Analyse des structures harmoniques dans les signaux audio : modéliser les variations de hauteur et d'enveloppe spectrale

Benoit FUENTES, Roland BADEAU, Gaël RICHARD

Institut Télécom,Télécom ParisTech, CNRS LTCI
37-39, rue Dareau - 75014 Paris - France
tel : +33 (0)1 45 81 77 77 - fax : +33 (0)1 45 89 79 06
Benoit.Fuentes@telecom-paristech.fr, Roland.Badeau@telecom-paristech.fr,
Gael.Richard@telecom-paristech.fr

Résumé – De nombreuses méthodes d'analyse et de décomposition intelligente des représentations temps-fréquence de signaux musicaux ont été développées ces derniers temps. Cependant, les outils utilisés ne sont pas forcément adaptés aux signaux polyphoniques dont les notes présentent des variations continues de fréquence fondamentale et d'enveloppe spectrale. Nous proposons un nouveau modèle d'analyse des structures harmoniques, permettant de considérer conjointement ces deux types de variations. Chaque note dans une transformée à Q constant est modélisée localement comme une somme pondérée de spectres harmoniques à bande étroite, et les paramètres du modèle sont estimés grâce à l'analyse probabiliste en composantes latentes. L'algorithme a été testé dans une tâche d'estimation de hauteur simple et les très bons résultats obtenus mettent en valeur la fiabilité et la robustesse du modèle proposé.

Abstract – Recently, numerous techniques have been developed for smart decomposition of time-frequency representations of audio. However, these methods are not necessarily well adapted for real music signals where each note can present continuous variations of both pitch and spectral envelope. A new model for the analysis of harmonic structures, which allows considering simultaneously those two kinds of variations, is introduced. Each note in a constant-Q transform is locally modeled as a weighted sum of narrowband harmonic spectra, and the model parameters are estimated by means of Probabilistic Latent Component Analysis. The algorithm has been tested in a task of monopitch estimation, and the very good results highlight the reliability and the robustness of the model.

1 Introduction

Récemment, de nombreuses techniques ont été proposées pour décomposer une représentation temps-fréquence d'un signal audio en éléments significatifs positifs ou nuls. Dans le cadre du traitement du signal musical, de telles décompositions peuvent servir dans de nombreuses applications, comme l'estimation de fréquences fondamentales multiples, la transcription automatique ou encore la séparation de sources. La méthode la plus utilisée pour effectuer ce type de décomposition est sans aucun doute la factorisation en matrices nonnégatives (*Non-negative Matrix Factorization* ou NMF), initialement imaginée pour l'analyse des images ou la factorisation de données [6], puis appliquée à l'analyse des signaux audio [8]. Elle consiste à décomposer chaque colonne du spectrogramme d'un signal comme une somme pondérée de spectres de base (ou atomes, noyaux). Cependant, la NMF ne prend pas en compte les spécificités propres aux signaux audio (on peut penser au caractère harmonique de nombreux événements musicaux, à des phénomènes de changement continu de fréquence fondamentale comme le *vibrato* ou encore à des variations temporelles d'enveloppe spectrale pour une même note de musique) et de nombreuses variantes ont donc été développées. Ainsi, dans [10], une contrainte d'harmonicité sur les atomes est ajoutée et on autorise un nombre fixe d'enveloppes spectrales différentes pour chaque pitch. Romain Hennequin propose quant à lui dans [4] un modèle dans lequel chaque atome est filtré par un filtre ARMA permettant ainsi de décrire les fortes variations de formes spectrales. Dans [5], le principe d'invariance par translation est exploité, ce qui permet avec peu d'atomes de modéliser les changements de fréquence fondamentale pour un même instrument, mais pas de prendre en compte les variations d'enveloppe spectrale au cours du temps.

