Décomposition parcimonieuse structurée des signaux audio de musique guidée par l'information experte

Hélène PAPADOPOULOS, Matthieu KOWALSKI

Laboratoire des Signaux et Systèmes
3, rue Joliot-Curie, 91192 Gif-sur-Yvette, France
helene.papadopoulos@lss.supelec.fr, matthieu.kowalski@lss.supelec.fr

Résumé – La construction, dans un cadre bayésien, d'*a priori* musicaux pour les décompositions parcimonieuses des signaux de musique est étudiée. Ces *a priori* reposent sur la connaissance musicale obtenue sur le signal, résumés dans des vecteurs de chroma contenant les douze notes de la gamme chromatique. L'originalité principale de ce travail est l'intégration de la connaissance musicale « experte » pour les décompositions parcimonieuses, plutôt que d'a priori « physique », comme les persistances en temps ou en fréquence, classiquement utilisés. Les décompositions obtenues sont comparées à l'état de l'art utilisant des approches physiques sur un problème de débruitage d'un bruit blanc gaussien. Les résultats obtenus en terme de rapport signal sur bruit sont équivalents. Les cartes de signifiance des coefficients temps-fréquence font apparaître plus clairement les structures harmoniques attendues.

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

On s'intéresse ici au problème de l'approximation parcimonieuse de signaux audio de musique dans des dictionnaires temps-fréquence appropriés. Une représentation d'un signal est dite parcimonieuse lorsque celui-ci peut être représenté par une combinaison linéaire des éléments du dictionnaire (appelés atomes) et que seuls un petit nombre des coefficients de la décomposition sont significativement non nuls. Ce problème a de nombreuses applications, notamment pour la compression [1], le débruitage [8] ou la séparation de sources [2].

Une particularité des signaux audio de musique est que, souvent, plusieurs types de composantes se superposent comme par exemple des composantes tonales (les partiels des notes, bien modélisés par des sommes de sinusoïdes variant lentement en amplitude et en fréquence) et transitoires (les attaques des notes, événements bien localisés en temps), ainsi que représenté figure 1. Ces diverses composantes peuvent avoir des comportements significativement différents et ne peuvent pas être représentées dans une même base. Par exemple les transitoires qui varient rapidement requièrent une fenêtre d'analyse courte tandis que les composantes tonales exigent des fenêtres longues. Les modèles *hybrides* [5] permettent une représentation simultanée des différentes couches. Dans ce cadre, on considère l'approximation d'un signal sur un dictionnaire construit comme l'union de deux bases MDCT (Modified Discrete Cosine Transform) avec des résolutions temps-fréquence différentes pour chaque couche. La décomposition d'un signal sur un tel dictionnaire surcomplet n'est pas unique. La parcimonie peut être alors utilisée comme critère de sélection.

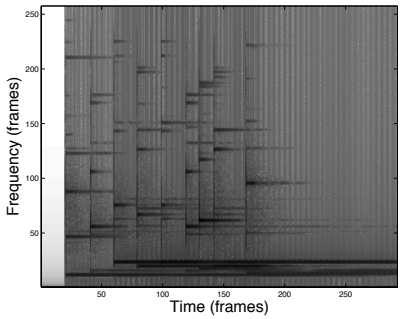

FIGURE 1 – Représentation temps-fréquence de l'extrait *Glockenspiel*.

On considère un modèle en trois couches de la forme $signal = tonals + transients + residual$ afin de décomposer un signal de musique $x \in \mathbb{R}^N$. La modélisation du signal repose à la fois sur un modèle pour les coefficients de la décomposition et un modèle de *cartes de signifiance* qui décrivent les positions des coefficients significatifs dans l'espace des indices.

Comme on l'a vu, les signaux de musiques sont très structurés et, idéalement, ces structures devraient se refléter dans la décomposition, de sorte que les coefficients aient une interprétabilité physique ou musicale dans la perspective de son analyse. On cherche alors une approximation qui, en plus d'être parci-

monieuse, soit aussi structurée en prenant en compte les dépendances entre les coefficients. Ces structures peuvent être modélisées directement sur les coefficients, cependant il peut être pratique de les modéliser directement sur les indices temps-fréquences directement plutôt que sur les coefficients eux-mêmes.

