



Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité " Signal et Images "

présentée et soutenue publiquement par

Romain HENNEQUIN

le 21 novembre 2011

Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale.

Modélisation des variations temporelles dans les éléments sonores.

Directeur de thèse : Bertrand DAVID

Co-encadrement de la thèse : Roland BADEAU

Jury

M. Bruno TORRÉSANI, Professeur, Université de Provence

M. Laurent DAUDET, Professeur, Institut Langevin

M. Éric MOULINES, Professeur, Télécom ParisTech

M. Paris SMARAGDIS, Assistant Professor, University of Illinois

M. Arshia CONT, Maître de conférence, IRCAM

M. Bertrand DAVID, Maître de conférence, Télécom ParisTech

M. Roland BADEAU, Maître de conférence, Télécom ParisTech

Rapporteur

Rapporteur

Président du jury

Examinateur

Invité

Directeur de thèse

Directeur de thèse

THÈSE

Télécom ParisTech

Grande école de l'Institut Télécom – membre fondateur de ParisTech

46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Résumé

Le principal objectif de cette thèse est de proposer des méthodes de décomposition de spectrogrammes de signaux musicaux reposant sur les redondances qui y sont naturellement présentes et sur lesquelles s'appuie tout auditeur pour comprendre et analyser ces signaux.

Les décompositions proposées sont issues des méthodes de factorisation non-négative telles que la Factorisation en matrices non-négatives (**NMF**). La **NMF**, puissante technique de réduction de rang de données non-négatives très utilisée pour décomposer des spectrogrammes musicaux, est réputée pour fournir une décomposition sur un petit nombre de motifs fréquentiels automatiquement extraits des données, qui ont généralement un sens perceptif.

La **NMF** ne permet cependant pas de modéliser de façon efficace certaines variations temporelles d'éléments sonores non-stationnaires communément rencontrées dans la musique. Cette thèse propose donc d'introduire des modèles génératifs de spectrogrammes musicaux basés sur des modèles simples et classiques de synthèse sonore (synthèse source/filtre, synthèse additive et synthèse par table d'onde) afin de pouvoir prendre en compte de telles variations.

Cette thèse se concentre sur deux types de variations courantes : d'une part, les variations d'enveloppe spectrale que l'on rencontre par exemple dans les sons d'instruments à cordes métalliques libres ou dans les sons modulés par la bouche (comme la voix chantée), d'autre part, les variations de fréquence fondamentale rencontrées par exemple dans des phénomènes tel que le vibrato ou la prosodie.

L'introduction de modèles de synthèse simples dans les méthodes de factorisation permet de proposer des décompositions capables de prendre en compte ces variations : l'utilisation d'un modèle de synthèse source/filtre permet de modéliser les variations spectrales de certains objets musicaux au cours du temps. L'introduction d'un modèle d'atomes harmoniques paramétriques inspiré de la synthèse additive permet de modéliser les variations de fréquence fondamentale. Cette première méthode manquant de robustesse, une seconde piste a été explorée : il s'agit d'un modèle inspiré de la synthèse par table d'onde qui utilise des transformations d'un unique atome de base afin de recréer toute la tessiture de chaque instrument.

Cette thèse propose ainsi de nouvelles méthodes de décomposition des spectrogrammes musicaux qui permettent d'obtenir une représentation intermédiaire en atomes sonores pouvant être utilisée pour diverses applications. Une application de séparation de sources informée par la partition musicale ainsi qu'une application de modification de notes isolées dans un mélange polyphonique sont ainsi présentées à la fin de ce mémoire.

Abstract

The main goal of this thesis is to propose new methods for musical spectrograms decomposition relying on the redundancies on which every listener bases his perception of music.

Proposed decompositions are derived from non-negative factorization methods such as Non-negative Matrix Factorization (**NMF**): **NMF** is a powerful rank reduction method of non-negative data commonly used to decompose musical spectrograms, which is renowned for providing a decomposition on a few frequency patterns (extracted from the data) that generally have a perceptual meaning.

However, **NMF** cannot efficiently model some kinds of temporal variations of non-stationary events usually found in musical spectrograms. This thesis proposes to introduce generative models of musical spectrograms relying on simple models of sound synthesis (source/filter synthesis, additive synthesis and wavetable synthesis) in order to take such variations into account.

This thesis focuses on two types of common variations: on the one hand, the spectral envelope variations that can be found, for instance, in plucked strings sounds or in singing voice, and on the other hand, the fundamental frequency variations found for instance in phenomena such as vibrato or prosody.

Introducing simple synthesis models in factorization methods makes it possible to propose decompositions able to model such variations: a source/filter model permits to take spectral variations of musical objects over time into account. A model of spectrogram with parametric harmonic atoms inspired by additive synthesis makes it possible to model fundamental frequency variations. Because of a lack of robustness of this first method, a second method (that deals with the variations of the fundamental frequency) is also proposed: the model is inspired by wavetable synthesis and uses transformations of a single atom in order to generate all the possible fundamental frequencies of each instrument.

Thus, this thesis proposes new methods for musical spectrograms decomposition which provide mid-level representations on a basis of atoms that can be used in several applications. A score informed source separation system and an application to single note editing in a polyphonic signal are then presented at the end of this thesis.

Table des matières

Résumé	3
Table des matières	5
Notations	9
Abréviations	11
Glossaire	13
1 Introduction	15
1.1 Décomposition des signaux musicaux	16
1.2 Contexte	16
1.3 Factorisation et modèles de synthèse	19
1.3.1 Synthèse source/filtre	20
1.3.2 Synthèse additive	20
1.3.3 Synthèse par table d'onde	21
1.4 Structure du document	21
2 Factorisation en matrices non-négatives	23
2.1 Présentation générale	23
2.1.1 Modèle	23
2.1.2 Fonction de coût	25
2.1.2.1 Divergences courantes	25
2.1.2.2 Divergence de Bregman et β -divergence	26
2.2 Unicité	27
2.2.1 Changement d'échelle et permutation	28
2.2.2 Extension/rétrécissement du cône polyédrique des solutions	28
2.2.3 Problème lié : impossibilité d'une factorisation exacte	30
2.3 Décomposition de spectrogrammes musicaux	31
2.3.1 Principe	31
2.3.2 Choix de l'exposant	32
2.3.2.1 Cas à deux composantes « indépendantes »	33
2.3.2.2 Autres cas	36
2.3.3 NMF et séparation de sources	37
2.4 Modélisation probabiliste	37

2.4.1	Modèles génératifs	38
2.4.1.1	Modèle gaussien	38
2.4.1.2	Modèle de Poisson	39
2.4.2	Analyse probabiliste en composantes latentes (PLCA)	39
2.5	Algorithmes	41
2.5.1	Algorithmes divers	41
2.5.1.1	Descente de gradient projeté	41
2.5.1.2	Méthode de Newton projetée	42
2.5.1.3	Moindres carrés alternés	42
2.5.1.4	Méthode non contrainte par reparamétrisation du problème	42
2.5.2	Mises à jour multiplicatives	43
2.5.2.1	Approche simple	44
2.5.2.2	Algorithme Majoration/Minimisation (MM)	45
2.5.2.3	Algorithme Espérance/Maximisation (EM)	47
2.5.2.4	Intérêts des algorithmes multiplicatifs	49
2.6	Variantes de la NMF et ajout de contraintes	50
2.6.1	Décompositions invariantes par translation	50
2.6.1.1	Décomposition invariante par translation temporelle : NMFD	51
2.6.1.2	Décomposition invariante par translation fréquentielle	52
2.6.2	Contraintes	53
2.7	Limitations de la NMF, variations temporelles	54
2.7.1	Variations d'enveloppes spectrales	54
2.7.2	Variations de fréquence fondamentale	55
3	Modélisation des variations d'enveloppe spectrale : modèle source/filtre et NMF	57
3.1	Modèle	58
3.1.1	Activation temps/fréquence	58
3.1.2	Paramétrisation source/filtre	59
3.2	Algorithme	60
3.2.1	Mise à jour des atomes	61
3.2.2	Mise à jour des activations globales	61
3.2.3	Mise à jour des filtres	62
3.2.4	Description globale et implémentation pratique	64
3.2.5	Dimension de l'espace des paramètres	65
3.2.6	Complexité algorithmique	66
3.2.7	Implémentation et choix de β	67
3.3	Exemples	67
3.3.1	Guimbarde	67
3.3.1.1	Description du signal décomposé	67
3.3.1.2	Expérience et résultat	68
3.3.2	Didgeridoo	70
3.3.2.1	Description du signal décomposé	70
3.3.2.2	Expérience et résultat	70
3.3.3	Clavecin	71
3.3.3.1	Description du signal décomposé	71

3.3.3.2	Expérience et résultat	71
3.3.4	Guitare avec pédale wah-wah	73
3.3.4.1	Description du signal décomposé	73
3.3.4.2	Expérience et résultat	73
3.3.5	Convergence de l'algorithme	75
3.4	Conclusion	75
4	Modélisation des variations de fréquence fondamentale	77
4.1	Spectrogramme paramétrique	77
4.1.1	Modèle	77
4.1.1.1	Atome harmonique paramétrique	78
4.1.1.2	Expression de g	79
4.1.1.3	Fonction de coût et nombre d'atomes	81
4.1.1.4	Atomes de NMF standard pour modéliser les parties non-harmoniques	81
4.1.2	Algorithme	81
4.1.2.1	Mise à jour de f_0	82
4.1.2.2	Mise à jour de \mathbf{H}	84
4.1.2.3	Mise à jour de \mathbf{a}	85
4.1.2.4	Mise à jour de \mathbf{W}' et \mathbf{H}'	85
4.1.2.5	Contraintes	86
4.1.2.6	Détails de l'implémentation	87
4.1.3	Exemple	87
4.1.4	Conclusion	91
4.2	Transformation des atomes	94
4.2.1	PLCA invariante par translation fréquentielle	95
4.2.2	Décomposition invariante par homothétie	97
4.2.3	Algorithme Espérance-Maximisation (EM)	100
4.2.3.1	Mise à jour de $P(z)$:	102
4.2.3.2	Mise à jour de $P_K(f' z)$:	102
4.2.3.3	Mise à jour de $P_I(\lambda_k, t z)$:	103
4.2.3.4	Mises à jour multiplicatives	104
4.2.3.5	Complexité algorithmique	105
4.2.4	Exemples	105
4.2.4.1	Exemple synthétique	105
4.2.4.2	Enregistrement réel	108
4.2.5	Conclusion	108
5	Applications et fusion des modèles	111
5.1	Séparation de sources informée par la partition	111
5.1.1	Modèle paramétrique de spectrogramme de mélange	113
5.1.1.1	Modèle de spectrogramme source	113
5.1.1.2	Modèle de spectrogramme de mélange	113
5.1.1.3	Modèle des éléments non harmoniques	114
5.1.2	Système de séparation	115
5.1.2.1	Initialisation à l'aide de la partition	116

5.1.2.2	Algorithme de décomposition	116
5.1.3	Résultats	117
5.1.3.1	Description de la base de données	118
5.1.3.2	Expérience	119
5.1.3.3	Résultats	119
5.1.4	Conclusion	122
5.2	Modification de notes isolées dans un signal polyphonique	122
5.2.1	Méthode de séparation	122
5.2.1.1	Méthode	122
5.2.1.2	Exemple	123
5.2.2	Modifications	124
5.3	Fusion des modèles paramétrique et source/filtre	126
5.3.1	Modèle mixte	127
5.3.2	Exemple de décomposition	128
Conclusion		133
Bibliographie		137
Table des figures		147
Liste des tableaux		151
Remerciements		153

Notations

\mathcal{S}	Ensemble
a	Scalaire
\mathbf{a}	Vecteur (colonne)
\mathbf{a}^T	Vecteur ligne transposé du vecteur \mathbf{a}
a_i	Coefficient i du vecteur \mathbf{a}
$\ \mathbf{a}\ _p$	Norme p du vecteur \mathbf{a} : $\ \mathbf{a}\ _p = \left(\sum_{i=1}^I a_i ^p \right)^{\frac{1}{p}}$
\mathbf{H}	Matrice
$[\mathbf{H}]_{ij}, h_{ij}, h_{i,j}$	Coefficient d'indice (i, j) de la matrice \mathbf{H} .
$h_{r,:}$	Vecteur égal à la r -ème ligne de la matrice \mathbf{H}
\mathbf{H}^T	Matrice transposée de la matrice \mathbf{H}
$\mathbf{W}\mathbf{H}$	Produit des matrices \mathbf{W} et \mathbf{H} : $[\mathbf{W}\mathbf{H}]_{ft} = \sum_{r=1}^R [\mathbf{W}]_{fr} [\mathbf{H}]_{rt}$
$\mathbf{A} \odot \mathbf{B}$	Produit d'Hadamard (élément par élément) des matrices \mathbf{A} et \mathbf{B} : $[\mathbf{A} \odot \mathbf{B}]_{ft} = [\mathbf{A}]_{ft} [\mathbf{B}]_{ft}$
$\frac{\mathbf{A}}{\mathbf{B}}$	Division terme à terme des matrices \mathbf{A} et \mathbf{B} : $\left[\frac{\mathbf{A}}{\mathbf{B}} \right]_{ft} = \frac{[\mathbf{A}]_{ft}}{[\mathbf{B}]_{ft}}$
$\mathbf{A}^{\odot\eta}$	Exponentiation terme à terme de la matrice \mathbf{A} : $[\mathbf{A}^{\odot\eta}]_{ft} = ([\mathbf{A}]_{ft})^\eta$
$\mathbf{1}_{M,N}$	Matrice de taille $M \times N$ dont tous les coefficients sont égaux à 1
$\dim \mathbf{W}$	Dimension de la matrice \mathbf{W} : si \mathbf{W} est de taille $F \times R$ alors $\dim \mathbf{W} = FR$

\mathcal{N}_c	Loi normale circulaire (complexe)
\mathcal{P}	Loi de Poisson
\propto	Suit une loi
d_{IS}	Distance d'Itakura-Saito scalaire
d_{KL}	Divergence de Kullback-Leibler scalaire
d_{EUC}	Distance euclidienne scalaire
d_β	β -divergence scalaire
d_ϕ	Divergence de Bregman scalaire associée à la fonction ϕ

$\mathbb{1}_{\mathcal{S}}$	Fonction indicatrice de l'ensemble \mathcal{S} : $\mathbb{1}_{\mathcal{S}}(x) = \begin{cases} 1 & \text{si } x \in \mathcal{S} \\ 0 & \text{sinon} \end{cases}$
----------------------------	---

δ_{ij}	Symbol de Kronecker : $\delta_{ij} = 1 \Leftrightarrow i = j$
$\lfloor x \rfloor$	Partie entière par défaut de $x \in \mathbb{R}$: $\max\{n \in \mathbb{Z} n \leq x\}$
$[x]$	Partie entière par excès de $x \in \mathbb{R}$: $\min\{n \in \mathbb{Z} n \geq x\}$
$[x]$	Arrondi à l'entier le plus proche de $x \in \mathbb{R}$: $\arg \min\{ x - n n \in \mathbb{Z}\}$
δ	Distribution de Dirac
$n!$	Factorielle de l'entier n : $n! = \prod_{k=1}^n k$

∇	Opérateur gradient
----------	--------------------

Abréviations

TFCT	Transformée de Fourier à Court Terme
TQC	Transformée à Q-Constant
NMF	Factorisation en matrices non-négatives (de l'anglais : <i>Non-negative Matrix Factorization</i>)
NMFD	Déconvolution de facteurs de matrices non-négatives (de l'anglais : <i>Non-negative Matrix Factor Deconvolution</i>)
PLCA	Analyse probabiliste en composantes latentes (de l'anglais : <i>Probabilistic Latent Component Analysis</i>)
DVS	Décomposition en Valeurs Singulières
EUC	Euclidien(ne)
KL	Kullback-Leibler
IS	Itakura-Saito
MM	Majoration/Minimisation
EM	Espérance/Maximisation
ADSR	Attack Decay Sustain Release. Modèle simple d'enveloppe temporelle à quatre phases : attaque, déclin, maintien et relâchement.
ARMA	Auto-Régressif(s) à Moyenne Ajustée
AR	Auto-Régressif(ve)
MA	à Moyenne Ajustée
SIR	Rapport signal à interférence (de l'anglais : <i>Signal to Interference Ratio</i>)
SAR	Rapport signal à artefact (de l'anglais : <i>Signal to Artifact Ratio</i>)
SDR	Rapport signal à distorsion (de l'anglais : <i>Signal to Distortion Ratio</i>)

Glossaire

Fréquence fondamentale : Pour un son périodique, la fréquence fondamentale est définie comme l'inverse de la période.

Harmonique : On dit qu'un signal sonore est harmonique s'il est périodique. Ce signal peut alors s'écrire comme une somme de sinusoïdes (appelées harmoniques) dont la fréquence est en rapport entier avec la fréquence fondamentale. Dans le domaine spectral, cette propriété se caractérise par la présence de pics spectraux au niveau de la fréquence fondamentale et de ses multiples entiers. Un motif spectral (ou un atome spectral) sera qualifié d'harmonique s'il présente une telle propriété.

Piano roll : Un *piano roll* est une représentation bi-dimensionnelle de la musique proche d'une partition mais à un niveau symbolique moindre. La dimension horizontale représente le temps et la dimension verticale la hauteur de note qui est graduée en demi-tons : les notes y sont repérées par des rectangles (dont la couleur peut éventuellement contenir une information sur la vitesse). Cette représentation est utilisée dans la plupart des séquenceurs MIDI car sa correspondance avec les fichiers MIDI est très simple.

MIDI : acronyme de *Musical Instrument Digital Interface*. Il s'agit d'un protocole de communication entre appareils de commande (clavier-maître par exemple) et appareils de musique électronique (synthétiseurs et effets par exemple). De nombreux messages de contrôle existent parmi lesquelles les messages de notes qui permettent de commander la production d'une note dans un synthétiseur. On retrouve notamment ces messages de notes dans les fichiers MIDI qui constituent donc une information proche de celle contenue dans une partition (à un niveau symbolique moindre) et qui permettent donc à un synthétiseur de générer intégralement un morceau.

Spectrogrammes : On travaille dans tout ce document sur des spectrogrammes qui, sauf précision contraire, sont issus d'une Transformée de Fourier à Court Terme (**TFCT**) du signal temporel.

Si \mathbf{x} est un signal temporel réel de longueur N , alors sa **TFCT** est une matrice dont chaque colonne est la transformée de Fourier discrète (dont on ne garde que les fréquences positives) d'une trame fenêtrée du signal \mathbf{x} . Ainsi elle est définie comme la matrice \mathbf{X} de taille $F \times T$, de coefficients :

$$[\mathbf{X}]_{ft} = \sum_{l=1}^L x_{l+(t-1)h} w_l e^{-i \frac{2\pi(f-1)(l-1)}{L}} \quad (1)$$

où h est le pas entre deux trames successives et \mathbf{w} est la fenêtre d'analyse, il s'agit d'un vecteur de longueur L . Dans ce document, on utilisera généralement une fenêtre de Hann ou une fenêtre de Hamming. On a $F = \lfloor \frac{L}{2} \rfloor + 1$ (les autres termes de la transformée de Fourier sont redondants pour un signal \mathbf{x} réel) et $T = \lfloor \frac{N}{h} \rfloor + 1$.

Dans tout ce document on utilise la notation \mathbf{X} quand on parle du spectrogramme complexe (c'est-à-dire la **TFCT**) et \mathbf{V} quand on parle d'un spectrogramme réel, généralement d'amplitude (module de la **TFCT**) ou de puissance (module au carré de la **TFCT**).

Chapitre 1

Introduction

Pour un être humain, il est généralement facile de décrire un extrait de musique : il est assez aisé d'identifier grossièrement les instruments qui le composent, d'en donner le genre, de donner une idée du tempo, et pour un auditeur un peu plus avancé, il est même possible de donner des informations beaucoup plus détaillées, comme de décrire les notes que joue chaque instrument. Ce genre de tâches peut s'avérer cependant beaucoup plus difficile pour un ordinateur, les signaux musicaux étant par nature complexes et la perception de ceux-ci étant difficile à appréhender. De nombreux travaux cherchent ainsi à extraire automatiquement de l'information d'un signal musical : information sur le rythme et le tempo [Scheirer, 1998], sur le genre du morceau [Allamanche *et al.*, 2004], sur les instruments présents [Essid, 2006], sur l'harmonie [Oudre *et al.*, 2009], sur la mélodie [Weil *et al.*, 2009], sur la position temps/fréquence de toutes les notes [Emiya *et al.*, 2010]... Tous ces travaux relèvent du domaine de recherche de l'extraction automatique d'information de la musique (souvent abrégé MIR, de l'anglais : *Music Information Retrieval*).

L'être humain est également capable de focaliser son attention sur un unique instrument dans un orchestre. Le domaine de recherche de la séparation de sources tente de reproduire cette capacité en essayant de séparer les signaux des différents instruments en présence [Virtanen, 2007, Smaragdis *et al.*, 2007].

Ces deux domaines de recherche (MIR et séparation de sources) ont tendance à se rapprocher et de plus en plus de travaux cherchent à allier les deux domaines, c'est-à-dire d'une part l'aspect représentation d'information compacte de haut-niveau capable de décrire l'extrait considéré et d'autre part l'aspect manipulation des divers éléments en présence, par exemple par le biais d'une représentation intermédiaire dite « mi-niveau » permettant des applications dans les deux domaines [Durrieu *et al.*, 2011].

Cette thèse s'inscrit dans cette tendance : son principal objectif est de proposer des méthodes de décomposition de signaux sonores musicaux reposant sur les redondances qui y sont naturellement présentes et sur lesquelles s'appuie tout auditeur pour « comprendre » ces signaux. Ces décompositions sont construites en introduisant des modèles génératifs de spectrogrammes basés sur des modèles simples de synthèse sonore. Cette thèse propose donc des décompositions *intelligentes* des signaux musicaux qui permettent d'obtenir une représentation intermédiaire en « atomes sonores », éléments constitutifs élémentaires de la musique. Ce type de représentation de la musique peut avoir de nombreuses applications notamment en séparation de sources sonores, en transcription automatique de partition ou dans le domaine de la transformation du son.

1.1 Décomposition des signaux musicaux

Les signaux musicaux possèdent de très importantes redondances et peuvent être généralement décrits avec beaucoup moins d'informations que leur forme d'onde : un morceau de musique est en effet généralement composé d'événements musicaux (par exemple des notes de musique, des sons de percussions...) qui se répètent au cours du temps. Notre perception de la musique est majoritairement influencée par ces événements (qui sont définis par une connaissance a priori de la musique que nous avons l'habitude d'écouter) et leur apparition répétée au cours du morceau. La perception est en effet principalement basée sur ce qu'on attend : on arrive à avoir une perception cohérente de la musique et à structurer cette perception grâce à la redondance qui y est présente. Une représentation proche de ce qui est perçu en termes d'événements sonores permet donc de comprendre et d'analyser la musique comme le ferait une personne. Il est donc très intéressant de pouvoir obtenir une telle représentation à partir d'une simple forme d'onde.

Les transformées temps/fréquence de type spectrogramme permettent de faire apparaître en partie les redondances perceptives même si celles-ci restent complexes. Ainsi dans la figure 1.1, on voit clairement apparaître une redondance forte entre certaines parties du spectrogramme.

De nombreux travaux cherchent à extraire automatiquement une structure qui explique bien le signal étudié tout en ayant réduit considérablement la quantité d'informations pour le décrire. On peut notamment citer deux types de méthodes : les méthodes de décomposition parcimonieuse [Mallat et Zhang, 1993, Chen *et al.*, 1998] qui cherchent à décomposer le signal sur un ensemble de signaux de base bien choisis pour pouvoir décrire le signal étudié, et les méthodes de factorisation, telles que la Factorisation en matrices non-négatives (NMF) [Lee et Seung, 1999], qui cherchent à extraire automatiquement des structures redondantes qui apparaissent dans les données analysées et qui peuvent donc être interprétées comme des méthodes de décomposition parcimonieuse pour lesquelles le dictionnaire est appris automatiquement à partir du signal. C'est sur ce deuxième type de méthodes que nous nous focalisons dans cette thèse.

1.2 Contexte

Si les méthodes de factorisation s'avèrent puissantes et robustes pour décomposer des signaux composés d'objets stationnaires, leur utilisation est plus délicate lorsqu'interviennent des éléments présentant des variations au cours du temps : plusieurs atomes sont alors nécessaires pour représenter un seul élément et il est généralement difficile de regrouper correctement ces atomes. De plus, même lorsque les éléments constitutifs du spectrogramme sont à peu près stationnaires, il est souvent utile de guider (d'« informer ») la décomposition afin d'obtenir une description réellement utilisable. De nombreux travaux ont donc cherché à dépasser ces limitations des méthodes de factorisation en proposant des techniques de décomposition alternatives.

Ainsi plusieurs types de décompositions ont été proposés : d'une part des décompositions utilisant des modèles physiques sous-jacents pour représenter les éléments des spectrogrammes musicaux, ce type de décomposition aboutissant généralement à des représentations « mi-niveau » qui peuvent être exploitées dans diverses applications, d'autre part des méthodes statistiques proposées récemment qui utilisent notamment des modèles de

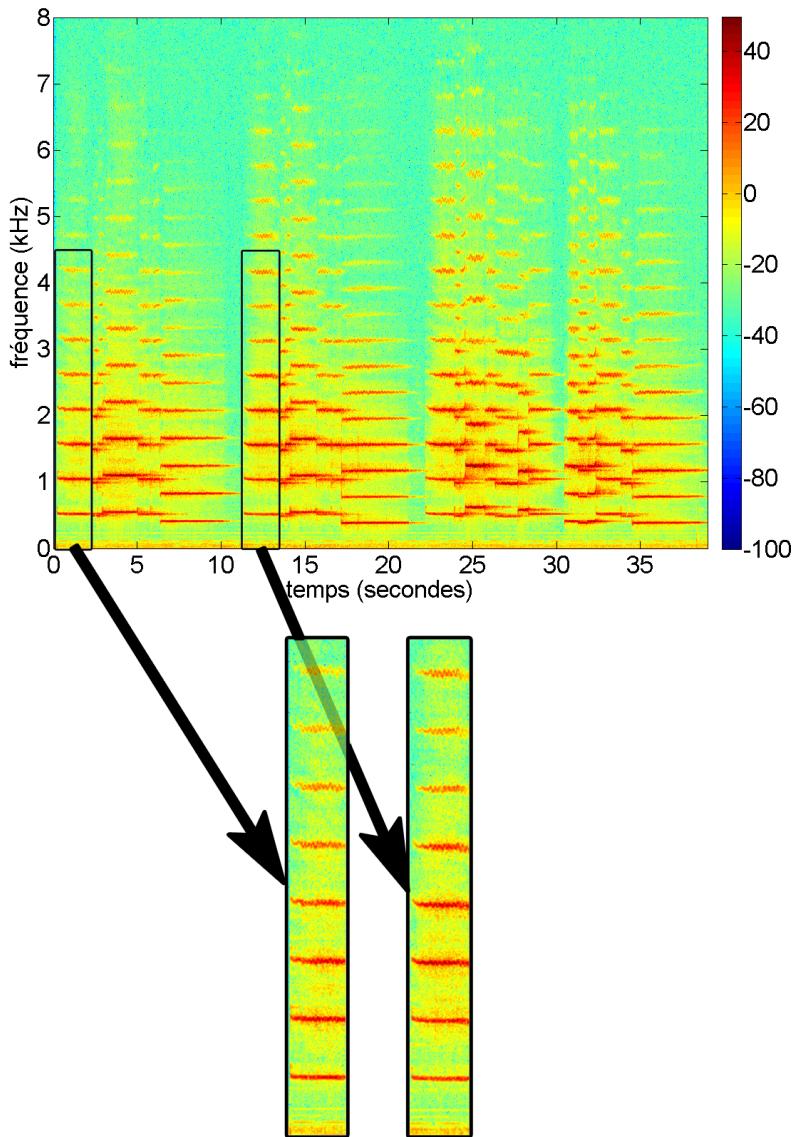


FIG. 1.1 – Spectrogramme de l'introduction de *Godfather Waltz* par Nino Rota. En bas : zoom sur des parties du spectrogramme, mettant en lumière l'existence de répétitions.