Le principal inconvénient des techniques mentionnées ci-dessus est qu'elles ne permettent pas facilement de modéliser les événements musicaux dont la hauteur et l'enveloppe spectrale varient conjointement au cours du temps. Le modèle HALCA (*Harmonic Adaptive Latent Component Analysis* ou analyse adaptative harmonique en composantes latentes) que nous proposons ici [1] permet de dépasser cette limitation. Il ne s'inscrit pas dans le cadre de la NMF mais de l'analyse probabiliste en composantes latentes (*Probabilistic Latent Component Analysis* ou PLCA) [7] et de son extension avec invariance par translation [9]. Cette technique est équivalente à la NMF, mais présente un formalisme probabiliste et offre l'avantage d'autoriser n'importe quel modèle de décomposition des données et

1. Ces travaux ont été en partie réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.

d'ajouter des a priori appropriés sur les paramètres.

Après avoir présenté deux outils importants (partie 2), nous introduisons le modèle HALCA (partie 3) ainsi qu'un a priori intéressant sur les paramètres du modèle (partie 4). Les parties 5 et 6 sont consacrées à l'évaluation et à la conclusion.

2 Les outils utilisés

Transformée à Q constant. La méthode d'analyse proposée s'effectue dans le domaine temps-fréquence. Pour cela, on applique en premier lieu une transformée à facteur de qualité Q constant (*Constant-Q Transform* ou CQT) [2] au signal audio à analyser. Il s'agit d'une représentation temps-fréquence d'un signal temporel dont la résolution fréquentielle est inversement proportionnelle à la fréquence d'analyse et dont l'échelle des fréquences est logarithmique. Grâce à cette dernière particularité, l'espacement entre deux partiels donnés d'une note harmonique reste identique quel que soit la hauteur. Un changement de hauteur peut alors être considéré comme une translation en fréquence des partiels, ce qui sera exploité dans le modèle HALCA.

Introduction à la PLCA. La PLCA est un outil probabiliste pour l'analyse de données positives (dans notre cas, les coefficients qui constituent la CQT V_{ft} d'un signal). Les observations V_{ft} sont considérées comme l'histogramme du tirage de N variables aléatoires indépendantes (f_n, t_n) (correspondant aux points fréquence-temps) distribuées selon la loi de probabilité discrète $P(f,t)$. La manière de modéliser $P(f,t)$ définit la façon dont les données seront décomposées. Ainsi, dans le modèle de base on introduit une variable cachée z et on pose $P(f,t) = \sum_z P(z)P(f|z)P(t|z) = \sum_z P(z,t)P(f|z)$, f et t étant supposés indépendants conditionnellement à z. $P(f|z)$ correspond alors au spectre des différents atomes de base et $P(z,t)$ aux activations temporelles de chaque atome. Dans le modèle de *Shift-Invariant* PLCA [9], de même que dans le modèle HALCA, le calcul de $P(f,t)$ résulte de la convolution de plusieurs densités de probabilité comme nous le verrons dans la section suivante. L'algorithme Espérance-Maximisation (EM) permet ensuite d'estimer l'ensemble des paramètres du modèle de $P(f,t)$.

3 Le modèle HALCA

On suppose que la CQT d'un signal polyphonique est constituée de la somme pondérée de plusieurs signaux monodiques (appelés « canaux » et désignés par la variable cachée c), soit $P(f,t) = \sum_c P(c)P(f,t|c)$. Chaque colonne de $P(f,t|c)$ représente alors le spectre d'une note de musique. Similairement à [10], afin de prendre en compte le caractère harmonique ainsi que la forme de l'enveloppe d'un tel spectre, nous le modélisons comme une somme pondérée de Z spectres harmoniques à bande étroite, appelés noyaux et notés $P_{Kh}(\mu|z)$, convoluée par une impulsion $P_{Ih}(i|t,c)$. Les noyaux possèdent tous la