On se limitera ici à une tache de débruitage sur des signaux de musiques, cette application servant de « preuve de concept » pour illustrer les nouveaux *a priori* musicaux présentés. Le modèle proposé s'inspire d'un modèle bayesien précédemment présenté par Févotte *et al.* dans [8], la différence essentielle étant la manière de modéliser les dépendances entre coefficients. En effet, dans cette contribution, le contenu musical sert directement à construire les *a priori* au lieu d'utiliser des intuitions physiques.

Comme proposé dans [9], la progression des accords ([10] [11]) fournit une information qui peut-être directement utilisé pour construire un *a priori* musical pour la couche tonale. On présente ici quelques extensions du modèle initial, par exemple l'utilisation des chroma [12] comme alternative à la progression des accords pour modéliser la couche tonale. Un modèle pour la couche transitoire utilisant l'information de position des « beats » [14] est aussi proposé. On pourra trouver plus de détails dans l'article [4].

2 Modèle

2.1 Représentation parcimonieuse

On considère un dictionnaire hybride D construit comme l'union de deux bases orthonormales $V = \{v_n, n = 1, \ldots, N\}$ et $U = \{u_m, m = 1, \ldots, N\}$, avec différentes résolutions temps-fréquences afin de représenter les couches tonales et transitoires du signal. Ici n and m sont des indices temps-fréquence, aussi notés $n = (q, \nu) \in [1, n_{ton}] \times [1, \ell_{ton}]$ or $m = (q, \nu) \in [1, n_{tran}] \times [1, \ell_{tran}]$ dans la suite. Le modèle est alors de la forme

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

où les cartes de signifiance Λ et Δ sont des petits sous-ensembles de $I = \{1, \ldots, N\}$ qui décrivent la position des coefficients significatifs de la décomposition, et r un résidu n'acceptant pas de décomposition parcimonieuse dans le dictionnaire.

Le modèle hybride est défini par un modèle de probabilités discrètes sur les cartes de signifiance et un modèle de probabilités sur les coefficients, conditionnellement aux cartes de signifiance. On re-écrit (1) comme :

$$x = \sum_{n \in I} \gamma_{ton,n} \alpha_n v_n + \sum_{m \in I} \gamma_{tran,m} \beta_m u_m + r \,. \tag{2}$$

où les $\gamma_{ton,n}$ et $\gamma_{tran,m} \in 0, 1$ sont des variables aléatoires indicatrices binaires pour les couches *tonales* et *transitoires* respectivement.

On construit l' *a priori* proposé sur un modèle Bernoulli-Gaussien :

$$p(\alpha_n | \gamma_{ton,n}, \sigma_{ton,n}) = (1 - \gamma_{ton,n}) \delta_0(\alpha_n) + \gamma_{ton,n} \mathcal{N}(\alpha_n | 0, \sigma^2_{ton,n}) \,. \tag{3}$$

où δ_0 est la fonction de Dirac et où on donne un modèle *a priori* Gamma-inverse aux variances $\sigma_{ton,n}$. On utilise un modèle similaire pour les coefficients β et γ_{trans}.

2.2 *A priori* pour les cartes de signifiance

On donne aux cartes de signifiance Λ et Δ des *a priori* structurés, en utilisant de l'information de contenu musical (information sémantique) que l'on peut extraire du signal en utilisant des algorithmes classiques issus de la communauté *Music Information Retrieval*.

2.2.1 Couche tonale

Pour modéliser la carte de signifiance correspondant à la couche tonale, on peut utiliser de l'information de contenu harmonique. Idéalement on aimerait indiquer précisément quelles sont les notes jouées à chaque instant, par exemple en se basant sur une partition. Or, en général, on ne dispose pas d'une telle transcription symbolique. De plus, une telle transcription serait incomplète puisqu'elle n'indiquerait pas par exemple quelles sont les harmoniques produites par les instruments. On propose de se baser à la place sur une représentation "moyen-niveau" du signal, qui, sans être une transcription exacte, caractérise son contenu harmonique. Ainsi, on peut utiliser la progression des accords (voir [9]). On décrit ci-dessous une alternative, où on utilise directement une représentation par *chromagramme* [12, 13]. Un vecteur de chroma est un vecteur à 12 dimensions qui représente, à un instant donné, l'importance des différentes hauteurs des 12 demi-tons de la gamme chromatique $(C, C\#, \ldots, B)$. On appelle *chromagramme* la succession de vecteurs de chroma calculés au cours du temps.