Markov cachés afin de structurer les variations temporelles.

Le premier type de méthodes introduit des modèles censés représenter les éléments de base en se basant sur des considérations physiques. La plupart des éléments rencontrés dans les spectrogrammes musicaux étant des notes de musique, qui sont des signaux harmoniques ou quasi-harmoniques, une des idées les plus simples, d'ailleurs réutilisée dans le chapitre 4 de cette thèse, est de considérer comme éléments de base des atomes harmoniques. Cette idée a été d'abord utilisée dans un cadre de décomposition parcimonieuse : on peut par exemple citer l'algorithme d'*Harmonic Matching Pursuit* [Gribonval et Bacry, 2003] qui reprend l'algorithme de *Matching Pursuit* [Mallat et Zhang, 1993] et l'adapte à des diction-

naires d'atomes harmoniques. Cette approche a été complétée dans [Leveau *et al.*, 2008, Leveau, 2007] qui aboutit à une représentation « mi-niveau » formée de molécules, agrégats d'atomes harmoniques appris sur une base de sons instrumentaux, censées correspondre à un objet sonore. Dans [Bello *et al.*, 2006], la musique de piano est décomposée sur 88 atomes harmoniques (les 88 notes du piano) d'abord extraits automatiquement du signal par des méthodes fréquentielles, puis servant de base pour la décomposition. Dans [Bertin *et al.*, 2007, Bertin, 2009, Vincent *et al.*, 2010], une extension de la Factorisation en matrices non-négatives (**NMF**) est proposée, dans laquelle la (quasi-)harmonicité des atomes est également imposée : les atomes « complexes » sont construits comme une combinaison linéaire d'atomes « simples » correspondant chacun à une sous-bande. Ce type d'approche permet notamment de garantir une enveloppe spectrale relativement lisse et de pouvoir prendre en compte une certaine inharmonicité qui peut être par exemple rencontrée dans le son du piano.

L'idée d'utiliser des sous-bandes est reprise par Durrieu [Durrieu *et al.*, 2010, 2009a] sous une forme un peu différente : celui-ci introduit un modèle source/filtre dans la **NMF** dans le but de modéliser les fortes variations d'enveloppe spectrale de la voix chantée. Ce type de modèle permet d'aboutir à une représentation mi-niveau (*cf.* [Durrieu *et al.*, 2011]) contenant simultanément des informations sur la fréquence fondamentale et le timbre.

Les décompositions utilisant des transformations des atomes reposent également sur une hypothèse sous-jacente d'harmonicité : dans les décompositions invariantes par translation fréquentielle [Schmidt et Mørup, 2006, Smaragdis *et al.*, 2008] qui sont utilisées pour décomposer des spectrogrammes à Q-constant, différentes notes d'un même instrument sont censées pouvoir être représentées par un même atome translaté (une translation correspondant alors à une transposition) : ce type de méthode permet de rendre encore plus compacte la représentation tout en permettant une prise en compte des variations de fréquence fondamentale au cours du temps. Nous avons d'ailleurs repris cette idée afin de l'adapter à des spectrogrammes classiques issus d'une Transformée de Fourier à Court Terme (**TFCT**) dans le chapitre 4.

C'est sur ce type de méthodes, qui utilisent une forte information *a priori* sur la façon dont sont constitués les spectrogrammes musicaux, que nous avons focalisé notre attention dans cette thèse, notre principal objectif étant de proposer des décompositions qui permettent la modélisation des variations temporelles des éléments sonores tout en aboutissant à une représentation intermédiaire du signal servant de base pour diverses applications, comme par exemple la séparation de sources ou la transcription.

Récemment de nouvelles méthodes ont également cherché à modéliser les variations temporelles en introduisant des modèles de Markov cachés dans les méthodes de factorisation [Ozerov *et al.*, 2009, Nakano *et al.*, 2010b, Mysore *et al.*, 2010] : dans ce type de méthode, une structure temporelle est apprise de façon statistique. Ainsi soit les atomes sont temporellement liés entre eux via des probabilités de transition pour chaque composante (comme c'est le cas dans [Ozerov *et al.*, 2009, Nakano *et al.*, 2010b]) et chaque atome a alors un « état », soit les atomes sont regroupés en dictionnaires et ce sont ces dictionnaires qui constituent chacun l'état d'une composante (comme c'est le cas dans [Mysore *et al.*, 2010]).

Il s'agit d'une voie assez différente de celle que nous avons choisi d'étudier. Il semble cependant qu'une utilisation jointe de l'approche que nous avons utilisée et de l'approche par modèle de Markov soit pertinente et constitue donc une piste future de travail.

1.3 Factorisation et modèles de synthèse

Les méthodes de factorisation issues d'un modèle de décomposition linéaire comme la **NMF** montrent de sévères limitations dès que les spectrogrammes à décomposer présentent des éléments variables dans le temps : par exemple lorsqu'une même note présente certaines variations spectrales (variations d'enveloppe spectrale, variations de fréquence fondamentale). Pourtant, même dans ce cas, il existe généralement une profonde redondance d'une trame à l'autre de cet élément. La figure 1.2 qui représente le spectrogramme d'une note isolée de violon contenant un important vibrato illustre bien ce problème : le vibrato induit d'importantes différences entre les trames successives (en particuliers pour les partiels de fréquence supérieure à 2000Hz), ainsi un modèle de décomposition linéaire comme la **NMF** nécessitera de nombreux atomes pour correctement décomposer cet extrait constitué d'un unique objet sonore et aboutira ainsi à une représentation difficile à interpréter. Pourtant une forte redondance apparaît clairement dans ce spectrogramme : les trames dans lesquelles la note est jouée présentent toutes une structure harmonique commune et les amplitudes de ces harmoniques restent quasiment constantes. Un unique paramètre varie au cours du temps : la fréquence fondamentale. Connaitre les variations de ce paramètre devrait donc permettre de retrouver la redondance.

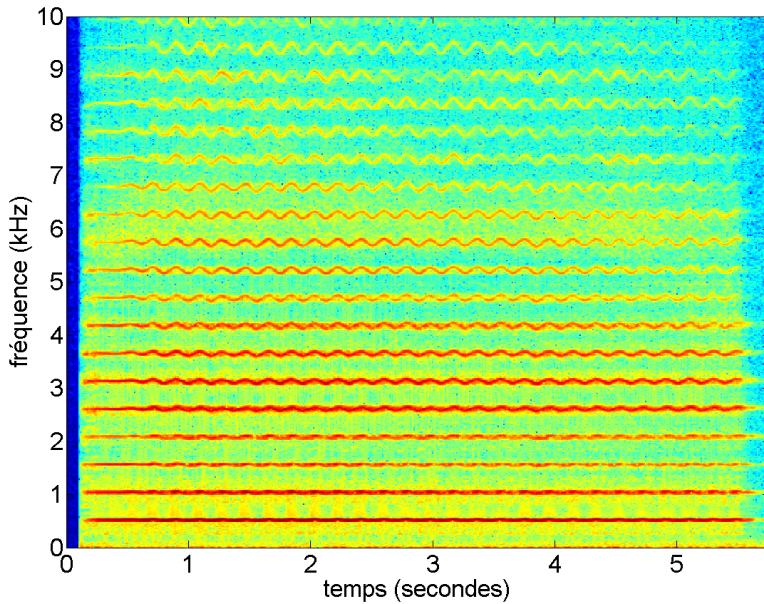


FIG. 1.2 – Spectrogramme d'une note (*Do4*) de violon contenant un important vibrato.

Ainsi, l'idée directrice suivie dans cette thèse est de structurer les méthodes de factorisation comme la **NMF** et ses dérivées afin de pouvoir prendre en compte les variabilités de certains éléments sonores qui sont couramment rencontrées dans des spectrogrammes musicaux. Afin de décrire ces signaux complexes, on va chercher à introduire des modèles capables de décrire cette variabilité : l'idée est d'utiliser des modèles de synthèse sonore simples. Ces modèles de synthèse sont utilisés dans le domaine temps/fréquence (ce qui se traduit dans nos méthodes par une synthèse des spectrogrammes). On cherche alors

à apprendre les paramètres de ces modèles de synthèse. Cette idée permet d'extraire du signal certaines redondances qui ne sont pas représentables avec un simple modèle linéaire comme la **NMF**. Ainsi, trois modèles de synthèse très simples, qui sont des modèles de base de la synthèse sonore très utilisés dans les synthétiseurs, ont été étudiés et introduits dans la **NMF** :

- le modèle de synthèse source/filtre,
- le modèle de synthèse additif,
- le modèle de synthèse par table d'onde.

Ces modèles sont ici utilisés uniquement comme une base de structuration des méthodes de factorisation et ne sont absolument pas destinés à faire de la synthèse dans ce travail.

Il est à noter que, dans cette thèse, nous avons exclusivement cherché à développer des modèles de décomposition de spectrogrammes « classiques » (c'est-à-dire issus d'une **TFCT**) car ceux-ci sont facilement inversibles, et peuvent donc éventuellement permettre une manipulation directe du son (séparation de sources, transformation du son), bien que nous nous inspirons parfois de méthodes conçues pour décomposer d'autres types de spectrogrammes (spectrogrammes à Q-constant par exemple).

1.3.1 Synthèse source/filtre

La méthode de synthèse source/filtre est basée sur un modèle de production dans lequel deux éléments interagissent : une source, productrice du son, et un résonateur filtrant la source. Par exemple dans le cadre de la production de la voix, les cordes vocales jouent le rôle de la source et le conduit vocal celui du résonateur. Ce type de synthèse a été implanté dans de nombreux synthétiseurs et notamment dans les premiers synthétiseurs commerciaux : il peut en effet être facilement mis en place avec des composants analogiques. Il consiste à générer une forme d'onde (généralement simple, périodique et riche en harmoniques) stationnaire qui est modifiée par un filtre dont les paramètres (généralement la fréquence de résonance) évoluent au cours du temps. Nous avons introduit ce modèle de synthèse dans la **NMF** afin de modéliser les variations d'enveloppe spectrale : les activations de la **NMF** sont remplacées par des filtres variant dans le temps et chaque motif spectral factorisé est censé représenter la source, dont la forme ne change pas au cours du temps. La méthode de décomposition proposée peut modéliser des variations d'enveloppe spectrale même si le modèle de production source/filtre ne correspond pas à la réalité physique du son considéré. Cette méthode fait l'objet du chapitre 3.

1.3.2 Synthèse additive

Le synthèse additive est le modèle de synthèse le plus simple et le plus ancien : il consiste simplement à additionner des sinusoïdes afin de recréer un son. Il est particulièrement adapté aux sons harmoniques (ou quasi-harmoniques) pour lesquels les fréquences des sinusoïdes sont réparties de façon harmonique. Dans le domaine spectral, si on se limite à des sons harmoniques, le modèle consiste donc à créer un atome harmonique : ce type de modèle est ainsi utilisé dans la méthode de décomposition présentée dans la section 4.1. Les amplitudes des harmoniques sont factorisées et les fréquences fondamentales des atomes estimées à chaque instant.

1.3.3 Synthèse par table d'onde

Le principe de la synthèse par table d'onde est simple : à partir d'une période (ou d'un petit nombre de périodes) d'un son harmonique, on produit le son par lecture périodique de cette période à une vitesse qui permet de créer la note souhaitée. Le son périodique obtenu est alors multiplié par une enveloppe d'amplitude. Dans le domaine spectral, cette opération consiste donc à disposer d'un motif spectral de base que l'on dilate ou compresse. C'est précisément ce type de modèle qui est utilisé dans la méthode proposée dans la section 4.2, dans laquelle le motif spectral est appris à partir du spectrogramme.

1.4 Structure du document

Le chapitre 2, principalement bibliographique, constitue une présentation détaillée de la NMF qui sert de base aux décompositions présentées dans les chapitres suivants. Le chapitre 3 présente une extension de la NMF qui permet de modéliser les variations d'enveloppe spectrale à l'aide d'un modèle source/filtre. Le chapitre 4 présente deux méthodes de décomposition de spectrogrammes permettant de modéliser les variations de fréquence fondamentale. Enfin, le chapitre 5 présente certaines applications de ces décompositions.

Chapitre 2

Factorisation en matrices non-négatives

La Factorisation en matrices non-négatives (**NMF**) est une puissante technique de réduction de rang des matrices à coefficients positifs ou nuls.

Il s'agit d'une technique de factorisation comme l'analyse en composantes principales [Hotelling, 1933], l'analyse en composantes indépendantes [Comon, 1994, Makino *et al.*, 2004], le codage parcimonieux (sparse coding) [Abdallah et Plumley, 2006], qui permet de réduire la dimension des données et donc d'expliquer les données par un petit nombre d'objets représentatifs. La **NMF** se distingue des autres techniques de factorisation par sa contrainte de non-négativité qui permet de factoriser des éléments qui font sens. Ainsi l'article de référence de la **NMF** [Lee et Seung, 1999] montre que l'utilisation de celle-ci sur une base de données d'images de visage permet d'extraire les différentes parties du visage : les éléments extraits correspondent en effet au nez, aux yeux, à la bouche... La **NMF** fournit ainsi une décomposition proche de ce qu'on perçoit. Lorsque la **NMF** est utilisée pour décomposer des spectrogrammes, des propriétés perceptives similaires apparaissent ce qui l'a rendue très populaire dans le domaine du traitement du son.

2.1 Présentation générale

2.1.1 Modèle

On dispose d'une matrice \mathbf{V} non-négative, c'est-à-dire dont tous les coefficients sont positifs ou nuls, de taille $F \times T$. La factorisation en matrices non-négatives (*cf.* [Lee et Seung, 1999]) approxime \mathbf{V} de la manière suivante :

$$\mathbf{V} \approx \mathbf{WH} \tag{2.1}$$

$$\text{i.e. } [\mathbf{V}]_{ft} \approx \sum_{r=1}^R [\mathbf{W}]_{fr} [\mathbf{H}]_{rt}, \tag{2.2}$$

où \mathbf{W} ($F \times R$) et \mathbf{H} ($R \times T$) sont des matrices non-négatives. L'approximation de l'équation (2.1) est généralement quantifiée au moyen d'une divergence (*cf.* section 2.1.2). Ceci revient à décomposer les vecteurs colonnes de \mathbf{V} sur les vecteurs colonnes de \mathbf{W} . R est

généralement choisi tel que $R(F + T) \ll FT$ (contrainte de réduction de la dimension) ou bien $R \ll \min(F, T)$ (contrainte de réduction du rang), ces deux contraintes étant en pratique équivalentes.

Une NMF simple est illustrée dans la figure 2.1.

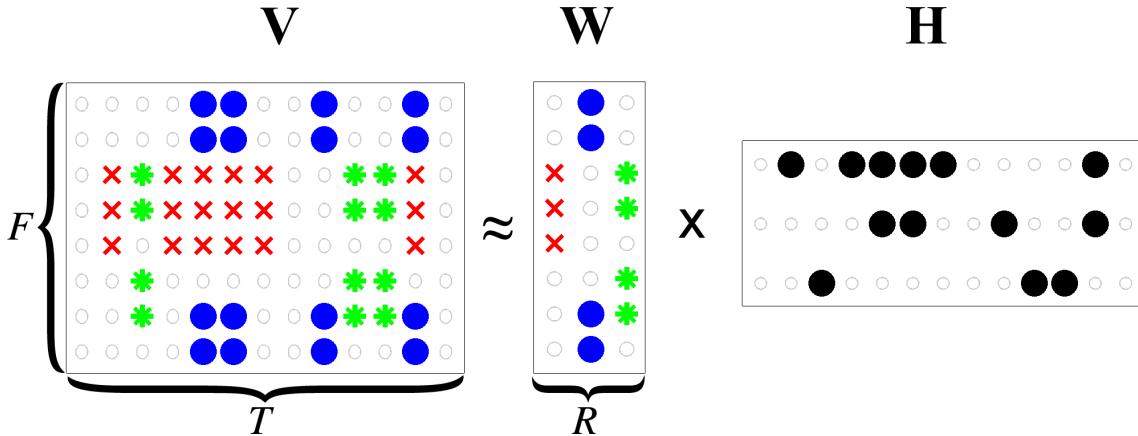


FIG. 2.1 – Factorisation en matrices non-négatives.

Les vecteurs colonnes de \mathbf{W} sont appelés atomes. La matrice \mathbf{H} est appelée matrice d’activation : la r -ème ligne de \mathbf{H} correspond en effet à l’activation du r -ème atome.

La factorisation en matrices non-négatives permet d’extraire automatiquement des motifs qui se répètent au cours du temps dans les données.

Contraintes de non-négativité : La contrainte de non-négativité des matrices \mathbf{W} et \mathbf{H} est la propriété fondamentale qui donne sa spécificité à la **NMF** parmi les autres méthodes de factorisation (telle que l’Analyse en Composantes Principales). De cette contrainte résultent les propriétés appréciables suivantes :

- Les atomes extraits sont dans le même espace (orthant positif) que les données non-négatives décomposées.
- Seules les combinaisons positives d’atomes sont possibles : ainsi il n’est pas possible qu’une partie d’un atome soit annulée par un autre atome et il n’existe donc pas de phénomène de création d’énergie noire.
- La contrainte de non-négativité est souvent présentée comme l’origine de la capacité de la **NMF** à fournir une décomposition perceptive des données. Notamment dans la décomposition de spectrogrammes musicaux, les éléments extraits ont généralement un sens perceptif et peuvent par exemple correspondre à des notes de musique.

La **NMF** a été utilisée dans de nombreux domaines : en traitement d’image [Lee et Seung, 1999, Hoyer, 2004], en fouille de données textuelles [Pauca *et al.*, 2004], en surveillance de messagerie électronique [Berry et Browne, 2005], en spectroscopie [Gobinet *et al.*, 2004]... On trouve, en particulier, de nombreuses applications dans le traitement du signal audio telles que la transcription automatique sur partition [Smaragdis et Brown, 2003, Paulus et Virtanen, 2005, Bertin *et al.*, 2007], la séparation de sources [Virtanen, 2007, Cichocki *et al.*, 2006c, Ozerov et Févotte, 2010], l’alignement sur partition [Cont, 2006] ou encore la restauration audio [Le Roux *et al.*, 2008b].

L'approximation (2.1) est généralement quantifiée par une distance (ou divergence) entre \mathbf{V} et \mathbf{WH} . Un algorithme de NMF est donc un algorithme de minimisation de cette divergence.

2.1.2 Fonction de coût

Généralement, la fonction de coût utilisée en NMF est exprimée au moyen d'une divergence D :

$$\mathcal{C}_{\mathbf{V}}(\mathbf{W}, \mathbf{H}) = D(\mathbf{V} || \mathbf{WH}).$$

Définition 2.1.1. On appelle divergence sur un ensemble \mathcal{S} une application $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$ telle que $\forall (X, Y) \in \mathcal{S}$ on a $D(X || Y) = 0 \Leftrightarrow X = Y$.

Les divergences utilisées en NMF sont, la plupart du temps, séparables :

$$\mathcal{C}(\mathbf{W}, \mathbf{H}) = D(\mathbf{V} || \mathbf{WH}) = \sum_{f=1}^F \sum_{t=1}^T d([\mathbf{V}]_{ft} || [\mathbf{WH}]_{ft}). \quad (2.3)$$

Une divergence ne vérifie pas nécessairement les propriétés de symétrie et d'inégalité triangulaire que doit vérifier une distance, mais une distance est bien un cas particulier de divergence.

2.1.2.1 Divergences courantes

De nombreuses divergences ont été utilisées en NMF dans la littérature. Les plus utilisées sont probablement :

- la divergence d'Itakura-Saito (IS) [Itakura et Saito, 1968] notée d_{IS} . Cette divergence est souvent appelée distance d'IS bien qu'elle ne vérifie ni la symétrie ni l'inégalité triangulaire.
- la divergence de Kullback-Leibler (KL) [Kullback et Leibler, 1951] notée d_{KL} .
- la distance Euclidien(ne) (EUC) notée d_{EUC} .

Ces divergences sont respectivement définies pour tout $(x, y) \in (\mathbb{R}_+)^2$ par :

$$\begin{aligned} d_{IS}(x|y) &= \frac{x}{y} - \log \frac{x}{y} - 1, \\ d_{KL}(x|y) &= x \log \frac{x}{y} + y - x, \\ d_{EUC}(x|y) &= \frac{1}{2}(x - y)^2. \end{aligned}$$

Remarque : La distance EUC est plus généralement définie sur \mathbb{R} par $d_{EUC}(x|y) = |x - y|$ ($D_{EUC}(\mathbf{x}||\mathbf{y}) = \sqrt{\sum_{i=1}^K (x_i - y_i)^2}$ sur \mathbb{R}^K). Cependant le facteur constant et la racine carrée ne changent pas le problème étudié (problème de minimisation) et il est donc équivalent de considérer la définition proposée ici, cette seconde définition ayant l'avantage de correspondre à la généralisation de la β -divergence (voir plus loin) et d'être différentiable pour $x = y$.

Certains auteurs ont proposé des études de la NMF pour des divergences généralisées incluant ces trois divergences classiques :

- La divergence de Csiszar [Cichocki et al., 2006b] qui est une généralisation de l' α -divergence d'Amari [Cichocki et al., 2007] : ces deux divergences incluent la divergence de **KL** et sa divergence duale.
- La divergence de Bregman [Dhillon et Sra, 2006] qui est une généralisation de la β -divergence, introduite dans [Eguchi et Kano, 2001] (dans un cadre différent de la **NMF**) reprise à travers une définition moins large par [Kompass, 2007] comme nous le démontrerons dans la section suivante. Ces deux divergences sont des généralisations des divergences **KL** et **IS** et de la distance **EUC**.

2.1.2.2 Divergence de Bregman et β -divergence

Les β -divergences ont souvent été présentées comme un type de divergence généralisée à part, alors qu'il s'agit d'une sous-classe des divergences de Bregman comme nous l'avons démontré dans [Hennequin et al., 2011a].

Définition 2.1.2. *La β -divergence est définie pour $\beta \in \mathbb{R} \setminus \{0, 1\}$ par :*

$$d_\beta(x|y) = \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta xy^{\beta-1}). \quad (2.4)$$

Pour $\beta \in \{0, 1\}$, la β -divergence est obtenue comme la limite de l'expression (2.4) et correspond respectivement à la divergence d'**IS** et à la divergence de **KL** :

$$\begin{aligned} d_0(x|y) &= d_{\text{IS}}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1, \\ d_1(x|y) &= d_{\text{KL}}(x|y) = x \log \frac{x}{y} + y - x. \end{aligned}$$

Définition 2.1.3. *Soit \mathcal{S} un convexe fermé et soit $F : \mathcal{S} \rightarrow \mathbb{R}$ une fonction de classe C^1 strictement convexe. La divergence de Bregman associée à la fonction F est une divergence sur \mathcal{S} définie pour tout $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$ par :*

$$D_F(\mathbf{x}||\mathbf{y}) = F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}) | \mathbf{x} - \mathbf{y} \rangle.$$

La convexité de F assure la positivité de D_F . On peut facilement restreindre cette définition sur \mathbb{R}^+ :

Définition 2.1.4. *Les divergences de Bregman sur \mathbb{R}^+ sont définies par :*

$$d_F(x|y) = F(x) - F(y) - F'(y)(x - y).$$

où F est une fonction de \mathbb{R}^+ dans \mathbb{R} de classe C^1 strictement convexe.

La divergence de Bregman correspond à la différence entre la valeur de la fonction et son développement limité au premier ordre comme l'illustre la figure 2.2 pour une divergence de Bregman scalaire. Par conséquent la divergence de Bregman n'est pas modifiée en ajoutant une fonction affine à F .

En choisissant :

$$F_\beta(x) = \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta} \quad (2.5)$$

pour $\beta \neq 0, 1$ et les limites par rapport à β de l'expression précédente en ces deux points (*i.e.* $F_0(x) = -\log x + x - 1$ et $F_1(x) = x \log x - x + 1$), on obtient de façon évidente $d_{F_\beta} = d_\beta$.

On a alors pour tout β , $F_\beta''(x) = x^{\beta-2}$, ce qui prouve bien que F_β est convexe.

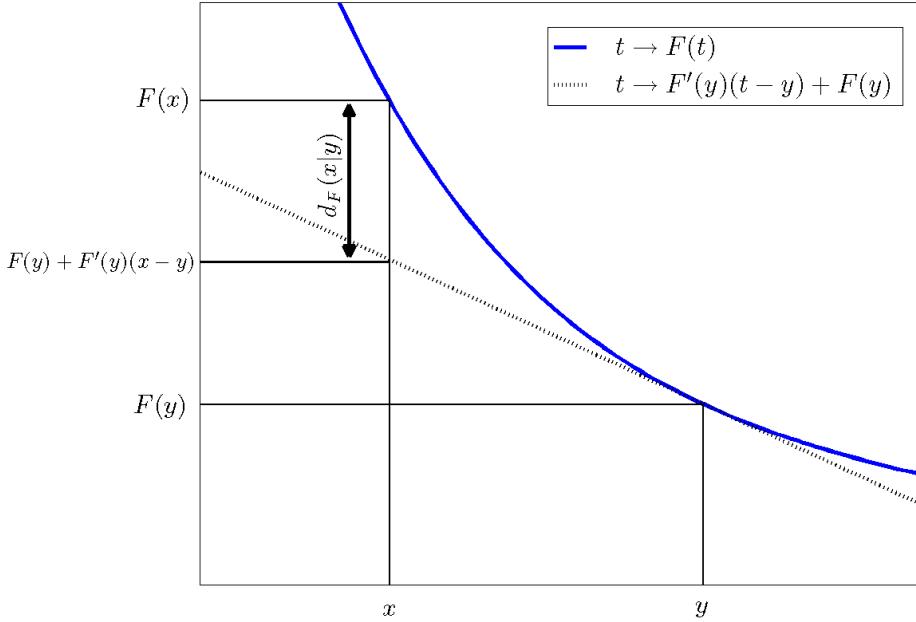


FIG. 2.2 – Représentation graphique de la divergence de Bregman associée à la fonction F entre les points x et y .

Remarque : Il existe une relation simple entre la β -divergence et l' α -divergence d'Amari. Cependant ces deux divergences sont distinctes.

L' α -divergence d'Amari est en effet définie par :

$$d_{\alpha}^{Amari}(x|y) = \frac{\alpha x + (1 - \alpha)y - x^{\alpha}y^{1-\alpha}}{\alpha(1 - \alpha)}.$$

On a alors de façon immédiate :

$$d_{\beta}(x|y) = y^{\beta-1}d_{\beta}^{Amari}(x|y).$$

2.2 Unicité

La factorisation de matrice présentée dans l'équation (2.1) n'est pas unique : il existe en effet de nombreux couples de matrices non-négatives $(\mathbf{W}', \mathbf{H}')$ (de même dimension que \mathbf{W} et \mathbf{H}) tels que $\mathbf{WH} = \mathbf{W}'\mathbf{H}'$. En effet pour toute matrice inversible \mathbf{Q} de taille $R \times R$ telle que \mathbf{WQ} et $\mathbf{Q}^{-1}\mathbf{H}$ sont non-négatives, on a bien $(\mathbf{WQ})(\mathbf{Q}^{-1}\mathbf{H}) = \mathbf{WH}$. On s'intéresse dans cette section uniquement à ce type de transformation et on suppose que les matrices \mathbf{W} et \mathbf{H} sont de rang plein, ce qui n'est pourtant pas forcément souhaitable comme le suggère l'exemple de la section 2.2.3, qui ne peut être factorisé de façon exacte qu'avec des matrices dégénérées.

2.2.1 Changement d'échelle et permutation

Le premier type de transformation qui ne change pas les valeurs du produit \mathbf{WH} est le cas où \mathbf{Q} et \mathbf{Q}^{-1} sont toutes deux non-négatives. On peut démontrer facilement que ces matrices sont exactement les matrices de permutation généralisées à coefficients non-négatifs (On peut par exemple en trouver une démonstration dans [Kaczorek, 2002, p.2]) : il s'agit des matrices inversibles qui ont exactement un coefficient non nul (et donc strictement positif) par colonne (et par ligne).