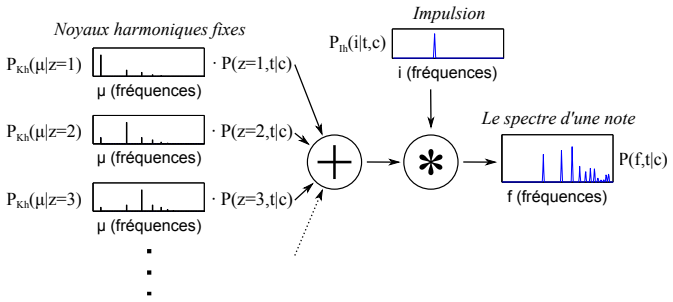

FIGURE 1 – Modèle de spectre harmonique pour un canal $c < C$ donné au temps t. Chaque noyau a la majorité de son énergie concentrée sur une harmonique donnée, multiple d'une fréquence fondamentale de référence, et le reste de l'énergie est partagé entre les harmoniques adjacentes.

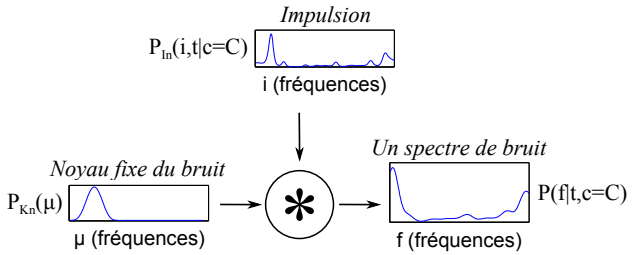

FIGURE 2 – Modèle de spectre de bruit (canal $c = C$) au temps t.

même fréquence fondamentale de référence mais ont leur énergie concentrée à différentes bandes de fréquence. L'impulsion est une distribution unimodale dont la valeur du mode tient compte de la hauteur de la note. Les coefficients de pondérations appliqués aux noyaux sont notés $P(z,t|c)$. De plus, afin de considérer la présence de bruit coloré, nous lui réservons le dernier canal C : $P(f,t|c=C)$ est modélisée comme la convolution d'une fenêtre régulière à bande étroite notée $P_{Kn}(\mu)$ et de l'impulsion $P_{In}(i,t|c=C)$ (le terme impulsion n'est plus approprié ici, mais nous le gardons par souci de cohérence). Les modèles de notes et de bruit sont illustrés dans les figures 1 et 2. Il est à noter que les noyaux $P_{Kh}(\mu|z)$ et $P_{Kn}(\mu)$ sont fixés dans l'algorithme et ne sont donc pas à estimer. Finalement, le modèle HALCA peut s'écrire comme :

$$P(f,t) = \sum_{c<C} P(c) \sum_{i,z} P(z,t|c) P_{Kh}(f-i|z) P_{Ih}(i|t,c) + P(C) \sum_i P_{Kn}(f-i) P_{In}(i,t|c=C). \quad (1)$$

Comme décrit dans [7], l'algorithme EM permet de trouver des règles de mise à jour pour les paramètres, telles qu'à chaque itération la log-vraisemblance des observations étant donné le modèle, définie par $L = \sum_{f,t} V_{ft} \ln(P(f,t))$, augmente :

$$P(c) \propto \sum_{z,f,i,t} V_{ft} P(i,z,c|f,t) \text{ pour } c < C \quad (2)$$

$$P(z,t|c) \propto \sum_{f,i} V_{ft} P(i,z,c|f,t) \text{ pour } c < C \quad (3)$$

$$P_{Ih}(i|t,c) \propto \sum_{f,z} V_{ft} P(i,z,c|f,t) \text{ pour } c < C \quad (4)$$

$$P(C) \propto \sum_{z,f,i,t} V_{ft} P(i,z,C|f,t) \quad (5)$$

$$P_{In}(i,t|c=C) \propto \sum_{f,z} V_{ft} P(i,z,C|f,t) \quad (6)$$

avec, pour $c < C$,

$$P(i,z,c|f,t) = \frac{P(c)P(z,t|c)P_{Kh}(f-i|z)P_{Ih}(i|t,c)}{P(f,t)} \quad (7)$$