Afin de sélectionner les atomes de la base MDCT correspondants au contenu harmonique du signal, on commence par construire une relation entre les positions des coefficients MDCT et les 12 demi-tons de la gamme chromatique. Étant donné un indice q, on note $\{a_k\}_{k=1,\ldots,12}$ l'amplitude de chaque bin $\{p^{chroma}_k\}_{k=1\ldots12}$ du vecteur de chroma normalisé calculé . On note aussi $\{p^{MDCT}_\nu\}_{\nu=1,\ldots,\ell_{ton}}$ les classes de demi tons pour chaque indice fréquentiel de la base MDCT. En supposant un accordage parfait à 440 Hz, un coefficient MDCT à la fréquence ν est converti en chroma p^{MDCT}_ν par :

$$p^{MDCT}_\nu = \left(12 \log_2 \frac{\nu}{440} + 69\right) \pmod{12} \,. \tag{4}$$

Remarquons qu'un chroma donné correspond à plusieurs indices fréquentiels consécutif (voir figure 2). De plus, en raison de l'échelle logarithmique utilisé dans la musique tonale occidentale, le nombre d'indices successifs correspondant à un chroma donné augmente avec les fréquences.

Le modèle chroma utilisé pour les indicatrices de la couche tonale est alors

$$P_\Lambda\{\gamma_{ton,(q,\nu)} = 1\} = \left\{ \begin{array}{ll} a_k & \text{if } \exists k \in [1,12] \mid p_\nu^{MDCT} = p_k^{chroma} \end{array} \right. . \quad (5)$$

On peut remplacer $\{p_k^{chroma}\}_{k=1..12}$ par un vecteur d'accord estimé sur le morceau de musique [11]. La figure 2 montre les cartes de signifiance pour la couche tonale, obtenus en se basant sur la progression des accords (gauche) et sur le chromagramme (droite) correspondant aux premières secondes du *Quatuor Op.127* de Beethoven.

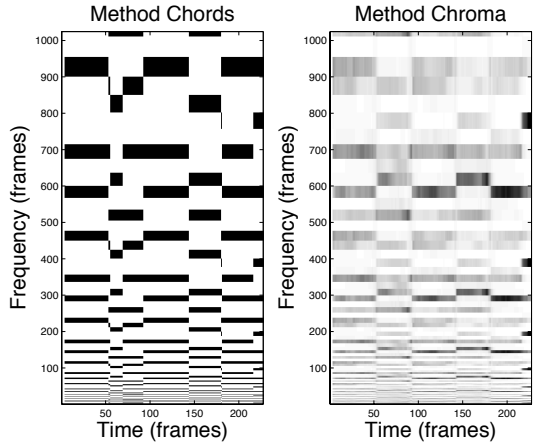

FIGURE 2 – Cartes de signifiance structurées pour la couche tonale, début du *Quatuor Op.127* de Beethoven.

2.2.2 Couche transitoire

Pour la couche transitoire, un modèle peut être construit en utilisant de l'information sur la structure temporelle du morceau analysé. L'unité de mesure étant les temps, dis aussi « beats », on se base sur l'idée que, dans une pièce de musique, la plupart des sons transitoires ont lieu sur les temps ou sur une subdivision des temps. Ainsi, par exemple, les sons percussifs sont souvent utilisés pour souligner la structure métrique ; dans un quatuor, les changements de coups d'archets ont lieu sur les changements de notes, et donc sont liés à la structure métrique.

On note $\{b_k\}_{k=1,\ldots,N_b}$ les N_b positions (en trames) des beats estimés en utilisant un algorithme de "beat tracking" [3] (on peut aussi considérer une subdivision des temps, par exemple les double-croches). On donne aux variables indicatrices de la couche transitoire les probabilités suivantes :

$$\forall \nu = 1, \ldots, \ell_{tran} \quad P_\Delta\{\gamma_{tran,(q,\nu)} = 1\} = \left\{ \begin{array}{ll} p_{tran} & \text{si } \exists k \in [1, N_b] \mid q = k \\ 1 - p_{tran} & \text{sinon} \end{array} \right. , \quad (6)$$

où $0 \leq p_{tran} \leq 1$. En pratique la valeur p_{tran} sera choisie proche de 1 de manière à donner une probabilité *a priori* élevée aux atomes correspondant à des positions de temps. La figure 3 illustre la carte de signifiance pour le signal *Glockenspiel* du test-set d'évaluation. À chaque position (temporelle) de beat, toutes les fréquences sont sélectionnées, résultant en des lignes verticales. On peut noter que la durée entre deux lignes consécutives n'est pas fixe, car il peut y avoir des variations de tempo.