Les matrices de permutation généralisées peuvent être décomposées en deux matrices plus simples : $\mathbf{Q} = \mathbf{DP}$ où \mathbf{D} est une matrice diagonale (correspondant à une opération de changement d'échelle) de diagonale strictement positive et \mathbf{P} une matrice de permutation. Ainsi ce type de transformation correspond à une permutation et un changement d'échelle des colonnes de la matrice de base \mathbf{W} (la transformation inverse étant appliquée aux lignes de la matrice d'activation \mathbf{H}).

Ce type de transformation n'est donc pas un problème puisqu'il suffit d'imposer une norme aux colonnes de \mathbf{W} (ou aux lignes de \mathbf{H}) pour avoir l'unicité vis-à-vis de la mise à l'échelle, et d'imposer un ordre sur les colonnes de \mathbf{W} (par exemple on peut imposer que ces colonnes soient classées dans l'ordre croissant par rapport à l'ordre lexicographique) pour avoir l'unicité vis-à-vis de la permutation.

2.2.2 Extension/rétrécissement du cône polyédrique des solutions

L'unicité de la **NMF** est un problème particulièrement compliqué qui est, entre autres, à l'origine des difficultés rencontrées pour obtenir des résultats théoriques sur la convergence des algorithmes d'estimation. Nous évoquons, dans cette section et la suivante, la notion d'unicité « standard » de la **NMF** (c'est-à-dire la notion d'unicité à homothétie et permutation près) et tentons de montrer que cette notion est complexe et peut s'avérer contre-intuitive.

Un deuxième type de transformation de \mathbf{W} et \mathbf{H} que l'on peut opérer fait appel à une interprétation un peu différente de la **NMF** : On peut en effet considérer le problème de **NMF** comme la recherche d'un cône polyédrique permettant d'approximer au mieux les vecteurs colonne de V . Les vecteurs colonne du produit \mathbf{WH} peuvent en effet être vus comme des points d'un cône polyédrique. Ce cône polyédrique $\mathcal{C}_{\mathbf{W}}$ est par définition le cône engendré (positivement) par les vecteurs colonne de \mathbf{W} c'est-à-dire :

$$\mathcal{C}_{\mathbf{W}} = \left\{ \sum_{i=1}^R h_i \mathbf{w}_i \mid (h_1, \dots, h_R) \in (\mathbb{R}^+)^R \right\},$$

où les \mathbf{w}_i sont les vecteurs colonne de \mathbf{W} . Remarque : on utilise ici le théorème de Minkowski-Weil qui stipule que les cônes polyédriques sont exactement les cônes engendrés par des familles finies de vecteurs.

Dans de nombreux cas ce cône peut être agrandi ou rétréci (en conservant un cône polyédrique) sans modifier le produit \mathbf{WH} . En effet, pour i et k fixés ($k \neq i$), si aucun des coefficients de \mathbf{w}_i n'est nul (il suffit en fait que pour tout j tel que $w_{ji} = 0$, on ait $w_{jk} = 0$) alors on peut appliquer la transformation :

$$\mathbf{w}'_i = \mathbf{w}_i - \lambda_k \mathbf{w}_k \quad \text{avec} \quad 0 < \lambda_k \leq \min_{j, w_{jk} \neq 0} \frac{w_{ji}}{w_{jk}}.$$

On obtient une nouvelle famille de vecteurs non-négatifs $\{\mathbf{w}_1, \dots, \mathbf{w}_{i-1}, \mathbf{w}'_i, \mathbf{w}_{i+1}, \dots, \mathbf{w}_R\}$, qui correspond aux vecteurs colonne de la matrice \mathbf{W}' . Cette nouvelle famille vérifie :

$$\mathcal{C}_{\mathbf{W}} \subsetneq \mathcal{C}_{\mathbf{W}'}$$

Démonstration. Si $x \in \mathcal{C}_{\mathbf{W}}$ alors il existe $(h_1, \dots, h_R) \in (\mathbb{R}^+)^R$ tel que $x = \sum_j h_j \mathbf{w}_j$. On a alors :

$$x = h_i \mathbf{w}_i + \sum_{j \neq i} h_j \mathbf{w}_j = h_i \mathbf{w}'_i + (h_i \lambda_k + h_k) \mathbf{w}_k + \sum_{j \neq i, k} h_j \mathbf{w}_j.$$

Comme $\lambda_k > 0$, on a $(h_i \lambda_k + h_k) > 0$ et par conséquent $x \in \mathcal{C}_{\mathbf{W}'}$. De plus, comme \mathbf{W} est de rang plein, \mathbf{w}'_i n'appartient pas à $\mathcal{C}_{\mathbf{W}}$ et donc, on a bien $\mathcal{C}_{\mathbf{W}} \subsetneq \mathcal{C}_{\mathbf{W}'}$. \square

Cette opération permet donc d'agrandir le cône sans changer le produit $\hat{\mathbf{V}} = \mathbf{WH}$. Cette opération est illustrée dans la figure 2.3 : la figure 2.3(a) représente un cône polyédrique (zone quadrillée en noir) que l'on peut obtenir par NMF de \mathbf{V} dont les vecteurs colonnes sont matérialisés par des croix violettes ; la figure 2.3(b) montre une extension possible du cône qui ne modifie pas la valeur de la matrice $\hat{\mathbf{V}}$.

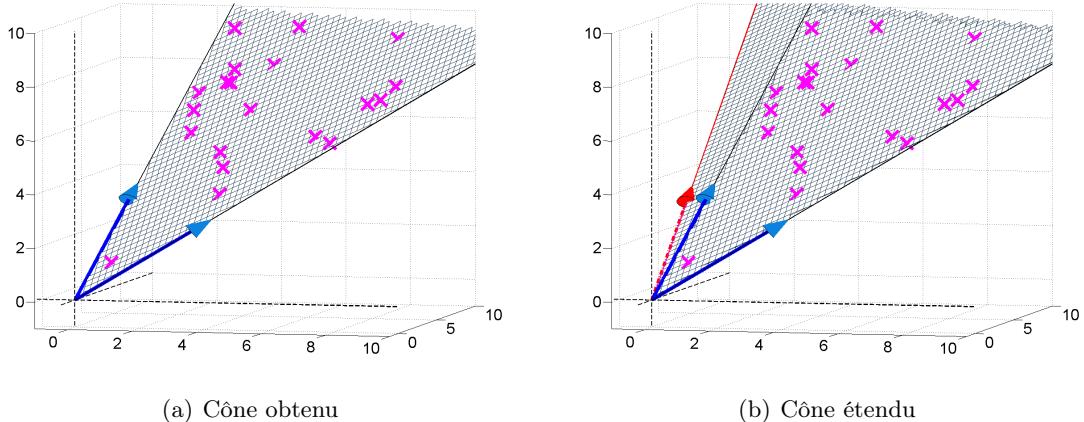


FIG. 2.3 – Extension du cône polyédrique : les flèches bleues pleines correspondent aux vecteurs colonnes de \mathbf{W} . La zone quadrillée est l'espace positivement engendré par ces vecteurs. Les croix violettes correspondent aux vecteurs colonnes de \mathbf{V} (à proximité de l'espace quadrillé). La flèche rouge en pointillé correspond à un vecteur obtenu par extension du cône.

Ce second type de transformation correspond à une matrice \mathbf{Q} dont les coefficients

peuvent être négatifs :

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots \\ 0 & 1 & \dots & 0 & \dots \\ \vdots & \vdots & \dots & \vdots & \dots \\ \vdots & \vdots & \dots & 1 & \dots \\ \vdots & \vdots & \dots & \vdots & \dots \\ \vdots & \vdots & \dots & -\lambda_k & \dots \\ \vdots & \vdots & \dots & \vdots & \dots \end{pmatrix} \quad k \\ i$$

Plus généralement, on peut appliquer des compositions de telles matrices ce qui correspond à des transformations du type :

$$\mathbf{w}'_i = \mathbf{w}_i - \sum_{k \neq i} \lambda_k \mathbf{w}_k,$$

où les λ_k sont des réels positifs tels que les coefficients de \mathbf{w}'_i sont tous positifs. Ce type de transformation peut être réalisé pour tout i .

Il peut également être possible de réduire le cône polyédrique $\mathcal{C}_{\mathbf{W}}$: les λ_k peuvent alors être négatifs. Il faut alors s'assurer que les colonnes de la matrice produit \mathbf{WH} restent bien dans le cône ce qui se traduit par des contraintes d'inégalités linéaires sur les λ_k . Cette opération revient en fait à inverser le rôle de \mathbf{W} et de \mathbf{H} .

Ce type de transformation (modification du cône polyédrique des solutions) est plus problématique vis-à-vis de l'unicité : on peut en effet se demander s'il existe un unique cône polyédrique minimal au sens de l'inclusion. $\text{Vect}(\mathbf{WH}) \cap (\mathbb{R}^+)^F$ (F est la dimension de l'espace) est en effet un cône polyédrique (intersection de deux cônes polyédriques) mais ce cône est généralement engendré par trop de vecteurs pour qu'on puisse utiliser cette famille génératrice pour \mathbf{W} (la factorisation ne serait alors plus une réduction de rang).

Par conséquent, il n'est généralement pas possible de trouver un cône polyédrique englobant tous les cônes polyédriques de solution et donc de garantir l'unicité de la **NMF**, comme le montre l'exemple de la section 2.2.3.

L'unicité standard ne peut donc être vérifiée que si le cône engendré par \mathbf{W} ne peut ni être agrandi (c'est-à-dire que les vecteurs de \mathbf{W} sont au bord de l'orthant positif) ni rétréci (ce qui veut dire que pour chaque vecteur colonne de \mathbf{W} , il existe un vecteur colonne de \mathbf{WH} qui lui soit colinéaire).

2.2.3 Problème lié : impossibilité d'une factorisation exacte

Un problème lié à la non-unicité des solutions est le problème de la **NMF** d'une matrice non-négative \mathbf{V} de rang K sur $R = K$ composantes. Lorsque la factorisation n'a pas de contraintes de non-négativité (comme dans une Analyse en Composantes Principales), le problème est évident et il existe toujours une factorisation exacte, c'est-à-dire telle que $\mathbf{V} = \mathbf{WH}$. En revanche, pour une **NMF**, il n'existe pas nécessairement de factorisation exacte dès que $2 < K < N$ (où N est la dimension de l'espace). Cette propriété assez contre-intuitive remet en cause l'idée d'estimer un bon rang K de factorisation à partir d'une

Décomposition en Valeurs Singulières (**DVS**) de la matrice \mathbf{V} . Le nombre de composantes réellement nécessaires est appelé *rang non-négatif* [Cohen et Rothblum, 1993].

Ainsi même si seules K composantes sont significatives dans la **DVS** (les autres valeurs singulières étant trop faibles pour être prises en compte) de la matrice \mathbf{V} , il n'est pas dit qu'une **NMF** à K composantes de la matrice \mathbf{V} apporte une approximation « correcte », comme le montre l'exemple suivant.

Exemple : La matrice $\mathbf{V} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$ est de rang 3. Pourtant, on peut démontrer

qu'il n'existe pas de cône engendré par une famille de 3 vecteurs non-négatifs contenant les vecteurs colonne de cette matrice. On a en effet $\mathcal{C}_{\mathbf{V}} = \text{Vect}\mathbf{V} \cap (\mathbb{R}^+)^4$ et pourtant aucun des vecteurs colonne de \mathbf{V} n'appartient au cône engendré par les trois autres vecteurs colonne de \mathbf{V} .

Ainsi les cônes respectivement engendrés par $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ et $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4\}$ ne sont pas comparables au sens de l'inclusion (aucun des deux n'est inclus dans l'autre). Pourtant leur intersection contient un cône engendré par au moins trois vecteurs : par exemple $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_5\}$, avec $\mathbf{v}_5 = (2, 1, 1, 2)^T$. Ce qui veut dire que lors d'une **NMF** à 3 éléments, si on obtient la matrice \mathbf{W} dont les vecteurs colonne sont $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_5\}$, alors il existe plusieurs manières d'agrandir le cône (on peut aussi bien aboutir à $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ qu'à $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4\}$) mais aucune n'aboutit à un cône minimal qui n'existe donc pas.

Le problème de l'unicité de la **NMF** est donc complexe. Ce problème semble pourtant assez peu étudié dans la littérature. On peut tout de même citer [Donoho et Stodden, 2004] qui s'y intéresse et notamment traite de l'interprétation géométrique de la **NMF** sous forme de recherche d'un cône polyédrique optimal et également [Cichocki et al., 2009].

2.3 Décomposition de spectrogrammes musicaux

2.3.1 Principe

La factorisation en matrices non-négatives appliquée à des spectrogrammes d'amplitude ou de puissance repose sur l'hypothèse que le spectrogramme est constitué de motifs fréquentiels élémentaires qui se répètent au cours du temps. On dispose d'un spectrogramme d'amplitude (ou de puissance) qui est donc une matrice non-négative \mathbf{V} ($F \times T$) : on cherche alors \mathbf{W} ($F \times R$) et \mathbf{H} ($R \times T$), matrices non-négatives vérifiant l'équation (2.1). Les colonnes \mathbf{w}_r de \mathbf{W} correspondent aux motifs fréquentiels de base, les lignes \mathbf{h}_r^T de \mathbf{H} correspondent aux activités temporelles de chacun de ces motifs. Ainsi $\mathbf{w}_r \mathbf{h}_r^T$ correspond à peu près au spectrogramme d'amplitude qu'on obtiendrait en extrayant tous les sons correspondant au motif \mathbf{w}_r (par exemple tous les sons correspondant à une même note). Un exemple de factorisation en matrices non-négatives d'un spectrogramme d'amplitude est présenté dans la figure 2.4 : il s'agit du spectrogramme d'un signal comprenant deux notes jouées l'une après l'autre puis simultanément. La similitude entre les trames successives durant une même note est flagrante et chaque trame de ce spectrogramme peut être facilement représentée comme une combinaison linéaire de deux motifs harmoniques (chaque trame correspondant à une note). La **NMF** (ici utilisée avec une divergence de **KL**) fournit

effectivement cette représentation (le nombre d'atomes étant fixé à 2) : les colonnes de \mathbf{W} sont bien des motifs harmoniques correspondant chacun à une note, et les lignes de \mathbf{H} mettent en valeurs les instants auxquels chaque note est jouée.

La perception de la musique par l'homme étant principalement basée sur la redondance (redondance interne à un morceau mais aussi redondance d'un morceau à un autre : la perception de cette redondance est alors culturelle) qui structure la musique, la **NMF**, qui permet d'extraire cette redondance, a donc le grand avantage de pouvoir fournir une description des spectrogrammes qui correspond approximativement à ce que l'homme peut percevoir.

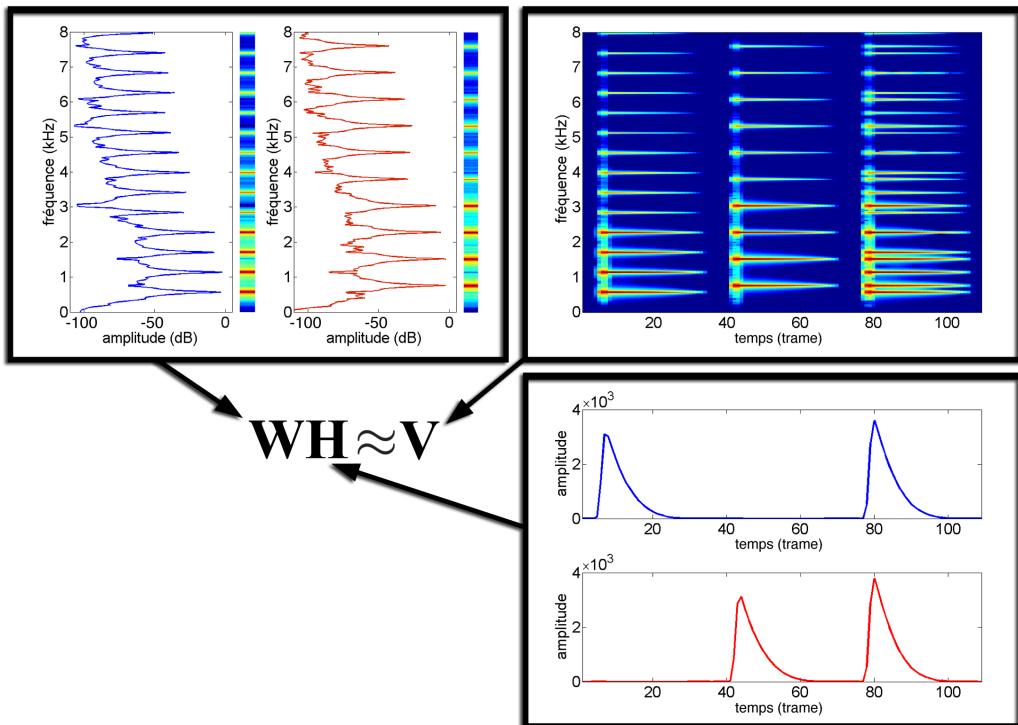


FIG. 2.4 – Factorisation en matrices non-négatives d'un spectrogramme.

2.3.2 Choix de l'exposant

Dans les factorisations en matrices non-négatives de spectrogrammes, on utilise généralement l'hypothèse que la somme des spectrogrammes d'amplitude des différentes composantes élevées à une certaine puissance est approximativement égale au spectrogramme d'amplitude du mélange élevé à cette même puissance :

$$|\mathbf{X}|^{\odot\alpha} \approx \sum_{k=1}^K |\mathbf{X}_k|^{\odot\alpha}, \quad (2.6)$$

où \mathbf{X} est le spectrogramme complexe issu d'une Transformée de Fourier à Court Terme (**TFCT**) du mélange et \mathbf{X}_k celui de la composante k .

Les choix les plus courants d'exposant sont $\alpha = 1$, auquel cas l'hypothèse s'applique aux spectrogrammes d'amplitude, et $\alpha = 2$ auquel cas l'hypothèse s'applique aux spectrogrammes de puissance (ce second cas se rencontre notamment dans le cadre du modèle gaussien présenté dans la section 2.4.1.1). L'hypothèse de l'équation (2.6) n'est en pratique jamais exactement vérifiée. Il peut donc être intéressant de savoir pour quel α cette hypothèse est « la moins fausse ». Si le choix de α est généralement motivé par le modèle utilisé, il semble que des estimations expérimentales d'un « bon » exposant soient rares : dans [Smaragdis *et al.*, 2009], le sujet est abordé très brièvement et les auteurs prétendent qu'un bon exposant est en pratique « plus proche de 1 que de 2 », mais sans donner plus de justification.

Cette section ne prétend pas donner une réponse indiscutable sur le choix de l'exposant α mais propose quelques pistes de réflexion. En pratique, la question ne se pose que pour les points temps/fréquence où plusieurs composantes (par exemple, plusieurs instruments) ont une énergie significative : en effet, si seule une composante (ou aucune) a une énergie significative l'approximation (2.6) est très bonne au point temps/fréquence considéré. Il est à noter qu'en musique, la plupart des points temps/fréquence ne contiennent qu'un instrument avec une énergie significative comme le suggère l'article [Parvaix et Girin, 2011], dans lequel une expérience montre qu'en moyenne, sur les morceaux utilisés, environ 90% de l'énergie provient de l'instrument prédominant dans le point temps/fréquence considéré.

On fait ici l'hypothèse que le spectrogramme complexe \mathbf{X} peut se décomposer linéairement sur une famille de « composantes » :

$$\mathbf{X} = \sum_{k=1}^K \mathbf{X}_k. \quad (2.7)$$

2.3.2.1 Cas à deux composantes « indépendantes »

On suppose que le spectrogramme de mélange est issu de deux composantes. On a donc la relation exacte suivante entre n'importe quel point temps/fréquence x de la TFCT du mélange et les points correspondants de la TFCT des composantes (notés x_1 et x_2), ces trois quantités étant considérées comme des variables aléatoires :

$$x = x_1 + x_2 = |x_1|e^{i\phi_1} + |x_2|e^{i\phi_2} = |x_1|e^{i\phi_1}\left(1 + \frac{|x_2|}{|x_1|}e^{i(\phi_2 - \phi_1)}\right).$$

Par conséquent, on a, pour tout α , la relation sur les amplitudes :

$$|x|^\alpha = |x_1|^\alpha \left|1 + \frac{|x_2|}{|x_1|}e^{i(\phi_2 - \phi_1)}\right|^\alpha.$$

Ces relations sont valables pour chaque point temps/fréquence indépendamment (les indices de temps et de fréquence ont été omis). On se demande donc pour quelle valeur de α , l'approximation suivante est la « meilleure » (dans un certain sens à préciser) :

$$|x|^\alpha = |x_1|^\alpha \left|1 + \frac{|x_2|}{|x_1|}e^{i(\phi_2 - \phi_1)}\right|^\alpha \approx |x_1|^\alpha + |x_2|^\alpha.$$

On note $\mu = \frac{|x_2|}{|x_1|}$ et $\phi = \phi_2 - \phi_1$. L'approximation peut se réécrire :

$$\frac{|1 + \mu e^{i\phi}|^\alpha}{1 + \mu^\alpha} \approx 1.$$

Cette expression est alors symétrique vis-à-vis de x_1 et x_2 . En effet, en échangeant les rôles de x_1 et x_2 , ce qui revient à remplacer μ par $\frac{1}{\mu}$ (et ϕ par $-\phi$), on obtient la même expression.

On a $|1 + \mu e^{i\phi}|^\alpha = (1 + \mu^2 + 2\mu \cos \phi)^{\frac{\alpha}{2}}$.

Supposons que la différence de phase ϕ suit une loi uniforme sur $[0, 2\pi]$. Cette hypothèse sera valable en toute généralité si les deux composantes sont issues d'objets physiques distincts n'interagissant pas physiquement entre eux. Elle est en revanche plus que discutable s'il s'agit de composantes issues d'une même source sonore.

On pourrait alors s'intéresser à l'espérance de $\frac{|1 + \mu e^{i\phi}|^\alpha}{1 + \mu^\alpha}$ pour les différentes valeurs de α (et de μ). Cependant, cette espérance ne donne pas une bonne idée de la validité de l'approximation. En effet pour $\alpha = 2$, on a pour tout μ :

$$\mathbb{E} \left\{ \frac{|1 + \mu e^{i\phi}|^2}{1 + \mu^2} \right\} = 1,$$

ce qui pourrait faire penser que $\alpha = 2$ est un bon choix pour que l'approximation soit à peu près respectée. Mais la distribution de $\left\{ \frac{|1 + \mu e^{i\phi}|^2}{1 + \mu^2} \right\}$ est en réalité faible en 1 et se concentre en fait principalement autour des extrema de la variable, comme le montre la figure 2.5 dans le cas particulier $\mu = 1$.

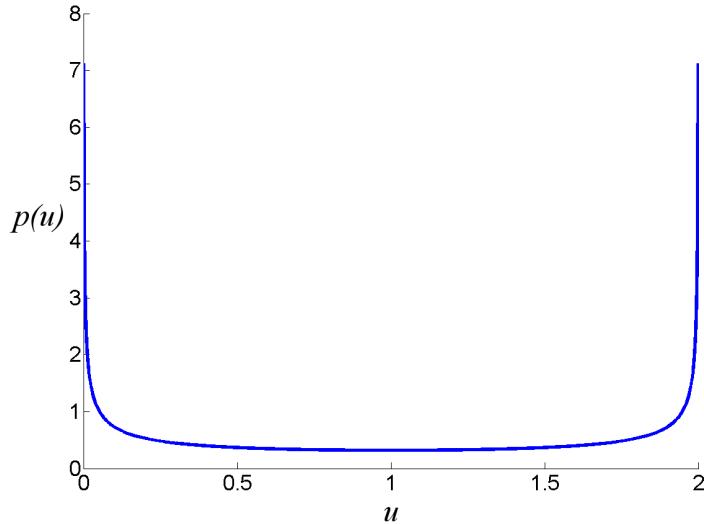


FIG. 2.5 – Distribution de $U = \frac{|1 + \mu e^{i\phi}|^\alpha}{1 + \mu^\alpha}$ pour $\mu = 1$ et $\alpha = 2$: la variable prend ses valeurs dans l'intervalle $[0, 2]$.

Il semble donc plus pertinent de s'intéresser aux valeurs de α pour lesquelles cette distribution se concentre autour de 1.

Le calcul de la densité de probabilité de la variable $U = \frac{|1 + \mu e^{i\phi}|^\alpha}{1 + \mu^\alpha}$ (qui prend ses valeurs

dans $[\frac{|1-\mu|^\alpha}{1+\mu^\alpha}, \frac{|1+\mu|^\alpha}{1+\mu^\alpha}]$) donne :

$$p(u) = \frac{(1 + \mu^\alpha)((1 + \mu^\alpha)u)^{\frac{2}{\alpha}-1}}{\mu\alpha\pi\sqrt{1 - \left(\frac{((1 + \mu^\alpha)u)^{\frac{2}{\alpha}} - (1 + \mu^2)}{2\mu}\right)^2}}.$$

On peut mesurer la concentration de la densité de U autour de 1 en estimant $\mathbb{E}\{(U-1)^2\}$ (sorte d'erreur quadratique moyenne) ou bien $\mathbb{E}\{|U - 1|\}$. Il semble difficile de calculer analytiquement l'expression de ces quantités pour tout α et pour tout μ . On peut cependant les estimer à l'aide de méthodes de Monte Carlo.

Critère $\mathbb{E}\{(U - 1)^2\}$: La figure 2.6 donne une estimation par méthode de Monte-Carlo de $\mathbb{E}\{(U - 1)^2\}$ pour différentes valeurs de μ et α . On constate que $\mathbb{E}\{(U - 1)^2\}$ ne prend des valeurs importantes que pour des valeurs de μ proches de 1 (approximativement entre 0.1 et 10), c'est-à-dire lorsque les deux composantes X_1 et X_2 ont une amplitude comparable. Ce résultat était prévisible : si une des composantes est négligeable par rapport à l'autre (i.e. $\mu \rightarrow 0$ ou $\mu \rightarrow \infty$) alors U tend vers 1 pour tout α . Si X_2 peut être négligé par rapport à X_1 alors on a bien pour tout α : $|X|^\alpha \approx |X_1|^\alpha \approx |X_1|^\alpha + |X_2|^\alpha$.

Pour μ proche de 1, on constate que les valeurs de α qui favorisent la concentration de U autour de 1 sont comprises entre 0.9 et 1.2. On constate que selon ce critère, il n'y a pas de valeur optimale de α pour tout μ : pour chaque μ , on trouve une valeur optimale différente.

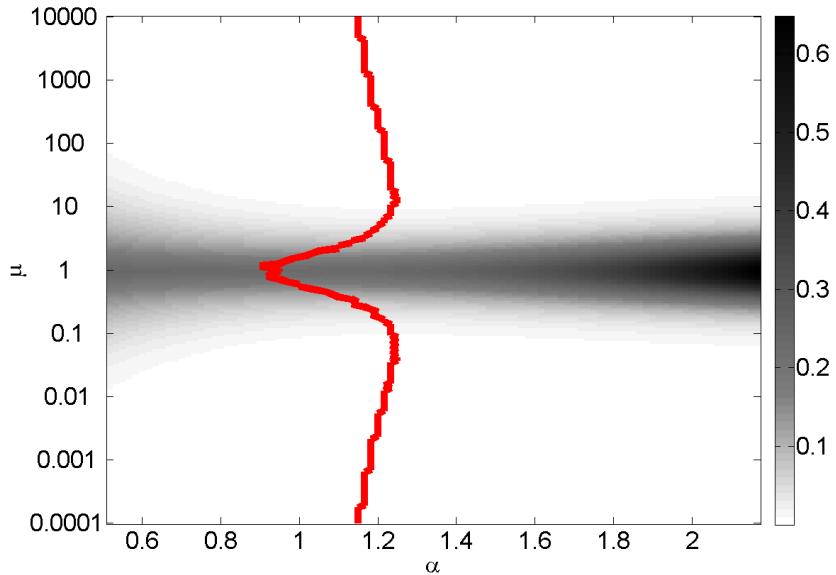


FIG. 2.6 – $\mathbb{E}\{(U - 1)^2\}$ en fonction de α et μ . La courbe rouge représente le α qui minimise $\mathbb{E}\{(U - 1)^2\}$ pour chaque μ .

