distributions of each parameter upon data x and the remaining parameters.

The Minimum Mean Square Estimates (MMSE) of the parameters θ can then be computed from the Gibbs samples $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}\}$ of the posterior distribution $p(\theta|x)$:

$$\hat{\theta}_{MMSE} = \int \theta p(\theta|x) d\theta \quad (9)$$

$$\approx \frac{1}{K} \sum_{k=1}^{K} \theta^{(k)} \quad (10)$$

[2] We limit the number of considered harmonics to 6 because many of the higher harmonics, which are theoretically whole number multiples of the fundamental frequency, are far from any note of the Western chromatic scale. This is especially true for the 7th and the 11th harmonics.

The MAP estimate can be computed by thresholding the values of the MMSE. In [8], all the values of the MMSE lower that 0.5 are threshold to 0 and all the values greater than 0.5 are threshold to 1.

We do not detail here the expression for the update steps of the parameters, details can be found in [8]. Time-domain source estimates are reconstructed by inverse transform of the estimated coefficients (inverse MDCT in our case). The denoised estimation is constructed by $\hat{x} = \alpha V + \beta U$.

4. RESULTS AND DISCUSSION

The aim of this section is to analyze the performances of the proposed approach for the task of audio denoising. For the sake of simplicity, we first focus in details on a monophonic signal, the *Glockenspiel*. We also provide additional numerical results and examples on short extracts of polyphonic music. The impact of the various parameters (tuning, harmonics, and priors settings) is also studied.

4.1 Experimental Setup

In this article, we present results assuming that the transcription is known (notes for the monophonic signal, chords for the polyphonic signals). The 5 musical excerpts of various music styles are described in Table 1. Our approach that incorporates musical priors for modeling the tonal layer is compared with the one presented in [8].

Table 1. Sound excerpts used for evaluation of the model. SR: sampling rate.

Name	SR (Hz)	Duration
Glockenspiel	44100	2s
Misery (Beatles)	11025	11s
Love Me Do (Beatles)	11025	5s
Beethoven String quartet Op.127 - 1	11025	11s
Mozart Piano Sonata KV310 - 1	11025	11s

Parameters: The length of the two MDCT bases are set to 1024 samples for the tonal layer and 128 samples for the transient layer, at a sampling rate of 44100Hz, and respectively to 256 and 32 samples at a sampling rate of 11025Hz [3]. The MMSE and MAP estimates of the parameters are computed by averaging the last 100 samples of the Gibbs sampler, run for 500 iterations.

Evaluation Measures: Artificial noisy signals are created by adding Gaussian white noise to the clean signal with various input SNRs. The case without additional noise WN (without noise) corresponds to a separation into two layers *transient* + *tonal*. Partials are expected to be recovered in the tonal layer while attacks or percussive sounds will be recovered in the transient layer. The results in terms of output SNR are summarized in Table 2 and provide an objective evaluation measure. However, although widely used for assessing algorithm performances, the SNR is not a completely relevant measure of distortion for audio signals. Subjective evaluation by listening to the signals is also required. The audio excerpts are available at: `http://`

[3] As underlined in [8, 14], better results are obtained using a very short window length for the transients ($\approx$ 3ms). The two window lengths must be significantly different enough to discriminate between tonals and transients

webpages.lss.supelec.fr/perso/kowalski/ismir11/ismir11.html.

Table 2. Resulting values of output SNRs (dB) for various input SNRs and without additional Gaussian noise (WN).

	Proposed approach				[8] approach			
SNR	WN	0	10	20	WN	0	10	20
Glockenspiel	71.2	14.1	21.3	28.5	70.2	15.7	22.5	29.2
Misery	42.3	7.0	13.0	20.9	44.4	7.3	13.3	21.1
Love Me Do	28.6	6.8	12.5	19.3	29.6	6.9	12.7	19.4
Beethoven	54.5	8.5	13.6	21.6	54.6	8.9	14.0	21.9
Mozart	62.6	9.3	15.4	23.4	60.9	9.8	15.9	23.9

Computational Performances: The algorithms are implemented in MATLAB and performed on a MacBook Pro Intel Core 2 Duo clocked at 2.4GHz with 2GB RAM. The computation time of the proposed method is similar to the one obtained with [8], ≈ 447 s for processing the *Glockenspiel* signal. The use of MCMC schemes generates high computational costs.

4.2 Results and Discussion

Concerning the quality of denoising, the results provided in Table 2 show that our model provides results that are comparable to state-of-the-art algorithms in terms of SNR: the difference between the presented method and the [8] are in general lower than 1 dB. However, noticeable differences may be perceived while listening to the sound files.

The main interest of the proposed model lies in the modeling of the tonal layer. Fig. 3 shows significance maps of the selected atoms (MAP estimates) for the *Glockenspiel* signal, in the WN case. As can be seen, the use of musical priors yields to a structure that better reflects the music content of the signal compared to the approach that use physical priors. The resolution of the tonal significance map is sharper. The partials of the notes clearly appear as thin horizontal lines and the beginning of the notes is very clear. One can also see that our method using musical priors provides sparser estimates of the significance map.