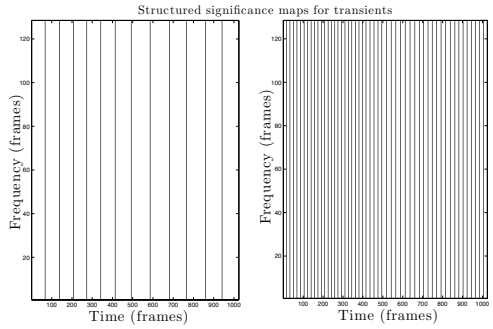

FIGURE 3 – Cartes de signifiance structurées pour la couche transitoire pour le signal *Glockenspiel*. Gauche : en considérant la position des temps. Droite : en considérant la position des double-croches.

3 Inférence

Les paramètres du modèle sont ensuite estimés par une méthode MCMC classique utilisant un échantillonneur de Gibbs.

4 Résultats

Base de données et mesures d'évaluation : On évalue l'approche proposée, en terme de rapport signal à bruit (SNR) et de parcimonie sur 5 signaux de musique de style et d'instrumentation variés : un signal monophonique *Glockenspiel*, deux extraits polyphoniques complexes des Beatles, *Misery* et *Love Me Do*, contenant de la voix et des percussions, le début du *Quatuor à cordes Op. 127* de Beethoven et le début de la *Sonate pour piano KV310* de Mozart.

Paramètres : La taille des deus bases MDCT est fixée à 1024 échantillons pour la couche tonale et 128 échantillons pour la couche transitoire, à un taux d'échantillonnage de 44100Hz. Les estimateurs MMSE et MAP des paramètres sont calculés en prenant la moyenne des derniers 100 échantillons de l'échantillonneur de Gibbs, après 500 itérations.

En raison du manque de place, on ne présente ici, Table 1, qu'une petite partie des résultats obtenus . Le lecteur intéressé trouvera une discussion détaillée et de nombreux exemples dans [4]. D'un point de vue débruitage, on obtient des résultat similaires avec trois méthodes (Prior musical pour la couche tonale, pour la couche tonale et transitoire et *F2008*)).

Cependant, les cartes de signifiance obtenues en utilisant les *a priori* musicaux font plus clairement apparaître les structures attendues (meilleure résolution des partiels, en particulier dans les basses fréquences, et attaques des notes plus nettes) que lorsqu'on utilise une approche basée sur des *a priori* physiques.

SNR_{in}	WN			0			10			20		
Method	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*
Gl.	71.35	71.49	70.22	14.13	13.86	15.74	21.37	20.98	22.45	28.59	28.24	29.22
Mi.	42.73	32.72	44.41	7.03	7.04	6.9	13.35	13.34	13.29	20.89	20.60	21.08
Lo.	28.32	27.14	29.61	6.80	6.72	6.77	12.85	12.95	12.72	19.27	19.72	19.35
Be.	54.72	35.97	54.64	8.63	8.65	7.71	14.52	14.54	14.03	22.05	21.49	21.99
Mo.	62.33	57.07	60.96	9.45	9.37	8.97	16.28	16.12	15.94	23.96	23.73	23.88

TABLE 1 – Comparaison des résultats obtenus avec l'approche proposée utilisant un prior musical pour la couche tonale seulement (cas *A*), pour les couches tonale et transitoire (cas *A + tr*) et [8] (cas *F2008*).

Ceci est illustré par la figure 4 où l'on présente les cartes de signifiance obtenues par le modèle proposé par Févotte *et al.* et la méthode chroma, pour un extrait de Mozart, auquel un bruit blanc gaussien a été ajouté (SNR de 10 dB).