Critère $\mathbb{E}\{|U - 1|\}$: La figure 2.7 donne une estimation par méthode de Monte-Carlo de $\mathbb{E}\{|U - 1|\}$ pour différentes valeurs de μ et α . Les remarques sur la concentration de U autour de 1 lorsqu'une des deux composantes est négligeable restent évidemment valables.

Il semble que selon ce deuxième critère, on obtienne la même valeur optimale de α pour tout μ . Cette valeur est approximativement 1.12.

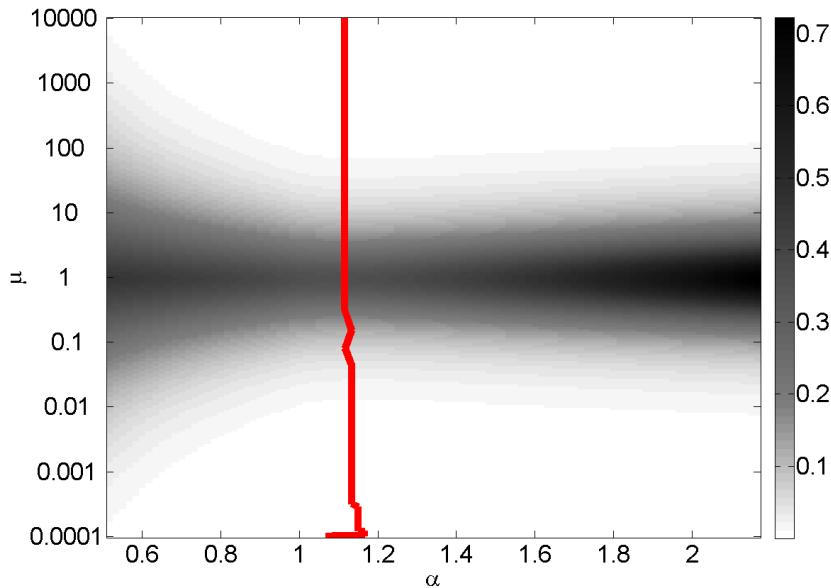


FIG. 2.7 – $\mathbb{E}\{|U - 1|\}$ en fonction de α et μ . La courbe rouge représente le α qui minimise $\mathbb{E}\{|U - 1|\}$ pour chaque μ . Remarque : la courbe rouge devrait être, en théorie, symétrique sur le graphe (la distribution de U est en effet inchangée en remplaçant μ par $\frac{1}{\mu}$). La légère asymétrie dans le graphe est due aux erreurs d'estimation.

Bien qu'il soit difficile de donner une réponse tranchée au problème du choix de l'exposant, les deux critères présentés, bien que discutables, suggèrent qu'une « bonne valeur » de α pour un mélange à deux composantes se trouve autour de 1.

2.3.2.2 Autres cas

Plus de 2 composantes : Il semble intuitif que si l'approximation (2.6) est bonne pour 2 composantes pour une certaine valeur de α , alors cette approximation restera bonne pour plus de composantes. En effet, pour 3 composantes, on aura :

$$|x_1|^\alpha + |x_2|^\alpha + |x_3|^\alpha \approx |x_1 + x_2|^\alpha + |x_3|^\alpha \approx |x_1 + x_2 + x_3|^\alpha,$$

et, par induction, pour tout K :

$$\left| \sum_{k=1}^K x_k \right|^\alpha \approx \sum_{k=1}^K |x_k|^\alpha.$$

Cependant, il ne s'agit pas d'une démonstration du fait que l'approximation n'est pas quantifiée.

Composantes non-indépendantes : Lorsque deux composantes sont issues d'un même objet physique (ou d'objets qui interagissent fortement entre eux), l'hypothèse faite sur les phases dans la section 2.3.2.1 ne semble plus justifiée. Il semble beaucoup plus difficile de faire une hypothèse générale réaliste dans ce cas. Le problème est en fait assez profond et revient à s'interroger sur la nature même d'une composante dans le modèle linéaire (2.7). Tant que les composantes ont réellement un sens physique (par exemple chaque composante correspond à un instrument), la nature d'une composante est bien délimitée. Quand ce n'est plus le cas, une composante n'est plus qu'un outil de travail, et il est alors extrêmement difficile de faire une hypothèse sur les phases des différentes composantes qui ait un réel sens physique.

2.3.3 NMF et séparation de sources

Bien que la phase ne soit pas prise en compte dans les NMF de spectrogrammes, il est possible de reconstruire un signal dans le domaine temporel correspondant aux différentes composantes factorisées par des techniques de masquages temps/fréquence. Supposons par exemple qu'on ait réalisé la NMF d'un spectrogramme d'amplitude \mathbf{V} , issue de la TFCT \mathbf{X} du signal \mathbf{x} : $\mathbf{V} \approx \mathbf{WH}$, et qu'on soit capable de regrouper les composantes en deux classes \mathcal{A} et \mathcal{B} correspondant chacune par exemple à une source sonore. On note alors $\mathbf{W}_\mathcal{A}$ la matrice des atomes associés à la classe \mathcal{A} , $\mathbf{H}_\mathcal{A}$ la matrice d'activation correspondante, $\mathbf{W}_\mathcal{B}$ la matrice des atomes associés à la classe \mathcal{B} et $\mathbf{H}_\mathcal{B}$ la matrice d'activation correspondante, de telle sorte qu'on a $\mathbf{V} \approx \mathbf{WH} = \mathbf{W}_\mathcal{A}\mathbf{H}_\mathcal{A} + \mathbf{W}_\mathcal{B}\mathbf{H}_\mathcal{B}$.

On peut alors reconstruire un signal temporel en faisant une TFCT inverse des matrices suivantes :

$$\begin{aligned}\mathbf{X}_\mathcal{A} &= \mathbf{X} \odot \frac{\mathbf{W}_\mathcal{A}\mathbf{H}_\mathcal{A}}{\mathbf{WH}} \xrightarrow{\text{TFCT inverse}} \tilde{\mathbf{x}}_\mathcal{A}, \\ \mathbf{X}_\mathcal{B} &= \mathbf{X} \odot \frac{\mathbf{W}_\mathcal{B}\mathbf{H}_\mathcal{B}}{\mathbf{WH}} \xrightarrow{\text{TFCT inverse}} \tilde{\mathbf{x}}_\mathcal{B}.\end{aligned}$$

On obtient alors des signaux temporels $\tilde{\mathbf{x}}_\mathcal{A}$ et $\tilde{\mathbf{x}}_\mathcal{B}$ qui, par linéarité de la TFCT inverse, vérifient $\mathbf{x} = \tilde{\mathbf{x}}_\mathcal{A} + \tilde{\mathbf{x}}_\mathcal{B}$.

Il est à noter que les matrices complexes $\mathbf{X}_\mathcal{A}$ et $\mathbf{X}_\mathcal{B}$ ne sont généralement pas les TFCT de signaux temporels et que l'opération de TFCT inverse n'est que formelle.

Ce principe se généralise facilement à un nombre quelconque de classes à séparer.

Cette méthode de masquage très générale est inspirée du filtrage de Wiener variable dans le temps qui fournit un cadre théorique solide : si $\mathbf{W}_\mathcal{A}\mathbf{H}_\mathcal{A}$ est la puissance spectrale estimée du signal $\mathbf{x}_\mathcal{A}$, alors $\tilde{\mathbf{x}}_\mathcal{A}$ est un estimateur de $\mathbf{x}_\mathcal{A}$ qui minimise l'erreur quadratique moyenne. Ces hypothèses sont rencontrées en NMF dans le cadre de la formulation probabiliste présentée dans la section 2.4.1.1 : dans ce cadre, la séparation de source par la méthode présentée ci-dessus consiste en un filtrage de Wiener.

Bien que cette dénomination soit ambiguë, on appellera par la suite filtrage de Wiener toute méthode de masquage temps/fréquence à masques continus.

2.4 Modélisation probabiliste

Le problème de factorisation en matrices non-négatives peut être formalisé en terme statistique soit via un modèle génératif [Févotte *et al.*, 2009, Virtanen *et al.*, 2008] comme

présenté dans la section 2.4.1, soit via un modèle de « comptage » [Shashanka *et al.*, 2007], comme présenté dans la section 2.4.2. Il existe également d'autres modèles statistiques pour la NMF (par exemple un modèle avec bruit additif gaussien [Schmidt et Laurberg, 2008] ou bruit multiplicatif de loi Gamma [Févotte *et al.*, 2009]) qui ne seront pas présentés dans ce document.

2.4.1 Modèles génératifs

2.4.1.1 Modèle gaussien

Dans ce modèle, la t -ème colonne du spectrogramme complexe \mathbf{X} , notée \mathbf{x}_t , est supposée être la somme de R variables latentes complexes \mathbf{c}_r^t :

$$\mathbf{x}_t = \sum_{r=1}^R \mathbf{c}_r^t.$$

Les vecteurs \mathbf{c}_r^t ($r = 1 \dots R$) sont indépendamment distribués et suivent une loi gaussienne complexe circulaire centrée de variance $h_{rt}\text{diag}(\mathbf{w}_r)$:

$$\mathbf{c}_r^t \propto \mathcal{N}_c(0, h_{rt}\text{diag}(\mathbf{w}_r)).$$

On note \mathbf{V} la matrice de coefficients $[\mathbf{V}]_{ft} = |[\mathbf{X}]_{ft}|^2$.

La log-vraisemblance de \mathbf{X} connaissant les paramètres \mathbf{W} (matrice formée des vecteurs \mathbf{w}_r) et $\mathbf{H} = (h_{rt})_{rt}$ est alors :

$$\mathcal{L}(\mathbf{X}|\mathbf{W}, \mathbf{H}) = \sum_{f=1}^F \sum_{t=1}^T \log \left(\mathcal{N}_c \left([\mathbf{X}]_{ft} | 0, \sum_{r=1}^R [\mathbf{W}]_{fr} [\mathbf{H}]_{rt} \right) \right).$$

Comme démontré dans [Févotte *et al.*, 2009], cette log-vraisemblance peut se réécrire :

$$\mathcal{L}(\mathbf{X}|\mathbf{W}, \mathbf{H}) = - \sum_{f=1}^F \sum_{t=1}^T d_{\text{IS}}([\mathbf{X}]_{ft} | [\mathbf{WH}]_{ft}) + c,$$

où c est une constante. Par conséquent, l'estimation du maximum de vraisemblance de \mathbf{W} et de \mathbf{H} est équivalente à la factorisation en matrices non-négatives de \mathbf{V} en utilisant comme fonction de coût la distance d'**IS**.

Remarque : Ce modèle est, semble-t-il, le seul à prendre en compte explicitement la phase des spectrogrammes même si cette information de phase n'est pas utilisée. Cependant, le modèle permet de générer n'importe quelle matrice complexe et ne prend pas en compte les contraintes liées au calcul d'un spectrogramme avec recouvrement temporel (lorsque les trames successives ne sont pas disjointes). Ainsi les spectrogrammes produits ne sont pas constants (*cf.* [Le Roux *et al.*, 2010]).

2.4.1.2 Modèle de Poisson

Ce modèle est conceptuellement assez proche du modèle gaussien présenté dans la section 2.4.1.1 mais il modélise directement des données non-négatives et ne prend donc pas en considération la phase des spectrogrammes complexes. Dans ce second modèle proposé dans [Virtanen *et al.*, 2008], on suppose que la colonne t de la matrice non-négative \mathbf{V} résulte de R composantes latentes :

$$[\mathbf{V}]_{f,t} = \sum_{r=1}^R c_{f,t}^r.$$

où les composantes $c_{f,t}^r$ sont indépendamment distribuées suivant une loi de Poisson :

$$c_{f,t}^r \sim \mathcal{P}(w_{fr}h_{rt}).$$

La loi de Poisson étant définie sur des entiers, on suppose donc que les composantes prennent des valeurs entières et donc que le spectrogramme prend des valeurs entières. Cette hypothèse n'est pas réellement un problème du fait qu'on peut multiplier le spectrogramme par un facteur très grand, de sorte que la précision numérique soit de l'ordre de l'entier et qu'approximer à l'entier le plus proche soit alors raisonnable.

La somme de variables indépendantes suivant une loi de Poisson suit elle-même une loi de Poisson de paramètre la somme des paramètres. Par conséquent, v_{ft} a pour probabilité :

$$v_{ft} \sim \mathcal{P}\left(\sum_{r=1}^R w_{fr}h_{rt}\right).$$

Par conséquent, la log-vraisemblance de \mathbf{V} est donnée par :

$$L(\mathbf{V}|\mathbf{W}, \mathbf{H}) = \sum_{f,t} -[\mathbf{WH}]_{f,t} + [\mathbf{V}]_{f,t} \log([\mathbf{WH}]_{f,t}) - \log(([V]_{f,t})!). \quad (2.8)$$

L'expression 2.8 est égale à une constante près à l'opposé de la divergence de **KL** entre \mathbf{V} et \mathbf{WH} . Par conséquent l'estimation du maximum de vraisemblance de ce modèle est équivalente à une **NMF** avec pour fonction de coût la divergence de **KL**.

2.4.2 Analyse probabiliste en composantes latentes (**PLCA**)

En **PLCA** [Shashanka *et al.*, 2007], la matrice de données non-négative $\mathbf{V} = (v_{ft})_{ft}$ est considérée comme un histogramme issu du tirage structuré des variables aléatoires f et t qui suivent une loi jointe $P(f, t)$. Suivant la forme donnée à $P(f, t)$, on peut obtenir des décompositions différentes. Lorsqu'on décompose un spectrogramme, la variable f est une variable de fréquence et la variable t une variable de temps.

Le modèle de base structure le tirage en introduisant une variable aléatoire « Composante » z cachée (latente). f et t sont alors supposées indépendantes connaissant z :

$$P(f, t|z) = P(f|z)P(t|z).$$

On a donc :

$$P(f, t) = \sum_{z=1}^Z P(z)P(f|z)P(t|z).$$

L'histogramme $\mathbf{V} = (v_{ft})_{ft}$ est donc supposé obtenu par N tirages indépendants de z suivant $P(z)$, chaque tirage de z étant suivi d'un tirage de f suivant $P(f|z)$ et de t suivant $P(t|z)$.

Les paramètres de ce modèle sont donc $P(f|z)$, $P(t|z)$ et $P(z)$ pour $t \in \{1, \dots, T\}$, $f \in \{1, \dots, F\}$ et $z \in \{1, \dots, Z\}$. On note θ l'ensemble de tous ces paramètres. En notant $(z_i, f_i, t_i)_{i \in \{1, \dots, N\}}$ l'ensemble des résultats des tirages effectués, la log-vraisemblance des observations est donnée par :

$$L(\theta) = \sum_{i=1}^N \log(P(f_i, t_i)). \quad (2.9)$$

Comme \mathbf{V} est l'histogramme obtenu par ces tirages, (f, t) apparaît v_{ft} fois dans la somme de l'équation (2.9) et donc cette somme peut se réécrire :

$$\begin{aligned} L(\theta) &= \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log(P(f, t)) \\ &= \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log \left(\sum_{z=1}^Z P(z) P(f|z) P(t|z) \right). \end{aligned} \quad (2.10)$$

On obtient alors facilement :

$$\begin{aligned} L(\theta) &= \sum_{f=1}^F \sum_{t=1}^T \left\{ v_{ft} \left(\log \left(\sum_{z=1}^Z P(z) P(f|z) P(t|z) \right) - \log v_{ft} \right) - P(f, t) + v_{ft} \right\} + c \\ &= - \sum_{f=1}^F \sum_{t=1}^T d_{KL} \left(v_{ft} \mid \sum_{z=1}^Z P(z) P(f|z) P(t|z) \right) + c, \end{aligned}$$

où $c = \sum_{f,t} \{v_{ft} \log v_{ft} - v_{ft}\} + \sum_{f,t} P(f, t) = \sum_{f,t} \{v_{ft} \log v_{ft} - v_{ft}\} + 1$ est une constante (c ne dépend pas des paramètres).

En posant par exemple $\mathbf{W} = (P(f|z))_{f,z}$ et $\mathbf{H} = (P(z)P(t|z))_{t,z}$, l'estimation du maximum de vraisemblance de ce modèle est donc équivalente à une **NMF** avec comme critère de coût la divergence de **KL**. On démontre aisément que le fait que $\mathbf{WH} = (P(f, t))_{f,t}$ soit normalisé (la somme des coefficients est égale à 1) contrairement à une **NMF** standard n'a pas d'influence sur le problème (le coefficient de normalisation optimal se calcule aisément et vaut $\sum_{f,t} v_{ft}$). Ainsi $P(f|z)$ correspond aux motifs spectraux normalisés et $P(t|z)$ aux activations normalisées de chaque composante. $P(z)$ correspond au poids relatif de chaque composante.

Comme dans le modèle de la section 2.4.1.2, le spectrogramme d'amplitude \mathbf{V} est censé ne prendre que des valeurs entières : ici encore, cette hypothèse n'est pas réellement un problème car on peut multiplier \mathbf{V} par un facteur arbitrairement grand sans changer la valeur de $\arg \min_{\theta} L(\theta)$. Par conséquent, en choisissant un facteur suffisamment grand, on peut ramener la précision numérique à l'ordre de l'entier. En pratique ce facteur disparaît dans les règles de mises à jour de l'algorithme Espérance-Maximisation utilisé pour minimiser la vraisemblance et n'a par conséquent pas à être pris en compte.

2.5 Algorithmes

Les fonctions de coût utilisées en NMF ne sont pas conjointement convexes par rapport à \mathbf{W} et \mathbf{H} mais dans certains cas (par exemple dans le cas où la divergence d de l'équation (2.3) est convexe par rapport à sa seconde variable, ce qui est le cas pour une β -divergence pour $\beta \in [1, 2]$) elles sont convexes par rapport à \mathbf{W} à \mathbf{H} fixée et convexes par rapport à \mathbf{H} à \mathbf{W} fixée. Une minimisation alternée par rapport à chacune de ces deux matrices est donc quasi-systématiquement envisagée.

Parmi les différents algorithmes rencontrés dans la littérature, on peut notamment citer :

- Algorithme à mises à jour multiplicatives (*cf. section 2.5.2*).
- Algorithme de descente de gradient projeté (*cf. section 2.5.1.1*).
- Algorithme de Newton projeté (*cf. section 2.5.1.2*).
- Algorithme de moindres carrés (*cf. section 2.5.1.3*).
- Algorithme de descente de gradient sur une reparamétrisation du problème (*cf. section 2.5.1.4*).

2.5.1 Algorithmes divers

2.5.1.1 Descente de gradient projeté

Il s'agit d'une méthode de gradient projeté classique avec projection sur l'orthant positif appliquée successivement à \mathbf{W} et \mathbf{H} . Les règles de mise à jour sont donc :

$$\mathbf{W} \leftarrow [\mathbf{W} - \eta_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{C}]^+,$$

$$\mathbf{H} \leftarrow [\mathbf{H} - \eta_{\mathbf{H}} \nabla_{\mathbf{H}} \mathcal{C}]^+,$$

où $\nabla_{\mathbf{A}}$ représente l'opérateur gradient par rapport à la matrice \mathbf{A} et $[\cdot]^+$ représente l'opérateur de projection sur l'orthant positif.

Les pas $\eta_{\mathbf{W}}$ (et $\eta_{\mathbf{H}}$) peuvent être choisis de plusieurs façons :

- par une recherche linéaire pour minimiser $\mathcal{C}(\mathbf{W} - \eta_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{C}, \mathbf{H})$. Lorsque la divergence utilisée pour la fonction de coût n'est pas définie en dehors de l'orthant positif, le pas ne peut prendre des valeurs que dans un intervalle borné de \mathbb{R} et par conséquent on peut utiliser une méthode de recherche par section d'or (*golden section search* [Forsythe et al., 1976]). Cette méthode est cependant assez lente (en termes de temps de calcul).
- par minimisation de la parabole tangente à $\mathcal{C}(\mathbf{W} - \eta_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{C}, \mathbf{H})$ (ce qui nécessite que cette fonction soit convexe), ce qui est nettement plus rapide qu'une méthode de recherche « exhaustive ».
- par utilisation d'un pas vérifiant les critères de Wolfe (comme dans [Lin, 2007] qui n'utilise en fait que la moitié des critères de Wolfe : le critère d'Armijo).
- par utilisation d'un pas fixe : le coût de calcul d'une itération est alors très faible, mais les résultats sont assez pauvres (la convergence est généralement lente) et rien n'assure que le critère soit décroissant à chaque itération. On peut cependant remarquer que pour une fonction de coût EUC, \mathcal{C} est α -convexe et de gradient K -lipschitzien donc des itérations de gradient projeté par rapport à \mathbf{W} seulement (resp. \mathbf{H} seulement) avec un pas compris entre 0 et $2\alpha/K^2$ convergent.

2.5.1.2 Méthode de Newton projetée

Il s'agit d'une méthode de deuxième ordre : la direction de descente est obtenue par la méthode de Newton à partir de la matrice Hessienne de la fonction de coût (*cf.* [Zdunek et Cichocki, 2007]) :

$$\begin{aligned}\mathbf{W} &\leftarrow [\mathbf{W} - (\nabla_{\mathbf{W}}^2 \mathcal{C})^{-1} \nabla_{\mathbf{W}} \mathcal{C}]^+, \\ \mathbf{H} &\leftarrow [\mathbf{H} - (\nabla_{\mathbf{H}}^2 \mathcal{C})^{-1} \nabla_{\mathbf{H}} \mathcal{C}]^+, \end{aligned}$$

où $\nabla_{\mathbf{W}}^2 \mathcal{C}$ est la matrice Hessienne de \mathcal{C} par rapport à \mathbf{W} .

Un pas peut éventuellement être ajouté pour assurer une meilleure décroissance.

Ce type de méthode nécessite que la fonction soit convexe par rapport à \mathbf{W} et par rapport à \mathbf{H} sinon l'algorithme peut converger vers un maximum local de la fonction de coût, ou diverger. La convergence théorique n'est assurée que dans des conditions assez strictes sur la divergence utilisée et uniquement dans un voisinage d'un point d'équilibre.

En pratique la convergence de ce type de méthode est beaucoup plus rapide en termes de nombre d'itérations qu'une descente de gradient projeté (la direction de descente obtenue à partir de l'approximation d'ordre 2 de la fonction de coût étant « meilleure » que le gradient). Cependant le coût d'une itération est nettement plus élevé pour cette méthode.

2.5.1.3 Moindres carrés alternés

Il s'agit de la première méthode proposée [Paatero et Tapper, 1994] pour un algorithme de **NMF**. Cette méthode vise à minimiser une fonction de coût construite à partir de la distance **EUC** : elle effectue une minimisation des moindres carrés projetée successivement par rapport à \mathbf{W} puis à \mathbf{H} . Cette méthode est donc équivalente à la méthode de Newton pour un coût **EUC** (section 2.5.1.2).

2.5.1.4 Méthode non contrainte par reparamétrisation du problème

Un des problèmes des algorithmes de **NMF** est lié à la contrainte de non-négativité. Il est possible de se débarrasser de cette contrainte en reparamétrisant le problème. En utilisant deux fonctions $f_{\mathbf{W}} : \mathbb{R} \rightarrow \mathbb{R}^+$ et $f_{\mathbf{H}} : \mathbb{R} \rightarrow \mathbb{R}^+$ appliquées composante par composante à des matrices \mathbf{A} et \mathbf{B} , on peut réécrire le problème de **NMF** de la façon suivante :

$$\underset{\mathbf{A}, \mathbf{B}}{\operatorname{argmin}} \quad \mathcal{C}(\mathbf{A}, \mathbf{B}) \quad \text{avec} \quad \mathcal{C}(\mathbf{A}, \mathbf{B}) = D(\mathbf{V} || f_{\mathbf{W}}(\mathbf{A})f_{\mathbf{H}}(\mathbf{B})), \quad (2.11)$$

ce qui revient à reparamétriser le problème de la façon suivante :

$$\mathbf{W} = f_{\mathbf{W}}(\mathbf{A}) \quad \text{et} \quad \mathbf{H} = f_{\mathbf{H}}(\mathbf{B}).$$

Pour que les problèmes fournissent des solutions équivalentes, il est nécessaire que les fonctions de reparamétrisation soient surjectives. Pour pouvoir utiliser des méthodes de minimisation classique, il est nécessaire que ces fonctions soient suffisamment régulières et notamment qu'elles soient continues. Par conséquent, la bijectivité des fonctions de paramétrisation n'est possible que de \mathbb{R} dans \mathbb{R}_+^* . La notation de l'équation (2.11) n'est alors plus correcte puisqu'on peut avoir un infimum qui n'est pas nécessairement atteint (la valeur nulle ne peut être atteinte). La bijectivité de la fonction de reparamétrisation permet de garder l'unimodalité par rapport à \mathbf{W} (resp. à \mathbf{H}) quand la fonction de coût

est unimodale par rapport à chacune de ses matrices. En revanche même si la fonction de reparamétrisation est convexe, rien n'assure que la fonction de coût le reste. Dans le cas d'une reparamétrisation non-bijective, l'unimodalité par rapport à \mathbf{W} et \mathbf{H} n'est plus assurée.

Le gradient de la fonction de coût \mathcal{C} par rapport à \mathbf{A} est :

$$\nabla_{\mathbf{A}} \mathcal{C}(\mathbf{A}, \mathbf{B}) = f'_{\mathbf{W}}(\mathbf{A}) \odot (\mathbf{D}\mathbf{H}^T),$$

où \mathbf{D} est la matrice de coefficients :

$$[\mathbf{D}]_{ft} = \frac{\partial d}{\partial y}(v_{ft}, \hat{v}_{ft}).$$

On a notamment pour une β -divergence :

$$\mathbf{D} = \hat{\mathbf{V}}^{\cdot, \beta-2} \odot (\hat{\mathbf{V}} - \mathbf{V})$$

et pour une divergence de Bregman D_ϕ :

$$\mathbf{D} = \phi(\hat{\mathbf{V}})'' \odot (\hat{\mathbf{V}} - \mathbf{V}).$$

Pour un algorithme de descente de gradient classique, la règle de mise à jour de \mathbf{A} s'écrit alors :

$$\mathbf{A} \leftarrow \mathbf{A} - \eta f'_{\mathbf{W}}(\mathbf{A}) \odot (\mathbf{D}\mathbf{H}^T)$$

et donc, la règle de mise à jour de \mathbf{W} s'écrit :

$$\mathbf{W} \leftarrow f_{\mathbf{W}}(\mathbf{A} - \eta f'_{\mathbf{W}}(\mathbf{A}) \odot (\mathbf{D}\mathbf{H}^T)).$$

En utilisant la fonction exponentielle comme fonction de reparamétrisation, on retrouve la mise à jour multiplicative des algorithmes SMART-NMF [Cichocki *et al.*, 2006a] :

$$\begin{aligned} \mathbf{W} &\leftarrow \exp(\mathbf{A} - \eta f'_{\mathbf{W}}(\mathbf{A}) \odot (\mathbf{D}\mathbf{H}^T)) \\ &= \mathbf{W} \odot \exp(-\eta \mathbf{W} \odot (\mathbf{D}\mathbf{H}^T)). \end{aligned}$$

Cependant dans [Cichocki *et al.*, 2006a], les pas ne sont pas les mêmes pour tous les coefficients : la direction de descente n'est donc pas l'opposée de celle du gradient et il ne s'agit donc pas d'une méthode de descente de gradient.