It should be noticed that, especially under low-input SNRs conditions, one may perceive some artifacts in the reconstructed signal with the method we propose. They are probably due to the fact that some high frequencies are captured by the transient basis rather than by the tonal basis. Future works should concentrate on modeling structured priors for the transient layer that are more adapted to the one proposed here for the tonal layer. However, in spite of these artifacts, one can find by listening to the signals that the sound of the reconstructed signals relying on musical priors is often "richer" than the one obtained with the approach used in [8]. Fig. 4 shows the significance maps of the selected atoms (MMSE estimates) for the *Mozart* signal, in the case $SNR_{in} = 10$dB. Again, the partials of the notes are better discriminated using musical priors, especially in low frequencies.

Indicator Variable Prior Set-up: The value p_{ton} in Eq. (7) has an effect on the above-mentioned artifacts produced by our model in low-input SNRs conditions. For instance, setting p_{ton} to 0.99 instead of 0.9 in the case of the *Glockenspiel* signal allows reducing the artifacts for $SNR_{in} = 10$dB. However, our experiments show that indicator variables corresponding to atoms that do not belong to the chord must not be set to 0. Setting p_{ton} to 1 results in reconstructed signals of very "poor" sound, as it can be assessed by listening tests. Output SNRs are also degraded. Setting $p_{ton} < 1$ allows taking into account imperfections of the chromagram given as input of the hybrid model (temporal imperfections due to windowing, discrepancy between the ideal model and reality, *etc.*).

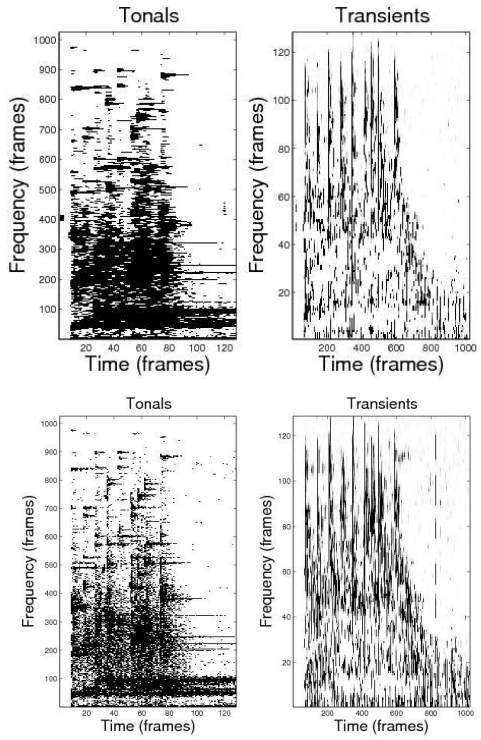

Figure 3. Significance maps of the tonal and transient bases (MAP estimates) for the *Glockenspiel* excerpt, case WN. Top: approach [8]. Bottom: proposed approach.

Impact of Tuning: Integrating tuning information in the model does not lead to improvement in terms of output SNR values, but yields to estimated significance maps that are more coherent with our model. Indeed, the "tubes" depend on the tuning and, even in case of "bad" tuning, the atoms are selected within the correct frequency regions.

Impact of Harmonics: We did not find any improvement when adding harmonics in our model. This may be partially explained by the fact that, in the polyphonic case, the contribution of a large part of the first 6 higher harmonics of a note is already taken into account in the significance map by the other notes. For instance, let us consider C major chord (C-E-G). The C note generates harmonics E and G. E and G are thus both actual played notes and harmonics. Their contribution is already partially taken into account in the significance map in the case of the model "without harmonics" already belong to the chord.

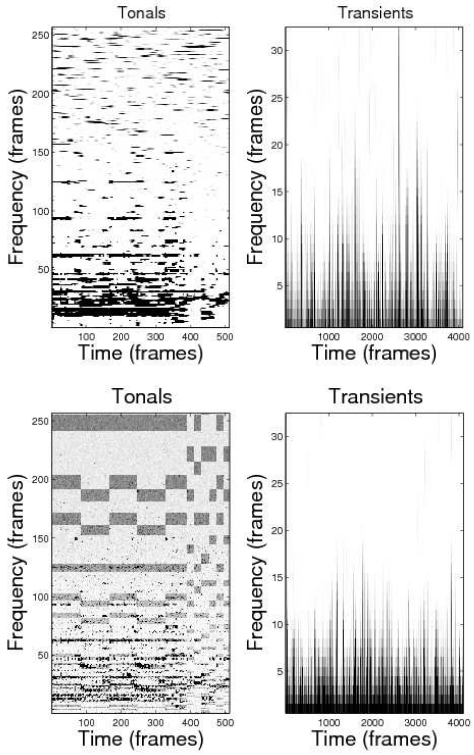

Figure 4. Significance maps of each basis (MMSE estimates) for the *Mozart* excerpt, case $SNR_{in} = 10$dB. Top: approach [8]. Bottom: proposed approach.