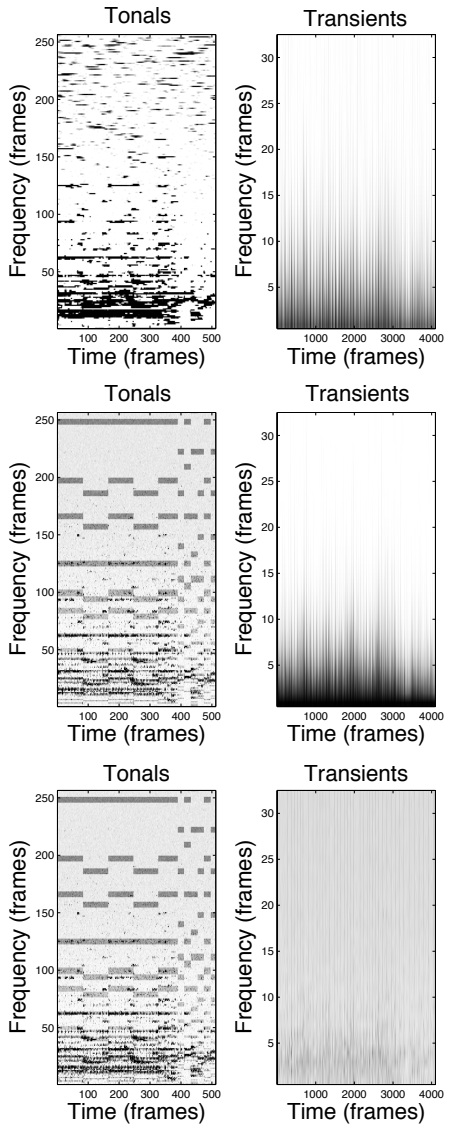

FIGURE 4 – Carte de signifiance pour chaque base (estimation au sens MMSE) pour l'extrait *sonate de Mozart*, et un rapport signal sur bruit de 10 dB en entrée. Haut : approche [8] ; Milieu : approche avec *a priori* musical (méthode chroma) pour la couche tonale seule. Bas : approche avec *a priori* musical pour la couche tonale (méthode chroma) et la couche transitoire.

5 Conclusion

La principale contribution de ce travail est de montrer que les *a priori* basés exclusivement sur la connaissance musicale sont une alternative viable aux modèles plus classiquement utilisés tels que les HMM, tout en apparaissant « naturels ». D'un point de vue performance en débruitage, les résultats obtenus sont comparables à l'état-de-l'art. Cependant, les cartes de signifiance obtenues font plus clairement apparaître les structures attendues, ces structures étant directement prises en compte dans le modèle. Les travaux futurs consisteront à tester ce type d'approche sur des problèmes plus difficiles comme la séparation de sources.

Références

[1] E. Ravelli, G. Richard and L. Daudet *Union of MDCT Bases for Audio Coding*, IEEE Trans. on Au. Sp., and Lang. Proc., 2008.

[2] L. Benaroya and F. Bimbot and R. Gribonval, *Audio source separation with a single sensor* IEEE Trans. on Au. Sp., and Lang. Proc., 2006.

[3] G. Peeters and H. Papadopoulos, *Simultaneous beat and downbeat-tracking using a probabilistic framework : theory and large-scale evaluation* IEEE Trans. on Au. Sp., and Lang. Proc., 2011.

[4] H. Papadopoulos and M. Kowalski *Sparse and structured decomposition of audio signals on hybrid dictionaries using musical priors*, Accepted in JASA.

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

[6] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval M. E. Davies, *Sparse Representations in Audio and Music : from Coding to Source Separation*, Proc. IEEE vol. 98, no 6, pp 995-12005, 2010

[7] P. Wolfe, S. Godsill, and W.-J. Ng . *Bayesian variable selection and regularization for time-frequency surface estimation*, J. of the Royal Stati. Soc. Serie B, 2004.

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

[9] H. Papadopoulos and M. Kowalski. *Sparse Signal Decomposition on Hybrid Dictionaries Using Musical Priors*, ISMIR 2011.

[10] A. Sheh and D. Ellis. *Chord segmentation and recognition using EM-trained HMM*, ISMIR 2003.

[11] H. Papadopoulos, and G. Peeters. *Joint estimation of chords and downbeats*, IEEE Trans. on Au. Sp., and Lang. Proc., 2011.

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

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

[14] A. Klapuri, A. Eronen, and J. Astola. *Analysis of the meter of acoustic musical signals*, IEEE Trans. on Au. Sp., and Lang. Proc. 2006.