et

$$P(i,z,C|f,t) = \frac{P(C)P_{Kn}(f-i)P_{In}(i,t|c=C)}{P(f,t)}, \quad (8)$$

$P(f,t)$ étant définit par l'équation (1). L'algorithme consiste alors à initialiser l'ensemble des paramètres, puis à itérer le nombre de fois voulu les équations (7) et (8), les différentes mises à jours (équations (2),(3),(4),(5) et (6)), et enfin la normalisation des paramètres pour que les probabilités somment à 1. Idéalement, au temps t et pour un canal $c < C$ donné, l'estimation de $P_{Ih}(i|t,c)$ est une densité de probabilité unimodale, la fréquence fondamentale de la note pouvant être directement déduite de la valeur du mode. Cependant, en l'absence de contrainte, ce paramètre ne converge pas nécessairement vers la solution souhaitée : dans la pratique on peut observer l'existence de maximas pour les valeurs de i correspondant à la hauteur exacte de la note ainsi qu'à toutes ses harmoniques supérieures. Afin de garder uniquement le maximum de plus basse fréquence, nous introduisons un a priori de minimum de variance asymétrique, tel que décrit dans la prochaine partie.

4 Ajout d'un a priori de minimum de variance asymétrique

Pour un canal $c < C$ et un temps donné t, soit $\boldsymbol{\theta}^{t,c}$ le vecteur de coefficients $\theta_i^{t,c} = P_{Ih}(i|t,c)$. Afin de contraindre $\boldsymbol{\theta}^{t,c}$ à être unimodal, nous utilisons un a priori de minimum de variance asymétrique qui force ses valeurs de variance et de moyenne à être faibles. Cet a priori repose sur une mesure adéquate, dépendante d'un paramètre α qui définit la force de l'asymétrie :

$$\begin{aligned} \operatorname{avar}_\alpha(\boldsymbol{\theta}^{t,c}) &= \sum_i \left(e^{\alpha i} - e^{\alpha \sum_i \hat{i}\theta_i^{t,c}} \right) \theta_i^{t,c} \\ &= \left(\sum_i e^{\alpha i} \theta_i^{t,c} \right) - e^{\alpha \sum_i i\theta_i^{t,c}} \text{ car } \sum_i \theta_i^{t,c} = 1. \end{aligned} \quad (9)$$

On peut prouver, grâce à la convexité de la fonction exponentielle, que $\operatorname{avar}_\alpha(\boldsymbol{\theta}^{t,c}) \geq 0$ et que $\operatorname{avar}_\alpha(\boldsymbol{\theta}^{t,c}) = 0 \Leftrightarrow \exists\, i_0 \mid \forall i, \theta_i = 1$ si $i = i_0$ et $\theta_i = 0$ sinon. Pour contraindre chaque impulsion à avoir une faible valeur de variance asymétrique, l'a priori suivant est utilisé pour l'ensemble Λ des paramètres :

$$P(\Lambda) = \sigma \prod_{t,c<C} \exp\left(-\beta \operatorname{avar}_\alpha(\boldsymbol{\theta}^{t,c})\right) \quad (10)$$

où β est un coefficient indiquant la force de l'a priori et σ un coefficient de normalisation. Pour utiliser cet a priori, l'étape maximisation de l'algorithme EM est remplacée par une étape de maximisation a postériori, et la mise à jour des impulsions (équation (4)) est remplacée par l'équation suivante :

$$\theta_i^{t,c} = \frac{\omega_i^{t,c}}{\beta\left(e^{\alpha i} - \alpha i e^{\alpha \sum_i \hat{i}\theta_i^{t,c}}\right) + \rho^{t,c}} \quad (11)$$

où $\omega_i^{t,c} = \sum_{f,z} V_{ft} P(i,z,c|f,t)$ et $\rho^{t,c}$ est un coefficient qui permet de garantir que $\sum_i \theta_i^{t,c} = 1$. S'il n'existe pas de solution analytique à cette équation, les simulations numériques ont montré que la méthode du point fixe, décrite dans l'Algorithme 1, converge toujours vers une solution. La figure 3 illustre l'effet de l'utilisation de cet a priori, ainsi que la croissance du critère de convergence au fil des itérations.