5. CONCLUSION AND FUTURE WORKS

In this article, we have presented a method for sparse decomposition of audio signals of music on overcomplete dictionaries made as union of two MDCT bases. We rely on previous works that consider dependencies between significant coefficients of the expansion. The originality of our approach is that we incorporate musical priors in the model. Our approach provides results whose quality corresponds to state-of-the-art approaches for the denoising task, and which show that our model that is adequate to fairly represent audio signals of music. The main contribution of the article is to show that the musical prior based on musical knowledge performs as well as more sophisticate prior as HMM and appears to be more "natural". The significance map corresponding to the tonal layer is coherent with the intrinsic content of music audio.

Future work will concentrate on fully integrating in the model chord estimation in an interactive fashion. The chromagram could be updated with the other parameters during MCMC inference in order to possibly improve the chord estimation. The prior we propose has a great potential of improvement in the future (for example, by using a time segmentation, a larger chord lexicon etc.)

As far as we know, the introduction of musical priors in hybrid models for spare decomposition is novel. The use of mid-level representation of audio - such as the chromagram, as proposed in this paper - or scores, if available, could be extended to many applications such as denoising, source separation, compression, coding and many others. Usually, only physical and mathematical criteria are taken into account. We believe that the use of musical information opens new interesting perspectives.

6. ACKNOWLEDGMENT

The authors would like to thank C. Févotte and all the authors of [8] for providing their source code.

7. REFERENCES

[1] S. Chen, D. David L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scient. Comp.*, 20, 1998.

[2] L. Daudet. Audio sparse decompositions in parallel Let the greed be shared ! . *IEEE Trans. Sig. Proc.*, 27, 2010.

[3] L. Daudet, S. Molla, and B. Torrésani. Towards a hybrid audio coder. In *WAA*, 2004.

[4] L. Daudet and B. Torrésani. Hybrid representations for audiophonic signal encoding. *Sig. Proc. J.*, 82, 2002.

[5] M.E. Davies and L. Daudet. Sparse audio representations using the MCLT. *Sig. Proc. J.*, 86(3), 2006.

[6] C. Févotte, l. Daudet, S. J. Godsill, and B. Torrésani. Sparse regression with structured priors : Application to audio denoising. In *ICASSP*, 2006.

[7] C. Févotte and S.J. Godsill. A Bayesian approach for blind separation of sparse sources. *IEEE Trans. Sp. Audio Proc.*, 14(6), 2006.

[8] C. Févotte, B. Torrésani, L. Daudet, and S.J. Godsill. Sparse Linear Regression With Structured Priors and Application to Denoising of Musical Audio. *IEEE Trans. Sp. Audio Proc.*, 16,, 2008.

[9] M.A.T. Figueiredo. Adaptive Sparseness for Supervised Learning. *IEEE Trans. Patt. Anal. Mac. Intel.*, 25, 2003.

[10] T. Fujishima. Real-time chord recognition of musical sound: a system using common lisp music. In *ICMC*, 1999.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 6, 1984.

[12] K.N. Hamdy, M. Ali, and A.H. Tewfik. Low Bit Rate High Quality Audio Coding With Combined Harmonic And Wavelet Representations. In *ICASSP*, 1996.

[13] M. Kowalski. Sparse regression using mixed norms. *Appl. Comp. Harm. Anal.*, 27, 2009.

[14] M. Kowalski and B. Torrésani. Random models for sparse signals expansion on unions of bases with application to audio signals. *IEEE Trans. Sig. Proc.*, 56,, 2008.

[15] S. Mallat. *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[16] S. Mallat and Z. Zhang. Matching Pursuit With Time-Frequency Dictionaries. *IEEE Trans. Sig. Proc.*, 41, 1993.

[17] S. Molla and B. Torésani. An hybrid audio scheme using hidden Markov models of waveforms. *Appl. Comp. Harm. Anal.*, 18, 2005.

[18] H. Papadopoulos and G. Peeters. Joint estimation of chords and downbeats. *IEEE Trans. Audio, Speech, Lang. Proc.*, 19(1), 2011.

[19] G. Peeters. Musical key estimation of audio signal based on HMM modeling of chroma vectors. In *DAFx*, 2006.

[20] E. Ravelli, G. Richard, and L. Daudet. Union of MDCT Bases for Audio Coding. *IEEE Trans. Audio, Speech, Lang. Proc.*, 16, 2008.

[21] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Serie B*, 58(1):267–288, 1996.

[22] T.S. Verma and T.H.Y. Meng. Extending Spectral Modeling Synthesis with Transient Modeling Synthesis. *Comp. Mus. J.*, 24, 2000.

[23] G.H. Wakefield. Mathematical representation of joint time-chroma distribution. In *ASPAAI*, 1999.

[24] C. Yeh, A. Roebel, and X. Rodet. Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals. *IEEE Trans. Audio, Speech, Lang. Proc.*, 18-6, 2010.