De même, en utilisant la fonction carré comme fonction de reparamétrisation, on retrouve l'expression du gradient donné dans [Chu *et al.*, 2004] et on peut en déduire la règle de mise à jour pour un algorithme de descente de gradient :

$$\mathbf{W} \leftarrow \mathbf{W} \odot (1 - 2\eta(\mathbf{D}\mathbf{H}^T))^{\odot 2}.$$

2.5.2 Mises à jour multiplicatives

Les algorithmes de mises à jour multiplicatives initialement proposés dans [Lee et Seung, 1999, 2000] pour des coûts EUC et KL sont très utilisés du fait de leur grande simplicité d'implémentation et de la rapidité de calcul des itérations. Des généralisations à des classes de divergence plus grandes ont été proposées dans [Dhillon et Sra, 2006] (divergence de Bregman) et [Cichocki *et al.*, 2006b] (β -divergence). On rencontre deux approches de ce type d'algorithme dans la littérature : une première approche simple et assez heuristique consiste à écrire la dérivée de la fonction de coût comme la différence de deux termes positifs et à construire la règle à partir de ces deux termes. Une seconde approche, plus élaborée, correspond aux algorithmes de type Majoration/Minimisation (MM).

2.5.2.1 Approche simple

Une approche simple mais heuristique des algorithmes à règles multiplicatives couramment utilisée en NMF [Lee et Seung, 1999, Févotte *et al.*, 2009, Smaragdis, 2004] consiste à écrire la dérivée de la fonction de coût par rapport au paramètre θ comme la différence de deux termes strictement positifs :

$$\frac{\partial \mathcal{C}}{\partial \theta} = p_\theta - m_\theta. \quad (2.12)$$

La règle de mise à jour de θ est alors :

$$\theta \leftarrow \theta \times \frac{m_\theta}{p_\theta}. \quad (2.13)$$

Ce type de mise à jour garantit que :

- θ reste non-négatif, puisque m_θ et p_θ sont positifs.
- θ devient constant si la dérivée partielle s'annule : en effet, on a alors $m_\theta = p_\theta$.
- θ évolue dans une direction de descente locale. En effet, si $\frac{\partial \mathcal{C}}{\partial \theta} > 0$ alors $\frac{m_\theta}{p_\theta} < 1$, par conséquent θ décroît et évolue donc dans la direction de $-\frac{\partial \mathcal{C}}{\partial \theta}$; réciproquement, si $\frac{\partial \mathcal{C}}{\partial \theta} < 0$ alors $\frac{m_\theta}{p_\theta} > 1$, par conséquent θ croît et évolue donc également dans la direction de $-\frac{\partial \mathcal{C}}{\partial \theta}$.
- θ est un point fixe de l'algorithme si et seulement si $\theta = 0$ ou la dérivée partielle s'annule.

Dans le cadre de la NMF, on exprime le gradient de la fonction de coût par rapport à \mathbf{W} (resp. \mathbf{H}) directement comme une différence de matrices à coefficients strictement positifs :

$$\nabla_{\mathbf{W}} \mathcal{C} = \mathbf{P}_{\mathbf{W}} - \mathbf{M}_{\mathbf{W}},$$

$$\nabla_{\mathbf{H}} \mathcal{C} = \mathbf{P}_{\mathbf{H}} - \mathbf{M}_{\mathbf{H}}.$$

Les règles de mise à jour sont alors :

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{M}_{\mathbf{W}}}{\mathbf{P}_{\mathbf{W}}},$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{M}_{\mathbf{H}}}{\mathbf{P}_{\mathbf{H}}}.$$

Remarque : Cette méthode est souvent qualifiée dans la littérature de descente de gradient à pas adaptatif. On peut en effet réécrire les règles multiplicatives sous forme additive ce qui rappelle une descente de gradient classique :

$$\mathbf{W} \leftarrow \mathbf{W} - \frac{\mathbf{W}}{\mathbf{P}_{\mathbf{W}}} \odot \nabla_{\mathbf{W}} \mathcal{C}.$$

Cette comparaison est cependant assez trompeuse car les pas ne sont pas les mêmes pour tous les coefficients du gradient et par conséquent la direction de descente ne correspond généralement pas à la direction (opposée) du gradient.

En utilisant une divergence de Bregman (associée à la fonction F) comme critère de coût, on obtient ainsi les règles proposées dans [Dhillon et Sra, 2006] :

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(F''(\mathbf{WH}) \odot \mathbf{V})\mathbf{H}^T}{(F''(\mathbf{WH}) \odot \mathbf{WH})\mathbf{H}^T},$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T(F''(\mathbf{WH}) \odot \mathbf{V})}{\mathbf{W}^T(F''(\mathbf{WH}) \odot \mathbf{WH})}.$$

En utilisant l'équation (2.5), on peut en déduire directement les règles de mise à jour proposées dans [Cichocki et al., 2006b] pour une β -divergence :

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{((\mathbf{WH})^{\odot(\beta-2)} \odot \mathbf{V})\mathbf{H}^T}{(\mathbf{WH})^{\odot(\beta-1)}\mathbf{H}^T}, \quad (2.14)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T((\mathbf{WH})^{\odot(\beta-2)} \odot \mathbf{V})}{\mathbf{W}^T(\mathbf{WH})^{\odot(\beta-1)}}. \quad (2.15)$$

Il a été démontré dans [Kompass, 2007] que ces dernières règles de mise à jour font chacune décroître la fonction de coût lorsque $\beta \in [1, 2]$: la démonstration est basée sur le fait que ces règles de mise à jour peuvent être interprétées comme les règles de mise à jour d'un algorithme **MM** (algorithme présenté section 2.5.2.2). Cette démonstration a récemment été étendue à $\beta \in [0, 1[$ dans [Févotte et Idier, 2011] et avec une version légèrement modifiée des règles de mise à jour (ajout d'un exposant) dans [Nakano et al., 2010a].

2.5.2.2 Algorithme MM

Il s'agit d'une approche plus propre pour déterminer les règles de mise à jour multiplicatives. La méthode consiste à majorer la fonction de coût par une fonction appelée fonction auxiliaire dont la minimisation peut être calculée de façon analytique. Les algorithmes de type **MM** englobent notamment l'algorithme Espérance/Maximisation (**EM**).

Une fonction auxiliaire au point θ_t , notée G_{θ_t} , doit vérifier les conditions suivantes :

- $G_{\theta_t}(\theta) \geq \mathcal{C}(\theta)$
- $G_{\theta_t}(\theta_t) = \mathcal{C}(\theta_t)$

On démontre aisément que la règle de mise à jour $\theta_{t+1} = \arg \min_{\theta} G_{\theta_t}(\theta)$ fait alors décroître le critère $\mathcal{C}(\theta)$. Ce principe est illustré dans la figure 2.8.

La difficulté pratique des algorithmes **MM** réside donc dans la construction de fonctions auxiliaires faciles à minimiser.

En **NMF**, l'algorithme **MM** a permis initialement de justifier l'utilisation des règles multiplicatives pour les coûts **KL** et **EUC**, [Lee et Seung, 2000].

La décroissance de la fonction de coût peut être une propriété souhaitable car elle assure la convergence du critère. Cependant, contrairement à ce qui est souvent écrit dans la littérature, elle n'assure pas la convergence des paramètres et rien n'assure que le point de convergence du critère en soit un minimum local. Les problèmes d'unicité de la **NMF** limitent les propriétés de convergence à une notion faible de stabilité de Lyapunov [Badeau et al., 2010].

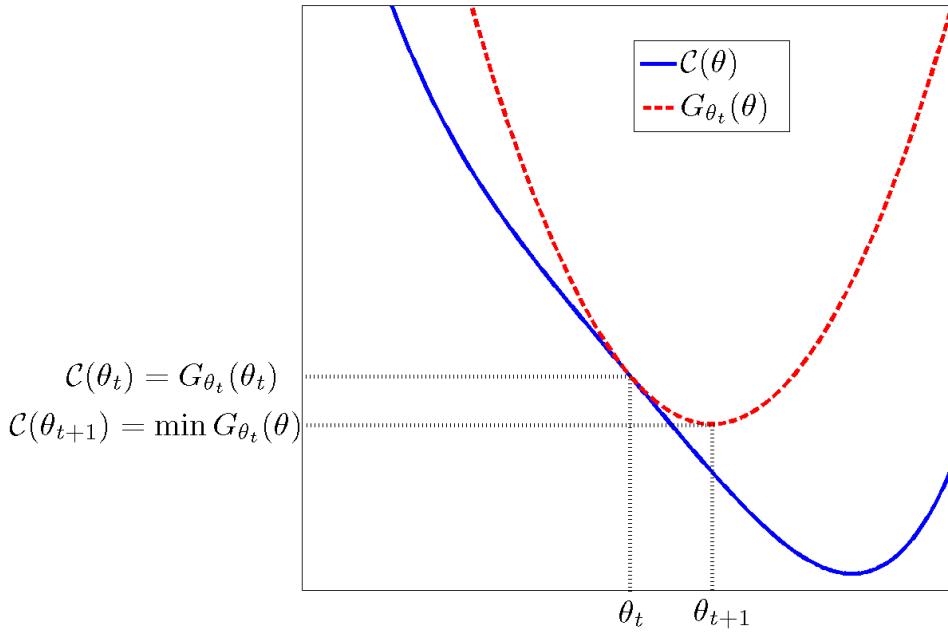


FIG. 2.8 – Règle de mise à jour dans un algorithme MM.

Il est à noter qu'il n'est absolument pas nécessaire que le critère soit décroissant pour que celui-ci converge et qu'un critère non-décroissant peut éventuellement converger beaucoup plus vite (*cf.* [Badeau *et al.*, 2010]), comme le montre la figure 2.9 : la figure représente l'évolution du critère (en l'occurrence une divergence d'**IS**) au cours des itérations d'un algorithme dont les règles de mise à jour sont les suivantes :

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\left(\frac{\mathbf{V}}{(\mathbf{WH})^{\odot 2}} \right) \mathbf{H}^T}{(\mathbf{WH})^{\odot -1} \mathbf{H}^T} \right)^{\odot \eta}, \quad (2.16)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T \left(\frac{\mathbf{V}}{(\mathbf{WH})^{\odot 2}} \right)}{\mathbf{W}^T (\mathbf{WH})^{\odot -1}} \right)^{\odot \eta}, \quad (2.17)$$

c'est-à-dire les règles de mise à jour classiques pour une divergence d'**IS** (issues des équations (2.14) et (2.15) avec $\beta = 1$) auxquelles on a ajouté un exposant η (qui peut être interprété comme un « pas »). Deux valeurs de l'exposant η ont été utilisées (les données factorisées sont synthétiques) : pour $\eta = 0.5$, le critère est décroissant (une preuve est donnée dans [Nakano *et al.*, 2010a]) mais la convergence est très lente à l'approche de la solution ; pour $\eta = 1.2$, le critère n'est pas décroissant (la première itération fait croître le critère, ce qui n'apparaît pas sur le graphe) mais le point de convergence semble être atteint en moins d'une centaine d'itérations. La rapidité de la convergence étant en pratique bien plus souhaitable que la décroissance du critère, il semble donc qu'il ne soit pas nécessaire de s'attacher à la décroissance du critère.

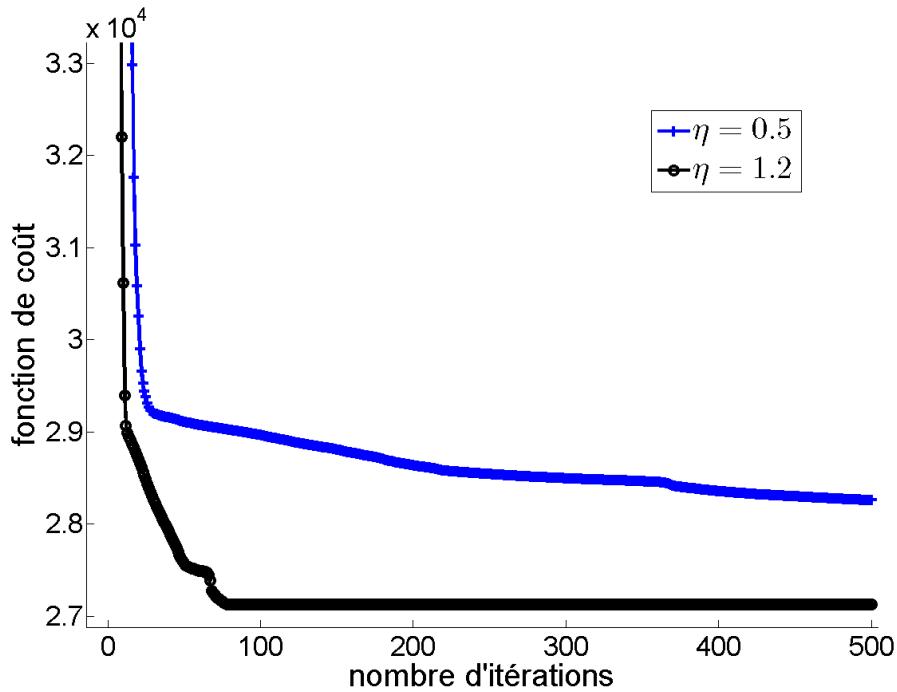


FIG. 2.9 – Evolution de la divergence d’IS pour deux exposants différents dans les règles multiplicatives : pour $\eta = 0.5$, le critère est décroissant (il existe une preuve théorique), pour $\eta = 1.2$, le critère n’est pas décroissant (ce n’est pas visible sur le graphe, mais la première itération fait croître le critère).

2.5.2.3 Algorithme EM

L’algorithme **EM** peut être utilisé dans les modélisations probabilistes de la **NMF** présentées dans la section 2.4 afin de maximiser la fonction de vraisemblance qui découle de ces modèles. Nous présentons ici l’utilisation de l’algorithme **EM** dans le cadre de la **PLCA** (*cf.* section 2.4.2) dont les règles de mise à jour peuvent facilement être mises sous forme multiplicative. Il a d’ailleurs été démontré dans [Shashanka *et al.*, 2008] que les règles de mises à jour obtenues avec l’algorithme **EM** sont équivalentes à celles proposées initialement dans [Lee et Seung, 1999].

Dans le modèle **PLCA** de base, la log-vraisemblance à minimiser est (*cf.* équation (2.10)) :

$$L((\bar{f}, \bar{t}); \theta) = \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log \left(\sum_{z=1}^Z P(z) P(f|z) P(t|z) \right),$$

où \bar{f} et \bar{t} représentent respectivement l’ensemble des tirages de f et de t .

En complétant cette log-vraisemblance avec la variable latente z , on obtient :

$$\begin{aligned} L((\bar{f}, \bar{t}, z); \theta) &= \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log(P(z)P(f|z)P(t|z)) \\ &= \sum_{f=1}^F \sum_{t=1}^T v_{ft} \{\log(P(z)) + \log(P(f|z)) + \log(P(t|z))\}. \end{aligned}$$

La variable latente z permet ainsi de séparer les paramètres $P(z)$, $P(f|z)$ et $P(t|z)$ dans l'expression de la vraisemblance. On peut alors calculer l'espérance de la vraisemblance complétée par rapport à z , connaissant f , t et le paramètre courant $\theta^{(c)}$ (ce calcul correspond à l'étape E de l'algorithme EM) :

$$\begin{aligned} Q(\theta|\theta^{(c)}) &= E_{z|f,t;\theta^{(c)}}(L((\bar{f}, \bar{t}, z); \theta)) \\ &= \sum_{z=1}^Z P(z|f,t,\theta^{(c)}) \sum_{f=1}^F \sum_{t=1}^T v_{ft} \{\log(P(z)) + \log(P(f|z)) + \log(P(t|z))\}. \end{aligned}$$

L'expression de $P(z|f,t,\theta^{(c)})$ est obtenue grâce au théorème de Bayes :

$$P(z|f,t,\theta^{(c)}) = \frac{P^{(c)}(f|z)P^{(c)}(t|z)P^{(c)}(z)}{P^{(c)}(f,t)},$$

où l'exposant $(.)^{(c)}$ fait référence à la valeur du paramètre courant.

L'étape M de l'algorithme EM consiste à maximiser $Q(\theta|\theta^{(c)})$ par rapport à θ , la règle de mise à jour du paramètre étant donnée par :

$$\theta^{(c+1)} = \arg \max_{\theta} Q(\theta|\theta^{(c)}).$$

θ étant constitué de probabilités ($P(z)$, $P(f|z)$ et $P(t|z)$), la minimisation est sujette à des contraintes de normalisation sur θ . Le lagrangien incluant ces contraintes est donné par :

$$H(\theta, \theta^{(c)}) = Q(\theta, \theta^{(c)}) + \mu(1 - \sum_z P(z)) + \sum_z \rho_z(1 - \sum_f P(f|z)) + \sum_z \tau_z(1 - \sum_t P(t|z)).$$

Remarques : Il faudrait en toute rigueur également introduire des contraintes de non-négativité sur ces probabilités. Il se trouve que ces contraintes sont en fait inactives et par conséquent, nous ne les faisons pas intervenir afin de ne pas alourdir les notations.

Les règles de mise à jour s'obtiennent alors en annulant les dérivées partielles du lagrangien. Par exemple pour $P(z)$, cela donne :

$$\frac{\partial H(\theta|\theta^{(c)})}{\partial P(z)} = \sum_{f,t} v_{ft} \frac{P(z|f,t,\theta^{(c)})}{P(z)} - \mu = 0.$$

En sommant cette égalité (multipliée par $P(z)$) sur z , on obtient :

$$\mu = \sum_{f,t,z} v_{ft} P(z|f, t, \theta^{(c)}),$$

puis la règle de mise à jour de $P(z)$:

$$P(z) \leftarrow \frac{\sum_{f,t} v_{ft} P(z|f, t, \theta^{(c)})}{\sum_{f,t,z'} v_{ft} P(z'|f, t, \theta^{(c)})}.$$

Le dénominateur est en fait juste une normalisation qui assure que $\sum_z P(z) = 1$.

On procède de manière analogue pour $P(f|z)$ et $P(t|z)$ et on obtient les règles de mise à jour :

$$P(f|z) \leftarrow \frac{\sum_t^t v_{ft} P(z|f, t, \theta^{(c)})}{\sum_{f',t} v_{f't} P(z|f', t, \theta^{(c)})},$$

$$P(t|z) \leftarrow \frac{\sum_f v_{ft} P(z|f, t, \theta^{(c)})}{\sum_{f,t'} v_{ft'} P(z|f, t', \theta^{(c)})}.$$

Ces règles de mise à jour peuvent aisément être mises sous forme matricielle multiplicative :

$$\mathbf{W} \xleftarrow{+\text{normalisation}} \mathbf{W} \odot (\mathbf{1}_{F,1}\mathbf{z}^T) \odot \left(\begin{pmatrix} \mathbf{V} \\ \hat{\mathbf{V}} \end{pmatrix} \mathbf{H}^T \right),$$

$$\mathbf{H} \xleftarrow{+\text{normalisation}} \mathbf{H} \odot (\mathbf{z}\mathbf{1}_{1,T}) \odot \left(\mathbf{W}^T \begin{pmatrix} \mathbf{V} \\ \hat{\mathbf{V}} \end{pmatrix} \right),$$

où \mathbf{W} est la matrice $F \times Z$ de coefficients $[\mathbf{W}]_{fz} = P(f|z)$, \mathbf{H} la matrice $Z \times T$ de coefficients $[\mathbf{H}]_{zt} = P(t|z)$, \mathbf{z} le vecteur de coefficients $[\mathbf{z}]_z = P(z)$ et $\hat{\mathbf{V}}$ est la matrice $F \times T$ de coefficients $[\hat{\mathbf{V}}]_{ft} = P(f, t)$. Ces mises à jour sont suivies d'une normalisation afin que les probabilités se somment à 1.

Ces règles sont alors très similaires à celles des équations (2.14) et (2.15) pour $\beta = 1$ (c'est-à-dire pour une divergence de KL).

Il est à noter que dans un algorithme EM au sens strict, les mises à jour de $P(z)$, $P(f|z)$ et $P(t|z)$ sont « simultanées » et $P(z|f, t', \theta^{(c)})$ n'est donc calculé qu'une seule fois par itération. En pratique il semble cependant que le fait de recalculer cette probabilité entre les mises à jour de $P(z)$, $P(f|z)$ et $P(t|z)$ (il s'agit alors d'un algorithme EM généralisé [Fessler et Hero, 1994]) accélère la convergence de l'algorithme.

2.5.2.4 Intérêts des algorithmes multiplicatifs

Les algorithmes à mises à jour multiplicatives sont, semble-t-il, les plus utilisés actuellement en NMF, en particulier en audio. Ces algorithmes présentent en effet un certain nombre d'avantages sur des algorithmes plus classiques (Newton, gradient projeté) :

- Les algorithmes à mises à jour multiplicatives sont faciles à mettre en oeuvre, et sont relativement rapides (beaucoup plus rapides que les méthodes d'ordre 2 et plus rapides que les algorithmes de gradient avec recherche de pas optimal).
- Il semble que pour des données à forte dynamique (ce qui est généralement le cas pour des spectrogrammes audio), les algorithmes à mises à jour multiplicatives donnent de meilleurs résultats, comme l'illustre la figure 2.10 qui présente l'évolution de la divergence de KL au cours des itérations de la NMF d'un spectrogramme musical pour trois types d'algorithmes. On constate que l'algorithme à mises à jour multiplicatives converge rapidement vers un minimum bien moindre que pour les deux autres algorithmes.

Une faiblesse de ce type d'algorithme est la lenteur de convergence à l'approche de la solution.

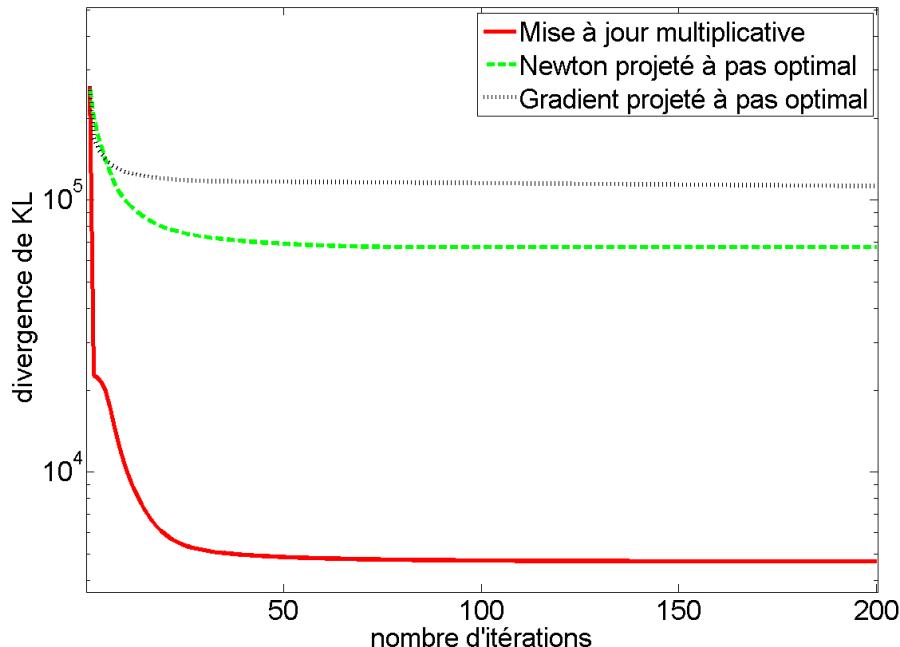


FIG. 2.10 – Evolution de la divergence de KL au cours des itérations pour trois types d'algorithmes de NMF : algorithme à mises jour multiplicatives, algorithme de Newton à pas optimal, algorithme de gradient projeté.

2.6 Variantes de la NMF et ajout de contraintes

2.6.1 Décompositions invariantes par translation

Il existe deux types de décompositions invariantes par translation : les décompositions invariantes par translation temporelle et les décompositions invariantes par translation fréquentielle.

2.6.1.1 Décomposition invariante par translation temporelle : NMFD

Dans [Smaragdis, 2004] une extension de la NMF invariante par translation temporelle est proposée. Dans cette méthode appelée Déconvolution de facteurs de matrices non-négatives (NMFD), des motifs temps/fréquence sont factorisés :

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{\tau=1}^L \mathbf{W}_\tau \overleftarrow{\mathbf{H}}^\tau, \quad (2.18)$$

où L est la longueur d'un motif temps/fréquence, \mathbf{W}_τ est un ensemble de matrices indexé par τ (qui peut être vu comme un tenseur) et $\overleftarrow{\mathbf{H}}^\tau$ correspond à la matrice d'activation \mathbf{H} dont on a translaté tous les coefficients de τ indices vers la gauche : l'équation (2.18) correspond donc à une opération de convolution. Il peut être intéressant pour l'interprétation de la décomposition de réécrire l'expression du tenseur des atomes de la façon suivante : $[\mathbf{U}_r]_{f\tau} = [\mathbf{W}_\tau]_{fr}$. Ainsi la matrice \mathbf{U}_r est l'atome bi-dimensionnel r et correspond alors à un événement musical temps/fréquence qui peut contenir des variations du contenu spectral au cours du temps. Ce type de décomposition est illustré dans la figure 2.11. Pour que cette décomposition ait du sens, il est nécessaire que les activations soient très parcimonieuses : ainsi les motifs temps/fréquence factorisés représentent bien un événement dans son ensemble. Un terme de contrainte de parcimonie sur les activations (*cf.* section 2.6.2) est donc généralement ajouté à la fonction de coût afin de favoriser cette propriété.

Cette méthode ne permet cependant pas de modéliser des variations entre les différentes occurrences d'un même événement : sa durée et l'évolution de son contenu spectral sont en effet fixes. Par conséquent, ceci limite l'utilisation de cette technique à la modélisation d'événements de durée fixe dont l'aspect est très proche d'une occurrence à l'autre (on peut penser par exemple à des éléments percussifs).

Ce type de méthode a également été utilisé directement dans le domaine temporel à l'aide d'un algorithme de type « semi-NMF » (*cf.* [Le Roux *et al.*, 2008a, Le Roux, 2009]), ce qui permet d'obtenir directement les formes d'ondes des différents éléments constitutifs du signal. Cependant, l'application de ces méthodes dans le domaine temporel s'avère très peu robuste pour la décomposition de signaux musicaux du fait de la variabilité de la forme d'onde d'une occurrence à l'autre d'un même événement sonore (par exemple plusieurs occurrences d'une même note) : de légères différences de phase entre les partiels pourtant inaudibles peuvent rendre la forme d'onde radicalement différente d'une occurrence à une autre.

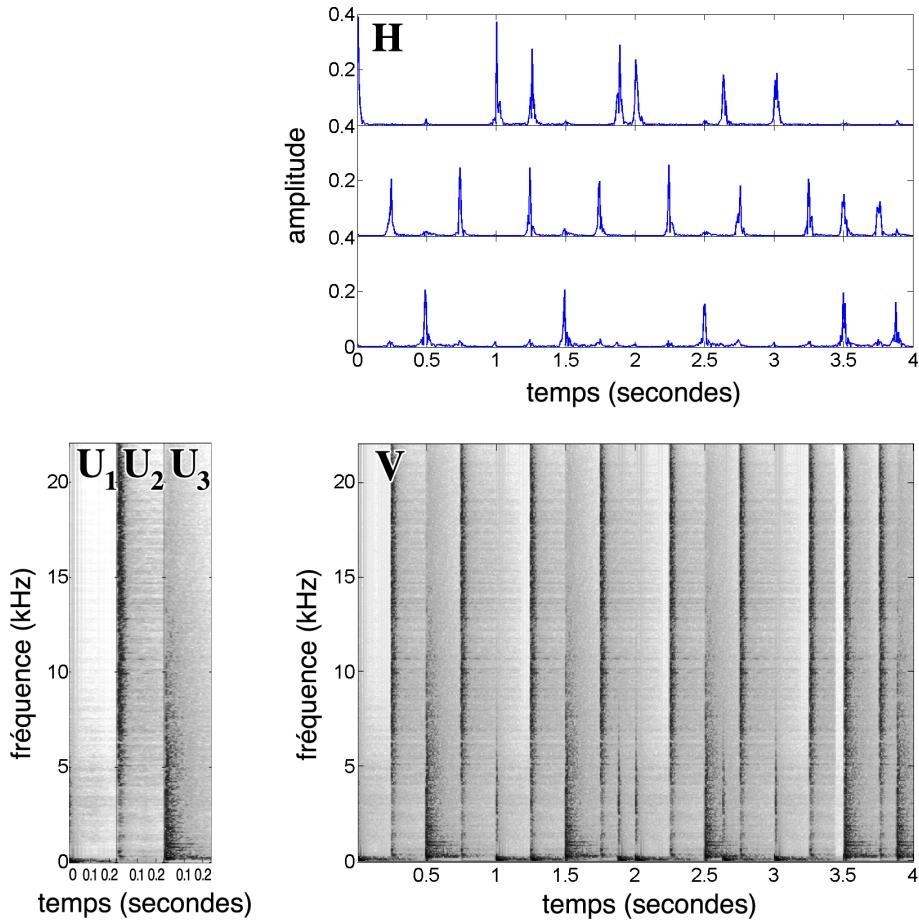


FIG. 2.11 – Décomposition invariante par translation temporelle du spectrogramme **V** (en-bas à droite) contenant des éléments percussifs (boucle de batterie). **U₁**, **U₂** et **U₃** (en-bas à gauche) sont les trois atomes temps/fréquence, correspondant chacun à un élément de batterie (respectivement, grosse caisse, charleston et caisse claire). Les activations **H** (en-haut à droite) doivent être très parcimonieuses pour que la décomposition ait de l'intérêt : ici les activations prennent effectivement une forme très impulsive au moment des attaques de l'élément de batterie correspondant.