Algorithme 1 : Méthode du point fixe

pour chaque t *et* c **faire**
- $\boldsymbol{\theta}^{t,c} \leftarrow \frac{\boldsymbol{\omega}^{t,c}}{\sum_i \omega_i^{t,c}}$;
- **répéter**
  - $m^{t,c} \leftarrow \sum_i i\theta_i^{t,c}$;
  - $\forall i, d_i \leftarrow \beta\left(e^{\alpha i} - \alpha i e^{\alpha m^{t,c}}\right)$;
  - trouver $\rho^{t,c}$ tel que $\sum_i \frac{\omega_i^{t,c}}{d_i + \rho^{t,c}} = 1$ et $\forall i, \frac{\omega_i^{t,c}}{d_i+\rho^{t,c}} \geq 0$ (il existe une unique solution qui peut être calculée à l'aide de n'importe quel algorithme de recherche de racines);
  - $\forall i, \theta_i^{t,c} \leftarrow \frac{\omega_i^{t,c}}{d_i+\rho^{t,c}}$;
- **jusqu'à** *convergence* ;

5 Evaluation

Dans un premier temps, et pour vérifier la pertinence du modèle HALCA, l'algorithme [2] a été évalué dans le cadre de l'estimation de hauteur simple. La base de données utilisée, Iowa [1], est constituée de 3307 notes isolées, provenant de divers instruments jouant sur l'ensemble de leur tessiture. Pour chaque signal, la CQT est calculée avec 36 points fréquentiels par octave, pour des fréquences allant de $27,5$ Hz à 6000 Hz, et avec un pas temporel de 20 ms. Puis, la CQT est analysée par notre algorithme avec un nombre de canaux fixé à 2 (un

2. Le code Matlab est disponible à l'adresse internet `http://perso.telecom-paristech.fr/~fuentes/shared_code/GRETSI_2011.zip`.

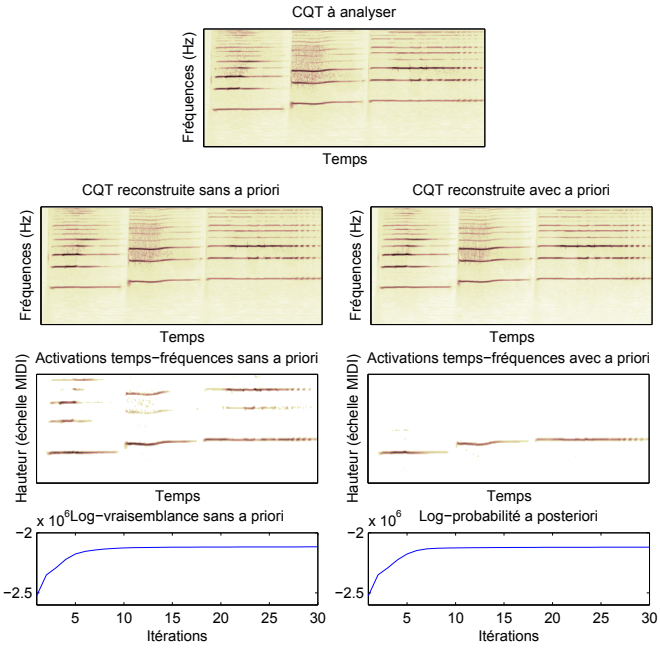

FIGURE 3 – Illustration de l'utilisation de l'a priori d'unimodalité : si les CQTs estimées restent quasiment inchangées, les activations temps-fréquence deviennent unimodales à chaque temps. Dans les deux cas, le critère de convergence croît au fil des itérations. Le signal d'entrée correspond à trois notes d'harmonica.

pour les notes et l'autre pour le bruit), un nombre de noyaux égal à 15 et la hauteur est déduite à chaque temps t grâce au maximum de la distribution $P_{Ih}(i|t, c = 1)$. La méthode est comparée à l'algorithme YIN [3], avec des fenêtres d'analyse de 100ms (nous avons utilisé le code développé par l'auteur [3]). Les résultats sont présentés dans la figure 4, et montrent que le modèle HALCA peut s'adapter de manière aveugle à tout type d'instrument (on remarquera ainsi les meilleures performances obtenues pour les bassons et hautbois, caractérisés par une très haute énergie dans les harmoniques supérieures).