2.6.1.2 Décomposition invariante par translation fréquentielle

Les décompositions invariantes par translation fréquentielle [Schmidt et Mørup, 2006, Smaragdis *et al.*, 2008] sont utilisées pour décomposer des spectrogrammes à Q-constant (qui ont une résolution fréquentielle logarithmique).

Si ce modèle de décomposition est très proche conceptuellement des décompositions invariantes par translation temporelle, la signification de la translation y est radicalement différente : dans un spectrogramme à Q-constant, une translation d'un motif spectral harmonique correspond en effet à une modification de la fréquence fondamentale :

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{\tau=1}^L \overset{\downarrow}{\mathbf{W}} \mathbf{H}_{\tau}, \quad (2.19)$$

où $\overset{\downarrow\tau}{\mathbf{W}}$ correspond à la matrice \mathbf{W} dont on a translaté tous les coefficients de τ indices vers le bas et \mathbf{H}_τ est un ensemble de matrices indexé par τ (qui peut être vu comme un tenseur).

La forme invariante par translation fréquentielle de la **PLCA** est présentée plus en détail dans la section 4.2.1 page 95 en introduction d'un nouveau modèle de décomposition.

2.6.2 Contraintes

La décomposition fournie par une **NMF** n'est pas toujours pertinente. Il est ainsi courant d'ajouter des contraintes par exemple sous forme d'un terme de pénalisation à la fonction de coût afin de favoriser certaines propriétés souhaitables pour les activations ou les atomes :

- Il est souvent souhaitable que les activations soient lisses notamment pour des éléments très stationnaires : si un atome est actif à un certain moment, il est fort probable qu'il soit également actif à l'instant suivant.
- Il peut être souhaitable que les activations de certains atomes soient très corrélées (lorsque ces atomes sont censés représenter un même objet) ou très décorrélées (lorsque ces atomes sont censés représenter des objets différents).
- Dans certains cas, il est souhaitable que les activations soient parcimonieuses.

Ces contraintes sont généralement introduites en ajoutant un terme de pénalité $\mathcal{C}_c(\mathbf{W}, \mathbf{H})$ à la fonction de coût. On obtient ainsi une fonction de coût du type :

$$\mathcal{C}(\mathbf{W}, \mathbf{H}) = D(\mathbf{V} || \mathbf{WH}) + \lambda \mathcal{C}_c(\mathbf{W}, \mathbf{H}).$$

Le terme de coût peut prendre les formes suivantes :

- Contrainte de parcimonie (proposée dans [Hoyer, 2004]) :

$$\mathcal{C}_c(\mathbf{H}) = \sum_{r=1}^R \frac{\sqrt{T} - \frac{\|\mathbf{h}_r\|_1}{\|\mathbf{h}_r\|_2}}{\sqrt{T} - 1}.$$

Il est à noter que dans [Hoyer, 2004], la contrainte n'est pas ajoutée de manière souple comme un terme de pénalisation mais est une contrainte forte : la valeur du terme de contrainte est en effet imposée à une valeur fixée par l'utilisateur. Il est cependant également possible d'introduire le terme de mesure de parcimonie proposé ci-dessus comme un simple terme de pénalisation dans la fonction de coût.

- Contrainte de décorrélation [Zhang et Fang, 2007] :

$$\mathcal{C}_c = \frac{1}{2} \left[\sum_{r=1}^R \log([\mathbf{HH}^T]_{rr}) - \log(\mathbf{HH}^T) \right].$$

- Contrainte de régularité temporelle [Virtanen, 2007] :

$$\mathcal{C}_c(\mathbf{H}) = \sum_{r=1}^R \frac{1}{\|\mathbf{h}_r\|_2^2} \sum_{t=2}^T (h_{r,t} - h_{r,t-1})^2.$$

Les termes de contraintes sont censés donner une certaine forme à la décomposition avec de meilleures propriétés. Plusieurs problèmes se posent lorsqu'on utilise un terme de contrainte :

- On peut opérer des changements d'échelle sur les solutions d'une **NMF** non contrainte sans changer la valeur de la fonction de coût (*cf.* section 2.2.1). Ainsi, il faut s'assurer qu'un terme de contrainte sur \mathbf{H} seulement (ou \mathbf{W} seulement) ne soit pas minimal quand \mathbf{H} (ou \mathbf{W}) tend vers 0 ou $+\infty$, afin d'éviter une convergence vers 0 d'une des matrices et vers $+\infty$ de l'autre et les problèmes numériques qui en découlent.
- Les termes de contraintes peuvent ne pas être convexes. Les contraintes de parcimonie sont par exemple systématiquement non convexes et ont leurs minima sur les bords. Ainsi, d'une part se pose la question du dosage de cette contrainte : pour un poids de cette contrainte nulle $\lambda = 0$, les solutions auront donc tendance à être plutôt « au centre de l'espace » (si elles sont déjà sur les bords, il semble inutile d'ajouter un terme pour les contraindre à y être), pour un poids infini, les solutions sont sur un axe. Ainsi l'ajout de la contrainte ne doit pas se faire au détriment de la pertinence de la décomposition. D'autre part, un terme non convexe de plus ajoute une difficulté supplémentaire dans le problème de minimisation.

Une méthode simple mais peu rigoureuse pour intégrer ces termes de pénalité dans les algorithmes à mise à jour multiplicative consiste à utiliser l'approche simple présentée dans la section 2.5.2.1.

2.7 Limitations de la **NMF**, variations temporelles

La simplicité du modèle de factorisation en matrices non-négatives entraîne un certain nombre de problèmes lorsqu'il s'agit de décomposer des spectrogrammes complexes : la **NMF** est en effet une technique de réduction de rang des données. Par conséquent, elle ne permet pas de prendre en compte de façon efficace certaines variations d'éléments non-stationnaires dans le spectrogramme : les variations d'enveloppe spectrale (section 2.7.1) et les variations de fréquence fondamentale (section 2.7.2) d'un même objet sonore sont ainsi mal modélisées par une **NMF** standard.

2.7.1 Variations d'enveloppes spectrales

Les variations importantes de forme spectrale au sein d'un même événement sonore (note de musique...) ne peuvent pas être modélisées de façon satisfaisante : de nombreux atomes sont nécessaires pour correctement représenter de tels événements et la décomposition perd alors son sens.

La **NMF** ne permet donc pas de prendre en compte l'évolution fréquentielle de chaque note et s'avère inefficace pour des sons présentant de fortes variations spectrales au cours du temps, même si la note présente une caractéristique redondante d'une trame à l'autre (par exemple une fréquence fondamentale fixe).

Cette limitation de la **NMF** est illustrée dans la figure 2.12 dans laquelle un son de guimbarde est décomposé : on constate que lorsque le nombre d'atomes est trop faible ($R = 1, 2$) la résonance n'est pas modélisée correctement. Un nombre d'atomes important ($R = 10$) est donc nécessaire pour correctement décomposer le spectrogramme : les atomes représentent alors chacun un bout de la résonance et n'ont alors plus vraiment de sens individuellement. L'utilisation de la **NMF** pour décomposer ce type de signal semble donc inadaptée.

Une solution à ce problème est proposée dans le chapitre 3.

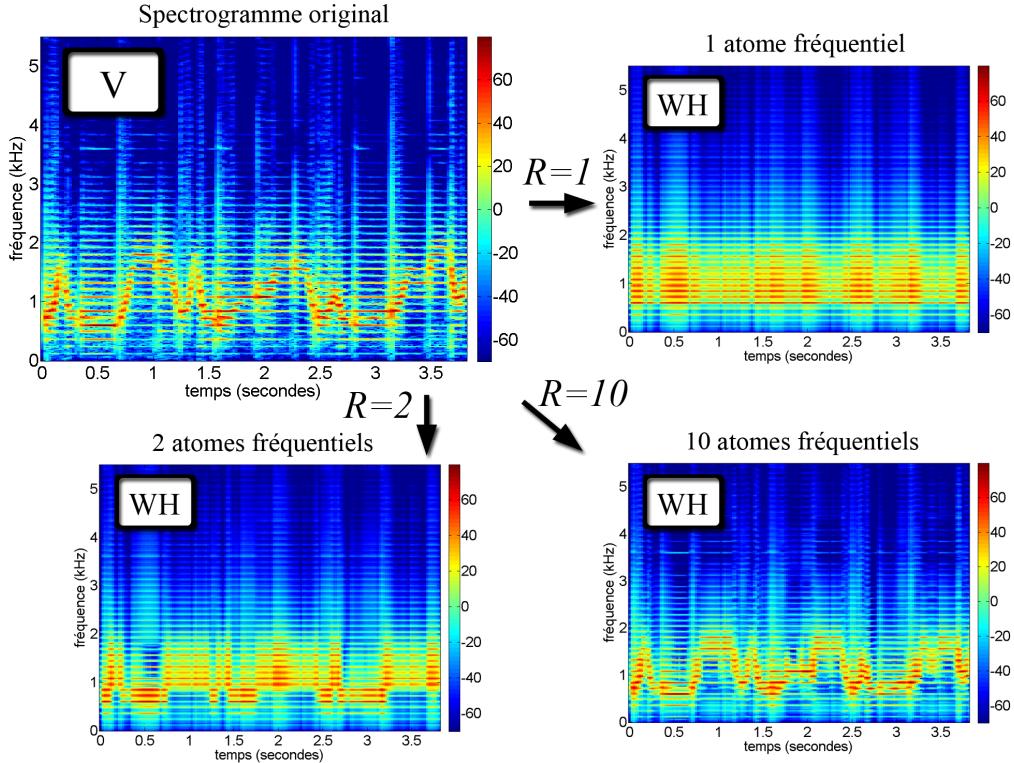


FIG. 2.12 – Décomposition d'un son de guimbarde à l'aide d'une **NMF** : en haut à gauche, spectrogramme original ; les autres spectrogrammes sont des spectrogrammes reconstruits à partir de la **NMF** de ce spectrogramme original en utilisant des nombres d'atomes R différents.

2.7.2 Variations de fréquence fondamentale

Les légères variations de fréquence fondamentale (par exemple rencontrées lors de vibrato) réduisent fortement la redondance d'une trame à l'autre et ne peuvent donc pas être prises en compte correctement par une méthode de réduction de rang comme la **NMF**.

Cette limitation de la **NMF** est illustrée dans la figure 2.13 : une note produite par un synthétiseur contenant un important vibrato est décomposée à l'aide d'une **NMF** pour $R = 1, 3$ et 10 atomes. Si la variation de fréquence fondamentale du vibrato reste faible, l'effet sur les harmoniques de haute fréquence est très important, et la redondance trame à trame est donc cassée par le vibrato. Pour $R = 1$ atome, le vibrato n'est pas du tout pris en compte par la **NMF** : seuls les premiers harmoniques sont correctement représentés. Pour $R = 3$ atomes, le vibrato commence à être pris en compte, mais la reconstruction dans les hautes fréquences reste mauvaise. Il est donc nécessaire d'avoir un nombre important d'atomes pour pouvoir prendre en compte ce phénomène : avec 10 atomes, le vibrato est très bien modélisé, mais encore une fois les atomes ont perdu leur sens individuel et possèdent une certaine redondance entre eux. La redondance dans ce type de spectrogramme est en fait trop subtile pour être correctement prise en compte par une **NMF** : seul un petit nombre de paramètres varie (en l'occurrence la fréquence fondamentale), le reste (enveloppe spectrale) est à peu près stationnaire.

Ce second problème est abordé dans le chapitre 4.

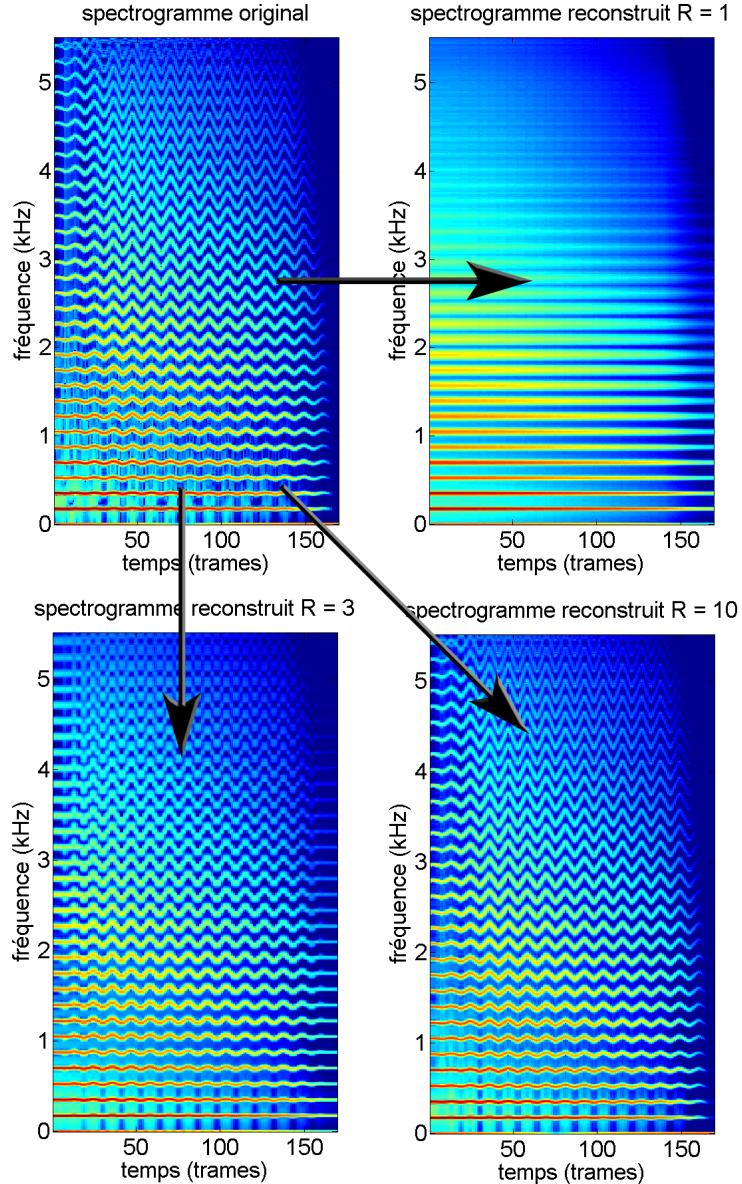


FIG. 2.13 – Décomposition du spectrogramme d'une note contenant du vibrato (spectrogramme original en haut à gauche) à l'aide d'une **NMF** : pour $R = 1$ atome (en haut à droite), $R = 3$ atomes (en bas à gauche) et $R = 10$ atomes (en bas à droite).

Chapitre 3

Modélisation des variations d'enveloppe spectrale : modèle source/filtre et NMF

Comme montré dans la section 2.7, la Factorisation en matrices non-négatives (**NMF**) classique ne s'avère efficace pour décomposer avec sens les spectrogrammes que lorsque les composants élémentaires (comme par exemple des notes de musique) du son analysé sont quasiment stationnaires, c'est-à-dire lorsque l'enveloppe spectrale de ces composants ne change pas au cours du temps. Néanmoins, il existe de nombreuses situations dans lesquelles ces composants élémentaires peuvent être fortement non-stationnaires. Dans ce chapitre seront étudiées les variations timbrales, c'est-à-dire les variations de la forme spectrale des composants au cours du temps. Ce type de variation peut par exemple être rencontré dans les instruments à cordes libres (pincées ou frappées) pour lesquels les partiels aigus disparaissent plus vite que les partiels graves ou bien dans la voix chantée (le son de différentes voyelles présente d'importantes dissimilarités spectrales).

Les variations de fréquence fondamentale qui sont rencontrées dans des phénomènes comme le vibrato ou la prosodie ne sont pas étudiées dans ce chapitre. Ces variations font l'objet du chapitre 4.

Lorsque d'importantes variations spectrales interviennent au sein d'un même élément sonore, la **NMF** classique doit utiliser plusieurs atomes individuellement dénués de sens pour décomposer un unique événement, ce qui nécessite un post-traitement (afin de regrouper les différentes parties d'un même élément [FitzGerald *et al.*, 2005]). Pour éviter ce problème, Smaragdis propose dans [Smaragdis, 2004] une extension de la **NMF** invariante par translation dans le temps (cf. section 2.6.1.1). Cette méthode ne permet malheureusement pas de modéliser des variations entre les différentes occurrences d'un même événement : sa durée et l'évolution de son contenu spectral sont en effet fixes. Par conséquent, ceci limite l'utilisation de cette technique à la modélisation d'événements de durée fixe dont l'aspect est très proche d'une occurrence à l'autre (on peut penser par exemple à des éléments percussifs).

Dans ce chapitre est présentée une extension de la **NMF** dans laquelle les activations temporelles deviennent dépendantes de la fréquence : cette approche peut être interprétée par le biais du classique paradigme source/filtre comme une factorisation source/filtre.

Notre méthode inclut des filtres Auto-Régressif(s) à Moyenne Ajustée (**ARMA**) dont les paramètres sont estimés à partir des données, associe un filtre variant dans le temps à chaque source et apprend les sources (atomes) d'une façon totalement non-supervisée. Cette méthode présente quelques similarités avec le travail de Durrieu [Durrieu *et al.*, 2009b, 2008] dans lequel un modèle source/filtre est également utilisé dans un cadre de **NMF** afin d'extraire la mélodie principale dans des morceaux de musique : ce modèle permet de représenter efficacement les importantes variations spectrales de la voix humaine. Cependant, l'approche proposée ici est assez différente puisque les sources sont apprises (dans le travail de Durrieu, elles sont fixes), un filtre est associé à chaque source (dans le travail de Durrieu, un unique filtre est utilisé car une seule source est supposée active à chaque instant), et le modèle de filtre utilisé est plus classique.

Dans la section 3.1, la décomposition source/filtre proposée est présentée comme une extension de la **NMF**. Dans la section 3.2 est présenté un algorithme itératif similaire à ceux utilisés en **NMF** pour calculer cette décomposition. Dans la section 3.3, quelques expériences de décomposition source/filtre de spectrogrammes sont présentées, et la décomposition obtenue avec l'approche proposée est comparée à la décomposition obtenue avec une **NMF** standard.

Le travail présenté dans ce chapitre a été publié dans [Hennequin *et al.*, 2011b, 2010a].

3.1 Modèle

3.1.1 Activation temps/fréquence

La **NMF** ne fournit pas une représentation efficace pour un son présentant une évolution spectrale importante : par exemple, les partiels aigus d'une note produite par un instrument à corde pincée décroissent plus rapidement que les partiels graves. Cette caractéristique ne peut pas être modélisée correctement avec un unique motif fréquentiel. Plusieurs atomes sont alors nécessaires pour décomposer une même note ce qui résulte en une décomposition moins pertinente : le sens de chaque atome est perdu et un atome ne correspond plus à un événement musical (comme une note) dans son ensemble.

Pour pallier cette limitation, nous avons proposé une extension de la **NMF** dans laquelle les activations temporelles sont remplacées par des activations temps/fréquence. Ainsi l'équation (2.1) (page 23) de base de la **NMF** est remplacée par :

$$[\mathbf{V}]_{ft} \approx [\hat{\mathbf{V}}]_{ft} = \sum_{r=1}^R [\mathbf{W}]_{fr} [\mathbf{H}(f)]_{rt}. \quad (3.1)$$

Les coefficients d'activation sont alors dépendants de la fréquence. Pour éviter que la dimension du problème soit plus importante que la dimension des données, il est nécessaire de paramétriser la dépendance de \mathbf{H} par rapport à f : nous avons choisi d'utiliser le modèle **ARMA** (section 3.1.2) pour sa généralité et sa capacité à modéliser des filtres complexes avec peu de coefficients, mais il serait tout à fait possible de considérer un autre modèle de filtres.

L'équation (3.1) peut être interprétée à l'aide du paradigme source/filtre : le spectre de chaque trame du signal est supposé correspondre à une combinaison linéaire de motifs fréquentiels (sources) filtrés. $[\mathbf{H}(f)]_{rt}$ correspond au filtre variable dans le temps associé à

la source r . La décomposition bénéficie ainsi de la polyvalence du modèle source/filtre qui est couramment utilisé pour représenter des objets sonores très divers.

3.1.2 Paramétrisation source/filtre

Nous avons choisi de paramétriser les activations temps/fréquence $h_{rt}(f)$ à l'aide du modèle ARMA. Il est à noter que ce type de modèles (plus précisément les filtres à Moyenne Ajustée (MA) qui constituent une sous-classe des filtres ARMA) a également été utilisé dans un cadre statistique afin de modéliser la forme spectrale de chaque note dans des systèmes d'estimation de fréquences fondamentales multiples [Badeau *et al.*, 2009].

$h_{rt}(f)$ a donc la forme suivante :

$$h_{rt}^{\text{ARMA}}(f) = \sigma_{rt}^2 \frac{\left| \sum_{q=0}^Q b_{rt}^q e^{-i2\pi\nu_f q} \right|^2}{\left| \sum_{p=0}^P a_{rt}^p e^{-i2\pi\nu_f p} \right|^2}, \quad (3.2)$$

où $\nu_f = \frac{f-1}{2(F-1)}$ est la fréquence normalisée associée à l'indice fréquentiel $f \in \{1, \dots, F\}$ (comme les signaux audio sont réels, seules les fréquences positives inférieures à la fréquence de Nyquist sont considérées). b_{rt}^q sont les coefficients de la partie MA du filtre et a_{rt}^p ceux de la partie Auto-Régressif(ve) (AR). σ_{rt}^2 est le gain global du filtre : afin d'éviter certains problèmes d'identifiabilité, le premier coefficient de tous les filtres est fixé à 1. Lorsque $P = Q = 0$, $h_{rt}^{\text{ARMA}}(f)$ ne dépend pas de f et la décomposition est alors une NMF standard d'activations σ_{rt}^2 .

En posant $\mathbf{a}_{rt} = (a_{rt}^0, \dots, a_{rt}^P)^T$ et $\mathbf{b}_{rt} = (b_{rt}^0, \dots, b_{rt}^Q)^T$, les activations temps/fréquence peuvent être réécrites :

$$h_{rt}^{\text{ARMA}}(f) = \sigma_{rt}^2 \frac{\mathbf{b}_{rt}^T \mathbf{T}(\nu_f) \mathbf{b}_{rt}}{\mathbf{a}_{rt}^T \mathbf{U}(\nu_f) \mathbf{a}_{rt}},$$

où $\mathbf{T}(\nu)$ est une matrice de Toeplitz de taille $(Q+1) \times (Q+1)$ avec $[\mathbf{T}(\nu)]_{pq} = \cos(2\pi\nu(p-q))$ et $\mathbf{U}(\nu)$ est construite de la même façon mais est de taille $(P+1) \times (P+1)$.

Démonstration. Pour tout $\mathbf{b}_{rt}^T \in \mathbb{R}^{Q+1}$ et pour tout $\nu_f \in \mathbb{R}$, on peut écrire :

$$\begin{aligned}
\left| \sum_{q=0}^Q b_{rt}^q e^{-i2\pi\nu_f q} \right|^2 &= \left(\sum_{q=0}^Q b_{rt}^q e^{-i2\pi\nu_f q} \right) \left(\sum_{q=0}^Q b_{rt}^q e^{i2\pi\nu_f q} \right) \\
&= \sum_{q=0}^Q \sum_{q'=0}^Q b_{rt}^q e^{-i2\pi\nu_f(q-q')} b_{rt}^{q'} \\
&= \sum_{q=0}^Q \sum_{q'=q+1}^Q b_{rt}^q (e^{-i2\pi\nu_f(q-q')} + e^{-i2\pi\nu_f(q'-q)}) b_{rt}^{q'} + \sum_{q=0}^Q (b_{rt}^q)^2 \quad (3.3) \\
&= \sum_{q=0}^Q \sum_{q'=q+1}^Q b_{rt}^q 2 \cos(2\pi\nu_f(q-q')) b_{rt}^{q'} + \sum_{q=0}^Q (b_{rt}^q)^2 \\
&= \sum_{q=0}^Q \sum_{q'=0}^Q b_{rt}^q \cos(2\pi\nu_f(q-q')) b_{rt}^{q'} \\
&= \mathbf{b}_{rt}^T \mathbf{T}(\nu_f) \mathbf{b}_{rt}.
\end{aligned}$$

La démonstration est analogue pour a_{rt} . \square

Il est possible de considérer un modèle **MA** ou bien un modèle **AR**, en prenant respectivement $P = 0$ et $Q = 0$. Il est à noter que $h_{rt}^{\text{ARMA}}(f)$ est toujours positif bien qu'il n'existe de contraintes de positivité ni sur \mathbf{b}_{rt}^q , ni sur \mathbf{a}_{rt}^p .

Le spectrogramme paramétrique donné dans l'équation (3.1) devient alors :

$$\hat{v}_{ft} = \sum_{r=1}^R w_{fr} \sigma_{rt}^2 \frac{\mathbf{b}_{rt}^T \mathbf{T}(\nu_f) \mathbf{b}_{rt}}{\mathbf{a}_{rt}^T \mathbf{U}(\nu_f) \mathbf{a}_{rt}}. \quad (3.4)$$

3.2 Algorithme

Afin de conserver une certaine généralité, nous utilisons dans ce chapitre une β -divergence comme fonction de coût :

$$\mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}) = \sum_{f=1}^F \sum_{t=1}^T d_\beta(v_{ft} | \hat{v}_{ft}),$$

où $[\boldsymbol{\Sigma}]_{rt} = \sigma_{rt}^2$, $[\mathbf{A}]_{rtp} = a_{rt}^p$ et $[\mathbf{B}]_{rtq} = b_{rt}^q$. L'expression de d_β est donnée dans l'équation (2.4) page 26.

La dérivée partielle de la fonction de coût par rapport à n'importe quelle variable θ (θ pouvant être n'importe quel coefficient de \mathbf{W} , $\boldsymbol{\Sigma}$, \mathbf{A} ou \mathbf{B}) est donnée par :

$$\frac{\partial \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma})}{\partial \theta} = \sum_{f=1}^F \sum_{t=1}^T \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}) \frac{\partial \hat{v}_{ft}}{\partial \theta}. \quad (3.5)$$

Nous déduisons de cette expression un algorithme à mise à jour multiplicative similaire à celui utilisé dans [Lee et Seung, 1999, Févotte et al., 2009, Smaragdis, 2004]. Dans un tel

algorithme itératif, la règle de mise à jour associée au paramètre θ est obtenue en écrivant la dérivée partielle de la fonction de coût par rapport à θ comme une différence de deux termes strictement positifs, comme décrit dans la section 2.5.2.1 page 44 : $\frac{\partial \mathcal{C}}{\partial \theta} = G_\theta - F_\theta$. La règle de mise à jour de θ est alors :

$$\theta \leftarrow \theta \times \frac{F_\theta}{G_\theta}. \quad (3.6)$$

Dans ce chapitre, nous considérons également une mise à jour simultanée de tout un vecteur pour les filtres : l'expression du gradient de \mathcal{C} par rapport à un vecteur $\boldsymbol{\theta}$ de coefficients de \mathbf{A} ou \mathbf{B} est alors similaire à l'expression (3.5), en remplaçant la dérivée partielle par un gradient $\nabla_{\boldsymbol{\theta}}$:

$$\nabla_{\boldsymbol{\theta}} \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}) = \sum_{f=1}^F \sum_{t=1}^T \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}) \nabla_{\boldsymbol{\theta}} \hat{v}_{ft}. \quad (3.7)$$

3.2.1 Mise à jour des atomes

La règle de mise à jour de w_{fr} est déduite de l'expression de la dérivée partielle de la fonction de coût par rapport à w_{fr} . La dérivée partielle du spectrogramme paramétrique $\hat{\mathbf{V}}$ (définie dans l'équation (3.4)) par rapport à $w_{f_0 r_0}$ est donnée par :

$$\frac{\partial \hat{v}_{ft}}{\partial w_{f_0 r_0}} = h_{r_0 t}^{\text{ARMA}}(f_0) \delta_{f_0 f}.$$

En remplaçant cette expression dans l'équation (3.5) avec $\theta = w_{f_0 r_0}$, on obtient la dérivée partielle de la fonction de coût par rapport à $w_{f_0 r_0}$:

$$\frac{\partial \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma})}{\partial w_{f_0 r_0}} = \sum_{t=1}^T h_{r_0 t}^{\text{ARMA}}(f_0) \hat{v}_{f_0 t}^{\beta-2} (\hat{v}_{f_0 t} - v_{f_0 t}).$$

Cette dérivée s'écrit donc comme la différence de deux termes strictement positifs :

$$G_{w_{f_0 r_0}} = \sum_{t=1}^T h_{r_0 t}^{\text{ARMA}}(f_0) \hat{v}_{f_0 t}^{\beta-1} \quad \text{et} \quad F_{w_{f_0 r_0}} = \sum_{t=1}^T h_{r_0 t}^{\text{ARMA}}(f_0) \hat{v}_{f_0 t}^{\beta-2} v_{f_0 t}.$$

On en déduit la règle de mise à jour de $w_{f_0 r_0}$:

$$w_{f_0 r_0} \leftarrow w_{f_0 r_0} \frac{F_{w_{f_0 r_0}}}{G_{w_{f_0 r_0}}}. \quad (3.8)$$

3.2.2 Mise à jour des activations globales

Les règles de mise à jour de σ_{rt}^2 sont calculées de façon analogue à partir de l'expression de la dérivée partielle de la fonction de coût \mathcal{C} par rapport à σ_{rt}^2 .

La dérivée partielle du spectrogramme paramétrique $\hat{\mathbf{V}}$ (défini dans l'équation (3.4)) par rapport à $\sigma_{r_0 t_0}^2$ est donnée par :

$$\frac{\partial \hat{v}_{ft}}{\partial \sigma_{r_0 t_0}^2} = w_{fr_0} \frac{\mathbf{b}_{r_0 t_0}^T \mathbf{T}(\nu_f) \mathbf{b}_{r_0 t_0}}{\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0}} \delta_{t_0 t}.$$

En remplaçant cette expression dans l'équation (3.5) avec $\theta = \sigma_{r_0 t_0}^2$, on obtient la dérivée partielle de la fonction de coût par rapport à $\sigma_{r_0 t_0}^2$:

$$\frac{\partial \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma})}{\partial \sigma_{r_0 t_0}^2} = \sum_{f=1}^F \frac{w_{fr_0}}{\sigma_{r_0 t_0}^2} h_{r_0 t_0}^{\text{ARMA}}(f) \hat{v}_{ft_0}^{\beta-2} (\hat{v}_{ft_0} - v_{ft_0}).$$

Cette dérivée s'écrit donc comme la différence de deux termes strictement positifs :

$$G_{\sigma_{r_0 t_0}^2} = \sum_{f=1}^F \frac{w_{fr_0}}{\sigma_{r_0 t_0}^2} h_{r_0 t_0}^{\text{ARMA}}(f) \hat{v}_{ft_0}^{\beta-1} \quad \text{et} \quad F_{\sigma_{r_0 t_0}^2} = \sum_{f=1}^F \frac{w_{fr_0}}{\sigma_{r_0 t_0}^2} h_{r_0 t_0}^{\text{ARMA}}(f) \hat{v}_{ft_0}^{\beta-2} v_{ft_0}.$$

On en déduit la règle de mise à jour de $\sigma_{r_0 t_0}^2$:

$$\sigma_{r_0 t_0}^2 \leftarrow \sigma_{r_0 t_0}^2 \frac{F_{\sigma_{r_0 t_0}^2}}{G_{\sigma_{r_0 t_0}^2}}. \quad (3.9)$$

On peut remarquer que lorsque $Q = 0$ et $P = 0$ (c'est-à-dire en l'absence de filtre), les règles de mise à jour de w_{fr} et σ_{rt}^2 sont les mêmes que celles données dans [Févotte et al., 2009] qui sont les règles d'une NMF standard en prenant une β -divergence comme fonction de coût ($\mathbf{W} = (w_{fr})_{fr}$ correspond à la matrice d'atomes et $\boldsymbol{\Sigma} = (\sigma_{rt}^2)_{fr}$ à la matrice d'activations).

3.2.3 Mise à jour des filtres

Les règles de mise à jour des coefficients des filtres sont calculées de façon assez similaire, mais ne se font plus élément par élément mais vecteur par vecteur : une règle de mise à jour est donc calculée pour chaque vecteur \mathbf{b}_{rt} et chaque vecteur \mathbf{a}_{rt} .

Mise à jour de \mathbf{b}_{rt} : L'expression du gradient du spectrogramme paramétrique \hat{v}_{ft} par rapport à $\mathbf{b}_{r_0 t_0}$ est donnée par :

$$\nabla_{\mathbf{b}_{r_0 t_0}} \hat{v}_{ft} = \delta_{t_0 t} \frac{2w_{fr_0} \sigma_{r_0 t_0}^2}{\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0}} \mathbf{T}(\nu_f) \mathbf{b}_{r_0 t_0}.$$

En remplaçant cette expression dans l'équation (3.7) avec $\boldsymbol{\theta} = \mathbf{b}_{r_0 t_0}$, on obtient le gradient de la fonction de coût par rapport à $\mathbf{b}_{r_0 t_0}$:

$$\nabla_{\mathbf{b}_{r_0 t_0}} \mathcal{C} = 2 \sum_{f=1}^F \frac{w_{fr_0} \sigma_{r_0 t_0}^2 \hat{v}_{ft_0}^{\beta-2} (\hat{v}_{ft_0} - v_{ft_0})}{\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0}} \mathbf{T}(\nu_f) \mathbf{b}_{r_0 t_0} = 2 \sigma_{r_0 t_0}^2 (\mathbf{R}_{r_0 t_0} - \mathbf{R}'_{r_0 t_0}) \mathbf{b}_{r_0 t_0},$$

où :

$$\mathbf{R}_{r_0 t_0} = \sum_{f=1}^F \frac{w_{fr_0} \hat{v}_{ft_0}^{\beta-1}}{\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0}} \mathbf{T}(\nu_f) \quad \text{et} \quad \mathbf{R}'_{r_0 t_0} = \sum_{f=1}^F \frac{w_{fr_0} \hat{v}_{ft_0}^{\beta-2} V_{ft_0}}{\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0}} \mathbf{T}(\nu_f).$$

Les matrices $\mathbf{R}_{r_0 t_0}$ et $\mathbf{R}'_{r_0 t_0}$ sont toutes deux définies positives sous de faibles hypothèses comme le montre le lemme suivant :

Lemme 3.2.1. *S'il existe au moins $Q + 1$ indices distincts f tels que $w_{fr_0} \hat{v}_{ft_0}^{\beta} \neq 0$ alors $\mathbf{R}_{r_0t_0}$ est définie positive.*

Démonstration. Comme le montre l'équation (3.3), on a pour tout $\mathbf{u} \in \mathbb{R}^{Q+1}$ et tout $\nu_f \in \mathbb{R}$: $\mathbf{u}^T \mathbf{T}(\nu_f) \mathbf{u} \geq 0$, donc $\mathbf{T}(\nu_f)$ est une matrice semi-définie positive. $\mathbf{R}_{r_0t_0}$ est alors clairement une matrice semi-définie positive puisque c'est une combinaison linéaire positive des matrices semi-définies positives $\mathbf{T}(\nu_f)$.

Il faut donc démontrer que $\mathbf{R}_{r_0t_0}$ est inversible. Nous allons démontrer que s'il existe plus de $Q + 1$ indices f différents tels que $w_{fr_0} \hat{v}_{ft_0}^{\beta-1} \neq 0$ alors $\mathbf{R}_{r_0t_0}$ est inversible.

Soit \mathbf{u} un vecteur du noyau de $\mathbf{R}_{r_0t_0}$ (considéré comme une matrice complexe). Alors :

$$\mathbf{u}^H \mathbf{R}_{r_0t_0} \mathbf{u} = \sum_{f=1}^F \frac{w_{fr_0} \hat{v}_{ft_0}^{\beta-1}}{\mathbf{a}_{r_0t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0t_0}} \mathbf{u}^H \mathbf{T}(\nu_f) \mathbf{u} = 0.$$

Comme tous les termes de cette somme sont positifs, ils sont tous égaux à 0. Par conséquent pour tout f tel que $w_{fr_0} \hat{v}_{ft_0}^{\beta-1} \neq 0$, $\mathbf{u}^H \mathbf{T}(\nu_f) \mathbf{u} = 0$ et \mathbf{u} est dans le noyau de $\mathbf{T}(\nu_f)$.

Comme l'image de $\mathbf{T}(\nu_f)$ est $\text{Vect}\{(1, e^{i2\pi\nu_f}, \dots, e^{i2\pi Q\nu_f})^T, (1, e^{-i2\pi\nu_f}, \dots, e^{-i2\pi Q\nu_f})^T\}$ (pour $\nu_f \neq 0$), \mathbf{u} est orthogonal à tous les vecteurs $(1, e^{i2\pi\nu_f}, \dots, e^{i2\pi Q\nu_f})^T$ où l'indice f est tel que $w_{fr_0} \hat{v}_{ft_0}^{\beta-1} \neq 0$.

S'il y a plus de $Q + 1$ tels indices f , alors il existe un ensemble $F = \{f_0, f_1, \dots, f_Q\}$ (où tous les f_k sont distincts) tel que pour tout $k \in \{0 \dots Q\}$, $w_{f_k r_0} \hat{v}_{f_k t_0}^{\beta-1} \neq 0$. Comme tous les f_k sont distincts et que pour tout $k \in \{1 \dots Q\}$ $\nu_{f_k} \in [0, 0.5]$, tous les $e^{i2\pi\nu_{f_k}}$ sont distincts. Par conséquent la famille de vecteurs $\mathcal{F} = \{(1, e^{i2\pi\nu_f}, \dots, e^{i2\pi Q\nu_f})^T\}_{f \in F}$ est une famille de Vandermonde et par conséquent est linéairement indépendante et constitue donc une base de \mathbb{C}^{Q+1} . Comme \mathbf{u} est orthogonal à tous les éléments de la famille \mathcal{F} qui est une base, \mathbf{u} est nul. Par conséquent, le noyau de $\mathbf{R}_{r_0t_0}$ est restreint au vecteur nul et $\mathbf{R}_{r_0t_0}$ est inversible et donc est définie positive.

Cette preuve peut être adaptée aux matrices \mathbf{R}'_{rt} , \mathbf{S}_{rt} et \mathbf{S}'_{rt} de façon évidente. \square

Cette hypothèse est en pratique toujours vérifiée dès que la trame d'indice t_0 n'est pas identiquement nulle : dans ce cas particulier, la décomposition est triviale, les gains des filtres $\sigma_{r_0t_0}$ peuvent être simplement fixés à 0, l'estimation des filtres n'ayant alors pas de sens. Pour $R'_{r_0t_0}$, l'hypothèse est très similaire.

Nous utilisons alors l'approche proposée dans [Badeau et David, 2008] afin de calculer la règle de mise à jour de la partie MA du filtre :

$$\mathbf{b}_{r_0t_0} \leftarrow \mathbf{R}_{r_0t_0}^{-1} \mathbf{R}'_{r_0t_0} \mathbf{b}_{r_0t_0}. \quad (3.10)$$

Comme $\mathbf{R}_{r_0t_0}$ et $\mathbf{R}'_{r_0t_0}$ sont toutes deux inversibles, $\mathbf{R}_{r_0t_0}^{-1}$ est bien définie et il est garanti que $\mathbf{b}_{r_0t_0}$ ne s'annule pas.

Mise à jour de \mathbf{a}_{rt} :

La règle de mise à jour de \mathbf{a}_{rt} est calculée de manière analogue à celle de \mathbf{b}_{rt} . Le gradient du spectrogramme paramétrique \hat{v}_{ft} par rapport à $\mathbf{a}_{r_0t_0}$ est donné par :

$$\nabla_{\mathbf{a}_{r_0t_0}} \hat{v}_{ft} = -2\delta_{tt_0} w_{fr_0} \frac{h_{r_0t_0}^{\text{ARMA}}(f)}{\mathbf{a}_{r_0t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0t_0}} \mathbf{U}(\nu_f) \mathbf{a}_{r_0t_0}.$$

En remplaçant $\boldsymbol{\theta}$ par $\mathbf{a}_{r_0 t_0}$ dans l'équation (3.7), on obtient alors l'expression du gradient de la fonction de coût par rapport à $\mathbf{a}_{r_0 t_0}$:

$$\nabla_{\mathbf{a}_{r_0 t_0}} \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}) = 2\sigma_{r_0 t_0}^2 (\mathbf{S}'_{r_0 t_0} - \mathbf{S}_{r_0 t_0}) \mathbf{a}_{r_0 t_0},$$

où :

$$\begin{aligned} \mathbf{S}_{r_0 t_0} &= \sum_{f=1}^F w_{f r_0} \hat{v}_{f t_0}^{\beta-1} \frac{\mathbf{b}_{r_0 t_0}^T \mathbf{T}(\nu_f) \mathbf{b}_{r_0 t_0}}{(\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0})^2} \mathbf{U}(\nu_f) \\ \text{et } \mathbf{S}'_{r_0 t_0} &= \sum_{f=1}^F w_{f r_0} \hat{v}_{f t_0}^{\beta-2} V_{f t_0} \frac{\mathbf{b}_{r_0 t_0}^T \mathbf{T}(\nu_f) \mathbf{b}_{r_0 t_0}}{(\mathbf{a}_{r_0 t_0}^T \mathbf{U}(\nu_f) \mathbf{a}_{r_0 t_0})^2} \mathbf{U}(\nu_f). \end{aligned}$$

Tout comme les matrices $\mathbf{R}_{r_0 t_0}$ et $\mathbf{R}'_{r_0 t_0}$, les matrices $\mathbf{S}_{r_0 t_0}$ et $\mathbf{S}'_{r_0 t_0}$ sont définies positives sous de faibles hypothèses.

On en déduit la règle de mise à jour pour la partie AR des filtres :

$$\mathbf{a}_{r_0 t_0} \leftarrow \mathbf{S}'_{r_0 t_0}^{-1} \mathbf{S}_{r_0 t_0} \mathbf{a}_{r_0 t_0}. \quad (3.11)$$

3.2.4 Description globale et implémentation pratique

Les règles de mise à jour (3.8), (3.9), (3.10) et (3.11) sont appliquées successivement à tous les coefficients de \mathbf{W} , tous les coefficients de $\boldsymbol{\Sigma}$, tous les coefficients de \mathbf{B} puis tous les coefficients de \mathbf{A} . Entre les mises à jours de chacune de ces matrices (et tenseurs), le spectrogramme paramétrique $\hat{\mathbf{V}}$ est recalculé : comme dans un algorithme de NMF standard, ce calcul entre chaque mise à jour est nécessaire afin de garantir la convergence.

Identifiabilité : Comme pour une NMF standard, la décomposition (3.4) qui minimise la fonction de coût n'est pas unique. Afin d'éviter certains problèmes d'identifiabilité, nous imposons des contraintes sur \mathbf{W} , $\boldsymbol{\Sigma}$, \mathbf{B} et \mathbf{A} :

- Pour tout r et t , on impose que \mathbf{b}_{rt} et \mathbf{a}_{rt} (considérés comme des polynômes) aient toutes leurs racines à l'intérieur du cercle unité.
- Pour tout r , nous imposons que $\|\mathbf{w}_r\| = 1$ pour une certaine norme $\|\cdot\|$.
- Pour tout r et t , nous imposons $b_{rt}^0 = 1$ et $a_{rt}^0 = 1$.

Ainsi, à la fin de chaque itération de notre algorithme, $\mathbf{b}_{r,t}$ et $\mathbf{a}_{r,t}$ sont transformés en remplaçant les racines hors du cercle unité par le conjugué de leur inverse (le gain est alors recalculé afin que $\hat{\mathbf{V}}$ ne soit pas modifié par cette transformation), chaque colonne de \mathbf{W} est normalisée, $\mathbf{b}_{r,t}$ et $\mathbf{a}_{r,t}$ sont divisés par leur premier coefficient et $\boldsymbol{\Sigma}$ est modifiée afin que \hat{v}_{ft} ne soit pas affecté par ces modifications : toutes ces transformations n'ont alors pas d'influence sur la valeur du spectrogramme paramétrique $\hat{\mathbf{V}}$.

Plusieurs choix de normalisation de filtres ont été testés : plutôt que d'imposer $b_{rt}^0 = 1$ et $a_{rt}^0 = 1$, on peut aussi imposer pour tout (r, t) que $\sum_f \mathbf{b}_{rt}^T \mathbf{T}(\nu_f) \mathbf{b}_{rt} = 1$ et $\sum_f \mathbf{a}_{rt}^T \mathbf{U}(\nu_f) \mathbf{a}_{rt} = 1$. Cette opération correspond à une normalisation de puissance qui fait donc plus sens. En pratique ces deux types de normalisation donnent des résultats très similaires.

Ces opérations sont récapitulées dans l’Algorithme 3.1. Dans le reste du chapitre, nous appellerons notre algorithme *factorisation source/filtre*.

Algorithme 3.1 : Factorisation source/filtre d’un spectrogramme

Données : \mathbf{V} , \mathbf{R} , \mathbf{Q} , \mathbf{P} , n_{iter} , β

Résultat : \mathbf{W} , Σ , \mathbf{B} , \mathbf{A}

Initialiser \mathbf{W} , Σ avec des valeurs positives

Initialiser \mathbf{B} , \mathbf{A} avec des filtres plats

pour $j = 1$ à n_{iter} **faire**

 Calculer $\hat{\mathbf{V}}$

pour chaque f et r **faire**

 Calculer $F_{w_{fr}}$ et $G_{w_{fr}}$

$$w_{fr} \leftarrow w_{fr} \frac{F_{w_{fr}}}{G_{w_{fr}}}$$

fin

 Calculer $\hat{\mathbf{V}}$

pour chaque r et t **faire**

 Calculer $F_{\sigma_{rt}^2}$ et $G_{\sigma_{rt}^2}$

$$\sigma_{rt}^2 \leftarrow \sigma_{rt}^2 \frac{F_{\sigma_{rt}^2}}{G_{\sigma_{rt}^2}}$$

fin

 Calculer $\hat{\mathbf{V}}$

pour chaque r et t **faire**

 Calculer \mathbf{R}_{rt} , \mathbf{R}'_{rt} , \mathbf{S}_{rt} et \mathbf{S}'_{rt}

$$\mathbf{b}_{rt} \leftarrow \mathbf{R}_{rt}^{-1} \mathbf{R}'_{rt} \mathbf{b}_{rt}$$

$$\mathbf{a}_{rt} \leftarrow \mathbf{S}'_{rt}^{-1} \mathbf{S}_{rt} \mathbf{a}_{rt}$$

fin

 Ramener les racines de tous les filtres à l’intérieur du cercle unité

 Diviser tous les coefficients de tous les filtres par le premier

 Normaliser \mathbf{W}

 Mettre à jour Σ en conséquence

fin

3.2.5 Dimension de l’espace des paramètres

Contrairement à la **NMF**, notre méthode n’est pas une technique de réduction de rang, ce qui permet de prendre en compte les variations temporelles d’un objet sonore. En revanche, la décomposition proposée réduit bien la dimension des paramètres et fournit donc une représentation compacte des données. Il est donc nécessaire de s’intéresser à la dimension des paramètres de notre décomposition.

La dimension des données originales est FT . Pour une **NMF** standard avec R atomes, la dimension totale des paramètres (somme des dimensions des matrices \mathbf{W} et \mathbf{H}) est $\dim \mathbf{W} + \dim \mathbf{H} = R(F+T)$. Avec notre décomposition, la dimension totale des paramètres est : $\dim \mathbf{W} + \dim \Sigma + \dim \mathbf{A} + \dim \mathbf{B} = RF + RT(P+Q+1)$. Par conséquent, on doit avoir $RF + RT(P+Q+1) \ll FT$, c’est-à-dire $R \ll T$ et $R(P+Q+1) \ll F$, ainsi P et Q doivent rester faibles. Comme en pratique $F \leq T$, la condition à respecter est la suivante :

$$R(P + Q + 1) \ll F.$$

Il faut remarquer que notre décomposition est censée réduire significativement le nombre d'atomes R nécessaire pour représenter correctement les données lorsque les différentes parties du son présentent d'importantes variations spectrales. Ainsi, la dimension totale des paramètres obtenus avec notre décomposition reste comparable à celle obtenue avec une **NMF** standard comme montré dans la section 3.3.

De plus, on peut remarquer qu'un grand nombre des coefficients des filtres sont inutiles et il n'est pas nécessaire de conserver ces valeurs : lorsque le gain global σ_{rt} d'un filtre devient très faible, les coefficients (\mathbf{b}_{rt} et \mathbf{a}_{rt}) n'ont plus de sens et sont donc inutiles : ils peuvent ainsi être supprimés sans affecter la valeur du spectrogramme paramétrique $\hat{\mathbf{V}}$.

On peut également noter que, dans la décomposition (3.4), tous les atomes sont associés à un filtre du même ordre : il serait cependant tout à fait possible de considérer un modèle plus large dans lequel les filtres n'ont pas les mêmes caractéristiques pour tous les atomes.

3.2.6 Complexité algorithmique

La complexité algorithmique de chaque itération de la factorisation source/filtre dépend de P , Q , R , F et T :

- Calcul de $\hat{\mathbf{V}}$: $\mathcal{O}((P + Q) RFT)$ opérations.
- Mise à jour de \mathbf{W} et Σ : $\mathcal{O}(RFT)$ opérations chacune.
- Mise à jour de \mathbf{B} : $\mathcal{O}(RT(FP + P^3))$ opérations.
- Mise à jour de \mathbf{A} : $\mathcal{O}(RT(FQ + Q^3))$ opérations.
- Normalisation/stabilisation : $\mathcal{O}(RT(F + P^3 + Q^3))$ opérations.

L'ordre des filtres devant rester très faible (P et Q sont typiquement de l'ordre de quelques unités), on a $P^2 < F$ et $Q^2 < F$. La complexité totale d'une seule itération de l'algorithme est donc de $\mathcal{O}((P + Q) RFT)$ opérations. Pour comparaison, la complexité d'une itération d'un algorithme multiplicatif de **NMF** est quant à elle de $\mathcal{O}(RFT)$ opérations.

Avec notre implémentation actuelle sous Matlab® , 100 itérations de notre algorithme appliqué à un spectrogramme de taille 1025×550 (correspondant à 6.5s de signal échantillonné à $f_s = 22050\text{Hz}$ avec des fenêtres de 2048 points et 75% de recouvrement) avec $R = 10$ atomes, $P = 2$ et $Q = 2$, durent à peu près 300s (sur un Intel®Core™2 Duo E8400 @3.00GHz). En comparaison, 100 itérations d'une **NMF** standard avec le même spectrogramme, le même nombre d'atomes ($R = 10$), mais $P = 0$ et $Q = 0$, sur le même ordinateur durent environ 9s : notre algorithme semble donc plus lent. En effet, l'inversion des matrices R_{rt} et S'_{rt} et le calcul de la réponse en fréquence des filtres sont particulièrement coûteux. Cependant, cette comparaison est désavantageuse pour notre algorithme, qui est censé utiliser moins d'atomes qu'une **NMF** standard : dans le cas présenté, notre algorithme estime en effet beaucoup plus de paramètres qu'une **NMF**. Si on compare le temps d'exécution avec le même nombre de paramètres, la différence est alors moindre : pour le même spectrogramme, 100 itérations de notre algorithme avec $R = 2$ atomes, $P = 2$ et $Q = 2$ (c'est-à-dire avec la même dimension totale des paramètres que pour une **NMF** avec $R = 10$ atomes), durent environ 60s ce qui est déjà beaucoup plus proche d'une **NMF**.

3.2.7 Implémentation et choix de β

Nous avons observé empiriquement la décroissance monotone de la fonction de coût et la convergence de l'algorithme pour $0 \leq \beta \leq 2$ sur un grand ensemble de tests : cette décroissance et cette convergence sont illustrées pour des cas particuliers dans la section 3.3.5.

Cependant, l'algorithme montre des comportements instables pour $1 \leq \beta \leq 2$: des instabilités numériques apparaissent lorsque les pôles de certains filtres se rapprochent du cercle unité. Ces instabilités deviennent fréquentes lorsque β se rapproche de 2 (ce qui correspond à une fonction de coût Euclidien(ne) (EUC)) ; cependant, d'après [Févotte et al., 2009], la dynamique importante des spectrogrammes audio est mieux représentée lorsque β se rapproche de 0. Afin d'éviter ces instabilités, nous contraignons le module des pôles à ne pas se rapprocher trop de 1 : la décroissance monotone de la fonction de coût n'est alors plus observée mais cela permet d'éviter des comportements non désirables de la décomposition (par exemple, éviter qu'un filtre d'ordre 2 ne devienne très résonnant en cherchant à s'adapter à un unique partielle).

Dans les exemples des sections suivantes, nous avons choisi $\beta = 0.5$ car pour cette valeur de β , les problèmes d'instabilité numérique nous ont semblé moindres et les résultats semblaient plus précis qu'avec $\beta = 0$ (distance d'Itakura-Saito (IS)).

3.3 Exemples

Dans cette section, plusieurs expériences sont présentées pour montrer que notre algorithme est bien capable de décomposer efficacement des sons ayant d'importantes variations spectrales. Tous les spectrogrammes utilisés dans cette section sont des spectrogrammes de puissance obtenus à partir de signaux enregistrés grâce à une Transformée de Fourier à Court Terme (TFCT).

Les algorithmes (NMF standard et factorisation source/filtre) ont été initialisés avec des valeurs aléatoires (à part pour les filtres, initialement plats) et ont été exécutés jusqu'à convergence apparente. Les algorithmes ont été relancés à plusieurs reprises (100 initialisations différentes) afin de maximiser les chances d'atteindre un « bon » minimum local de la fonction de coût. Les solutions atteintes étaient cependant très similaires d'une initialisation à l'autre en termes d'aspect global du spectrogramme reconstruit.

3.3.1 Guimbarde

3.3.1.1 Description du signal décomposé

Dans cette section, notre algorithme est utilisé pour décomposer un court extrait de guimbarde. La guimbarde est un petit instrument de musique constitué d'une languette métallique qui produit un son harmonique modulé par la bouche du musicien : le son produit est donc harmonique (avec une fréquence fondamentale fixe) et présente une forte résonance qui peut varier au cours du temps. On peut très clairement séparer ce son en une partie source, la languette métallique, dont le comportement est très redondant au cours du temps et une partie filtre, la résonance créée par la bouche et le conduit vocal, qui varie fortement au cours du temps. La partie redondante peut-être factorisée et la partie variable peut être modélisée par une activation temps/fréquence.