6 Conclusion

Nous avons présenté un nouveau modèle de décomposition harmonique de représentations temps-fréquence qui permet de prendre en compte conjointement les variations de hauteur et d'enveloppe spectrale pour un même instrument. De plus un nouvel a priori a été introduit pour contraindre les activations temps-fréquence à être unimodales. Dans un premier temps, cet algorithme a été testé dans une tâche d'estimation de hauteur simple et a obtenu d'excellents résultats. Dans la suite, nous prévoyons d'enrichir notre modèle grâce à l'ajout de nouveaux a priori et de le tester dans une tâche d'estimation de hauteurs multiples.

3. Disponible à l'adresse suivante : `http://en.pudn.com/downloads158/sourcecode/speech/detail704391_en.html`

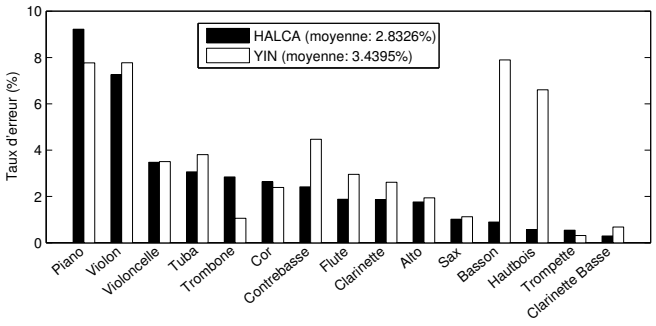

FIGURE 4 – Résultats de l'estimation de hauteur simple : taux d'erreur moyen pour chaque instrument de la base de données.

Références

[1] University of Iowa musical instrument sample database. `http://theremin.music.uiowa.edu/index.html`.

[2] J. BROWN : Calculation of a constant Q spectral transform. *JASA*, 89(1):425–434, janvier 1991.

[3] A. de CHEVEIGNE et H. KAWAHARA : Yin, a fundamental frequency estimator for speech and music. *JASA*, 111(4):1917–1930, 2002.

[4] R. HENNEQUIN, R. BADEAU et B. DAVID : NMF with time-frequency activations to model non stationary audio events. *IEEE Transactions on Audio Speech and Language Processing*, 19(4):744–753, mai 2011.

[5] M. KIM et S. CHOI : Monaural music source separation : nonnegativity, sparseness and shift-invariance. pages 647–624, Charleston, SC, USA, mars 2006.

[6] D.D. LEE et H.S. SEUNG : Learning the parts of objects by non-negativity matrix factorization. *Nature*, 401(6755):788–791, octobre 1999.

[7] M.V. SHASHANKA : *Latent variable framework for modeling and separating single-channel acoustic sources*. Thèse de doctorat, Boston University, Boston, MA, USA, août 2007.

[8] P. SMARAGDIS et J.C. BROWN : Non-negative matrix factorization for polyphonic music transcription. *Dans IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, NY, octobre 2003.

[9] P. SMARAGDIS, B. RAJ et M.V. SHASHANKA : Sparse and shift-invariant feature extraction from non-negative data. *Dans Actes d'ICASSP (International Conference on Acoustics, Speech and Signal Processing)*, pages 2069–2072, Las Vegas, Nevada, USA, avril 2008.

[10] E. VINCENT, N. BERTIN et R. BADEAU : Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio Speech and Language Processing*, 18(3):528–537, mars 2010.