En revanche, les fortes variations spectrales ne peuvent être correctement représentées par un unique atome dans une **NMF** standard.

La figure 3.1(a) représente le spectrogramme de l'extrait : le son produit est bien harmonique et la résonance variante dans le temps apparaît clairement dans le spectrogramme. On peut donc considérer que ce signal est composé d'un unique événement comprenant de fortes variations spectrales, et essayer de le décomposer avec un unique atome ($R = 1$). La fréquence d'échantillonnage de l'extrait est $f_s = 11025\text{Hz}$. Nous avons choisi une fenêtre de Hann de 1024 échantillons (93ms) avec un recouvrement de 75% pour le calcul de la **TFCT**.

3.3.1.2 Expérience et résultat

Le spectrogramme de l'extrait est décomposé d'une part avec un algorithme de **NMF** standard pour $R = 1$ atome et $R = 10$ atomes, d'autre part avec la factorisation source/filtre pour $R = 1$ atome, avec un filtre **AR** d'ordre 2 ($Q = 0$ et $P = 2$). Les spectrogrammes reconstruits sont respectivement représentés dans les figures 3.1(b), 3.1(c) et 3.1(d).

Bien que la guimbarde soit jouée seule dans le spectrogramme analysé, la **NMF** standard a besoin de plusieurs atomes pour décomposer correctement le spectrogramme de puissance. Avec 1 atome, la **NMF** n'est pas capable de représenter correctement la résonance variable (figure 3.1(b)). Avec 10 atomes, la résonance apparaît correctement, mais les atomes ne font pas sens : chaque atome représente une partie du son et est difficilement interprétable individuellement de façon perceptive.

La factorisation source/filtre permet quant à elle de représenter efficacement les variations spectrales du son (figure 3.1(d)) avec un unique atome : la résonance variable du son original est bien modélisée.

Le motif fréquentiel obtenu est représenté dans la figure 3.2(a) : ce motif est bien harmonique. L'activation temps/fréquence de cet atome est représentée dans la figure 3.2(b) : la trajectoire de la résonance apparaît très clairement. Par conséquent, la décomposition fournie par notre algorithme semble donner une représentation plus significative du spectrogramme considéré que la **NMF**.

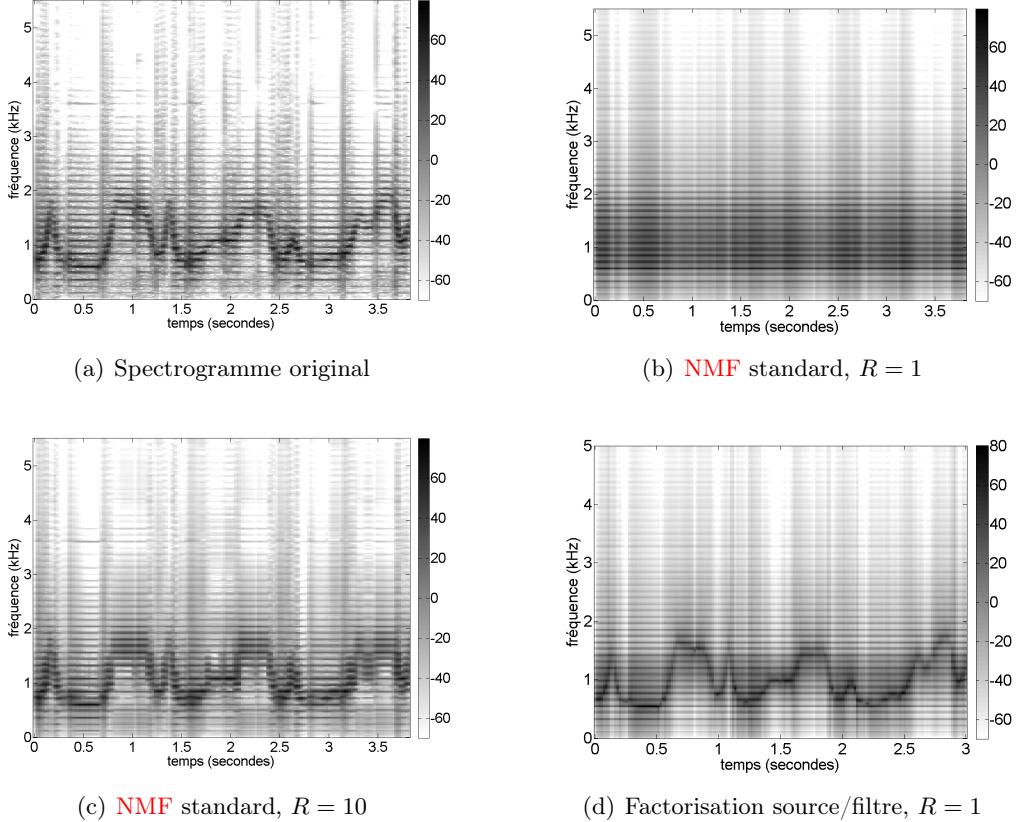


FIG. 3.1 – Spectrogramme de puissance original de l'extrait de guimbarde 3.1(a) et spectrogrammes reconstruits 3.1(b), 3.1(c) et 3.1(d).

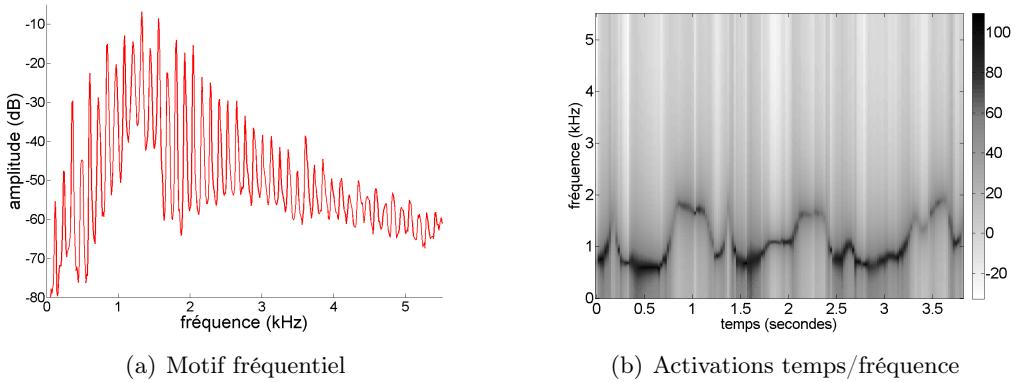


FIG. 3.2 – Factorisation source/filtre ($R = 1$, $Q = 0$ et $P = 2$) du spectrogramme de puissance du son de guimbarde.

3.3.2 Didgeridoo

3.3.2.1 Description du signal décomposé

Notre algorithme est ici appliqué à un court extrait de didgeridoo. Le didgeridoo est un instrument à vent ethnique provenant du nord de l'Australie. Cet instrument produit un son continu modulé généré par la vibration des lèvres de l'instrumentiste. Les modulations sont produites par la position de la bouche et la gorge du musicien ce qui lui permet de contrôler plusieurs résonances.

La figure 3.3(a) représente le spectrogramme de l'extrait : le son produit est quasiment harmonique (avec un peu de bruit) et une importante résonance variante dans le temps apparaît clairement dans le spectrogramme. On peut donc considérer que ce signal est composé d'un unique événement comprenant de fortes variations spectrales, et essayer de le décomposer avec un unique atome ($R = 1$). La fréquence d'échantillonnage de l'extrait est $f_s = 11025\text{Hz}$. Nous avons choisi une fenêtre de Hann de 1024 échantillons (93ms) avec un recouvrement de 75% pour le calcul de la TFCT.

3.3.2.2 Expérience et résultat

Le spectrogramme de l'extrait est décomposé d'une part avec un algorithme de NMF standard pour $R = 1$ atome et $R = 5$ atomes, d'autre part avec la factorisation source/filtre pour $R = 1$ atome, avec un filtre AR d'ordre 3 ($Q = 0$ et $P = 3$). Les spectrogrammes reconstruits sont respectivement représentés dans les figures 3.3(b), 3.3(c) et 3.3(d).

Bien que le didgeridoo soit joué seul dans le spectrogramme analysé, la NMF standard a besoin de plusieurs atomes pour décomposer précisément le spectrogramme de puissance. Avec 1 atome, la NMF n'est pas capable de représenter correctement la résonance variable (figure 3.3(b)). Avec 5 atomes, certaines variations spectrales apparaissent (figure 3.3(c)), mais la trajectoire de la résonance reste très floue. De plus, les atomes ne font pas sens : chaque atome représente une partie du son global et est difficilement interprétable de façon perceptive individuellement. Enfin, la dimension totale des paramètres est importante ($FR + RT = 3290$).

La factorisation source/filtre permet quant à elle de représenter efficacement les variations spectrales du son (figure 3.3(d)) avec un unique atome, tout en gardant une dimension des paramètres relativement faible ($FR + TR(Q+1) = 1093$) : la résonance variable du son original est bien modélisée, et l'erreur totale \mathcal{C} est plus petite que celle obtenue avec une NMF standard avec $R = 5$. Dans ce cas, la décomposition proposée est donc plus efficace et pertinente qu'une NMF standard.

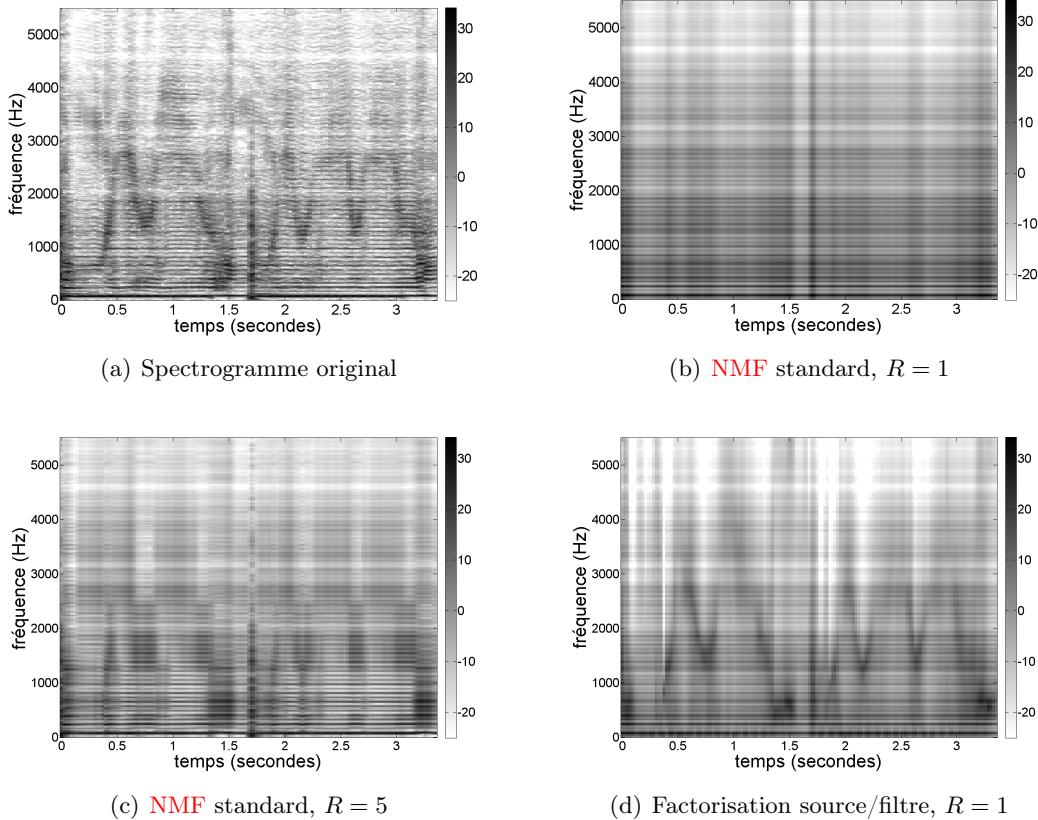


FIG. 3.3 – Spectrogramme de puissance original de l'extrait de didgeridoo 3.3(a) et spectrogrammes reconstruits 3.3(b), 3.3(c) et 3.3(d).

3.3.3 Clavecin

3.3.3.1 Description du signal décomposé

Dans cette section, notre algorithme est appliqué à un court extrait de clavecin, composé de deux notes différentes (*Do2* et *Mib2*) : d'abord, le *Do2* est joué seul, puis le *Mib2*, et enfin, les deux notes sont jouées simultanément. Le spectrogramme de l'extrait est représenté dans la figure 3.4(a). Comme pour de nombreux instruments à corde libre, les partiels de haute fréquence des sons produits par un clavecin décroissent plus rapidement que les partiels de basse fréquence. Ce phénomène apparaît très clairement dans le spectrogramme en forme de L de la figure 3.4(a) (le L est ajouté sur la figure pour représenter la forme caractéristique des notes). La fréquence d'échantillonnage de l'extrait est $f_s = 44100\text{Hz}$. Nous avons choisi une fenêtre de Hann de 2048 échantillons (46ms) avec 75% de recouvrement pour la TFCT.

3.3.3.2 Expérience et résultat

Le spectrogramme a été décomposé à l'aide d'un algorithme de NMF standard avec $R = 2$ atomes (1 atome par note) et $R = 6$ atomes, et à l'aide la factorisation source/filtre

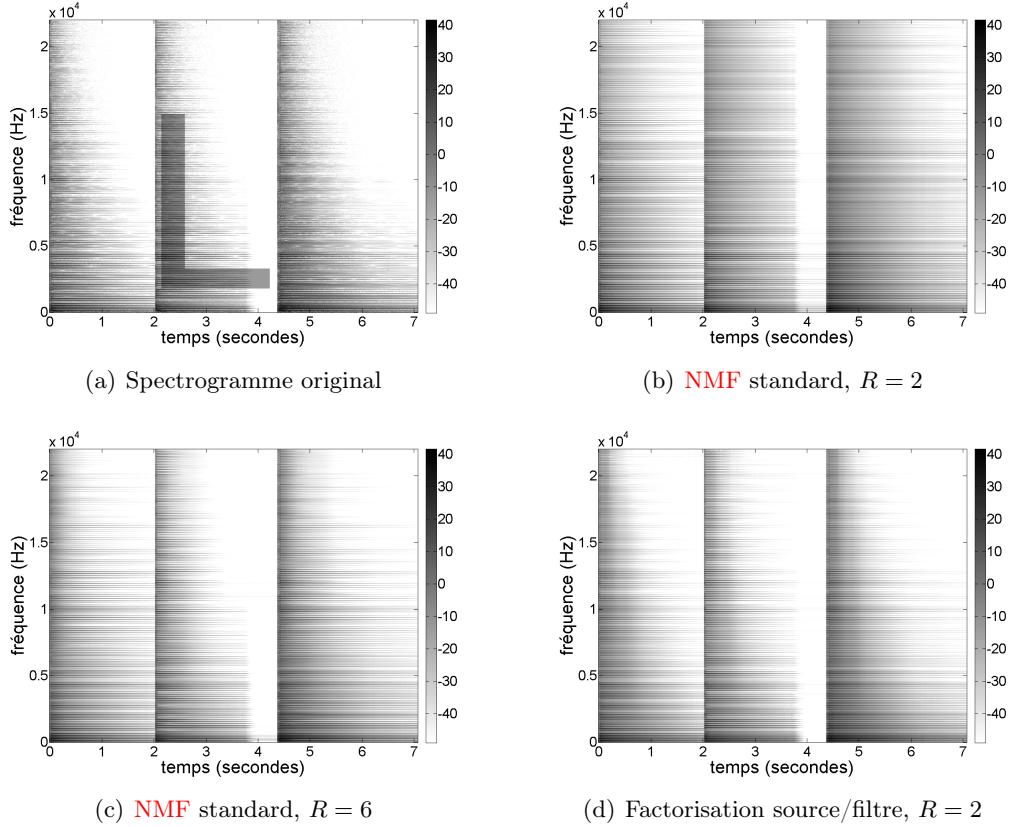


FIG. 3.4 – Spectrogramme original de puissance de l'extrait de clavecin 3.4(a) et spectrogrammes reconstruits 3.4(b), 3.4(c) et 3.4(d).

avec $R = 2$ atomes, et un filtre ARMA de paramètre $Q = 1$ et $P = 1$. Les spectrogrammes reconstruits sont représentés dans les figures 3.4(b), 3.4(c) et 3.4(d).

La NMF standard nécessite plusieurs atomes par note pour décomposer correctement le spectrogramme de puissance en forme de L : avec seulement deux atomes (un par note jouée), la décroissance rapide des hautes fréquences n'apparaît pas du tout (figure 3.4(b)). Avec 6 atomes, l'atténuation des partiels de haute fréquence apparaît (figure 3.4(c)), mais chaque atome est alors une partie d'une note et n'a pas vraiment de sens perceptif.

Le modèle ARMA inclus dans notre algorithme permet d'obtenir une bonne description de la forme globale du spectrogramme. 2 atomes (1 par note) sont suffisants pour représenter avec précision le spectrogramme original : chaque atome est harmonique (figure 3.5(a)) et correspond à une note et la décroissance rapide des partiels de haute fréquence est clairement bien modélisée par le modèle ARMA comme le montrent les activations temps/fréquence $h_{rt}^{\text{ARMA}}(f)$ représentées dans la figure 3.5(b). La dimension totale des paramètres fournis par notre algorithme ($FR + TR(Q + P + 1) = 5704$) reste plus faible que celle obtenue avec une NMF standard avec 6 atomes ($FR + RT = 9804$) et l'erreur globale \mathcal{C} entre spectrogramme original et reconstruit est approximativement la même que celle donnée par une NMF standard avec $R = 6$.

Par conséquent, la décomposition fournie par la factorisation source/filtre semble don-

ner une représentation ayant physiquement plus de sens que celle d'une **NMF** standard.

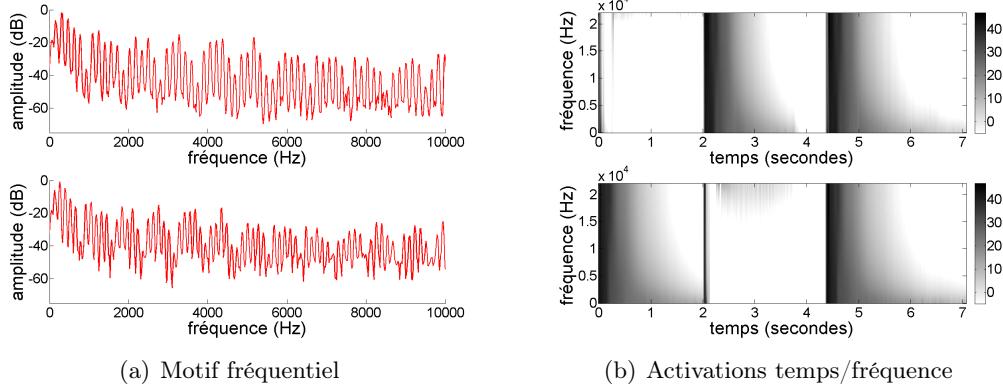


FIG. 3.5 – Factorisation source/filtre ($R = 2$, $Q = 1$ et $P = 1$) du spectrogramme de puissance de l'extrait de clavecin.

3.3.4 Guitare avec pédale wah-wah

3.3.4.1 Description du signal décomposé

Dans cette section, notre algorithme est utilisé pour décomposer un court extrait de guitare électrique traité par une pédale wah-wah. La pédale wah-wah est un effet très utilisé sur la guitare électrique. Elle consiste en un filtre résonant, dont la fréquence de résonance est contrôlée au moyen d'une pédale. Cette effet est nommé en raison de la ressemblance des sons produits avec l'onomatopée « Wah ». Le son d'une note de guitare électrique traité par une pédale wah-wah dont on modifie la position de la pédale présente de fortes variations spectrales et ne peut donc pas être correctement représenté par un unique atome dans une **NMF** standard.

Comme l'effet produit par une pédale wah-wah est bien modélisé par un filtre **AR** avec 2 pôles complexes conjugués, nous avons choisi de décomposer l'extrait avec $Q = 0$ et $P = 2$. L'extrait décomposé, dont le spectrogramme est représenté dans la figure 3.6(a), est composé de trois notes différentes jouées successivement (la première note est rejouée une seconde fois à la fin de l'extrait). Chaque note peut être vue comme un motif harmonique qui est filtré par un filtre résonant, la fréquence de résonance variant entre 400Hz et 1200Hz : cette résonance apparaît très clairement dans le spectrogramme de puissance. La fréquence d'échantillonnage de l'extrait est $f_s = 11025\text{Hz}$. Nous avons choisi une fenêtre de Hann de 1024 échantillons (93ms) avec 75% de recouvrement pour la **TFCT**.

3.3.4.2 Expérience et résultat

Comme le son analysé présente d'importantes variations spectrales, la **NMF** standard nécessite de nombreux atomes pour décomposer efficacement le spectrogramme de puissance de la figure 3.6(a). Ainsi, il n'y a plus de correspondance immédiate entre note et atome, et la décomposition ne correspond plus à l'analyse qui pourrait être faite du son par un auditeur humain. La figure 3.6(b) représente le spectrogramme de puissance reconstruit

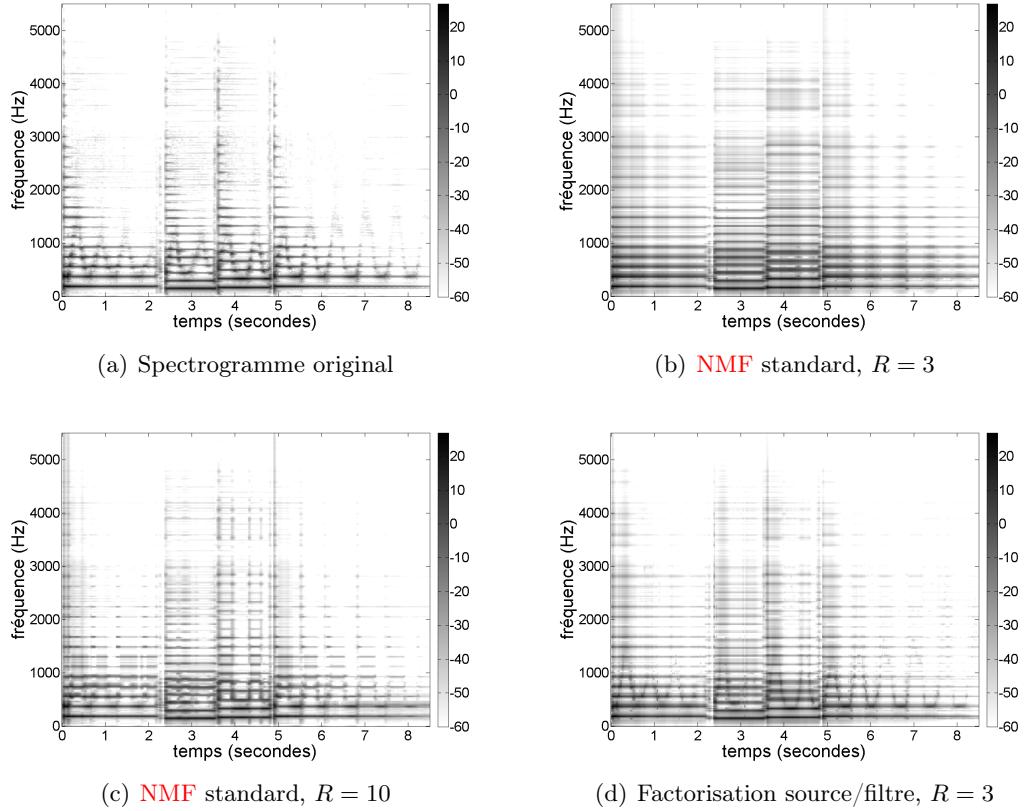


FIG. 3.6 – Spectrogramme de puissance original de l'extrait de guitare électrique traité par une pédale wah-wah 3.6(a) et spectrogrammes reconstruits 3.6(b), 3.6(c) et 3.6(d).

à partir de la **NMF** du spectrogramme original avec 3 atomes et la figure 3.6(c) avec 10 atomes. Avec 3 atomes, la **NMF** n'est pas capable de modéliser la résonance de l'effet. Avec 10 atomes, la résonance apparaît, mais le signal n'est pas décrit de façon « intelligente » : chaque atome est une partie d'une note et n'a par conséquent pas de sens perceptif clair. De plus la dimension totale des paramètres du problème est plus grande ($FR + RT = 8790$).

Avec une factorisation source/filtre, les fortes variations spectrales de chaque note peuvent être représentées avec précision en utilisant des filtres **AR** d'ordre 2 ($Q = 0$ et $P = 2$) comme montré dans la figure 3.6(d). Ainsi, 3 atomes (un pour chaque note) sont suffisants pour modéliser correctement le spectrogramme original. En effet, l'erreur globale de reconstruction \mathcal{C} entre le spectrogramme original et le spectrogramme reconstruit obtenue avec la factorisation source/filtre est approximativement la même que celle obtenue avec une **NMF** standard avec 10 atomes et environ la moitié de celle obtenue avec une **NMF** standard avec 3 atomes. Chaque atome est harmonique et correspond à une note, et la résonance de la pédale wah-wah apparaît clairement. La dimension totale de la représentation obtenue avec la factorisation source/filtre reste environ deux fois moindre que celle d'une **NMF** avec 10 atomes : $FR + R(Q + 1)T = 4833$. La décomposition proposée est donc capable d'extraire un spectre stationnaire « moyen » pour chaque note de

guitare (contenu dans \mathbf{W}) du son non-stationnaire produit par l'effet de la pédale wah-wah, cette non-stationnarité étant quant à elle décrite par les activations temps/fréquence. Les 3 motifs fréquentiels (atomes de \mathbf{W}) obtenus sont représentés dans la figure 3.7(a) : chaque motif est harmonique avec sa propre fréquence fondamentale, et correspond donc à une note (la NMF standard avec 3 atomes fournit des motifs similaires). Les activations temps/fréquence ($h_{rt}^{\text{ARMA}}(f)$) sont représentées dans la figure 3.7(b) : la résonance de la pédale wah-wah apparaît clairement aux moments auxquels les notes sont jouées. Par conséquent, la décomposition fournie par notre algorithme semble donner une représentation plus significative du spectrogramme considéré.

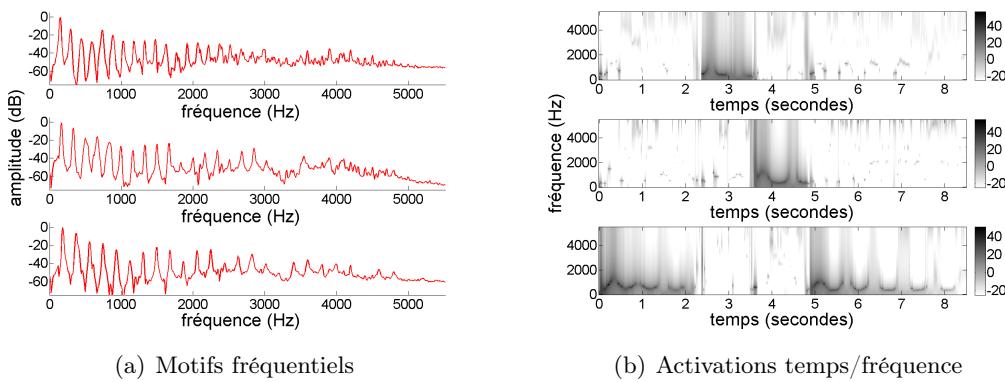


FIG. 3.7 – Factorisation source/filtre ($R = 3$ et $P = 2$) du spectrogramme de puissance du son de guitare électrique traité par une pédale wah-wah.

3.3.5 Convergence de l'algorithme

L'évolution de la fonction de coût au cours des itérations pour une factorisation source/filtre est représentée dans la figure 3.8 avec 8 différentes initialisations aléatoires, pour la décomposition des extraits présentés dans les sections 3.3.3 (extrait de clavecin) et 3.3.4 (extrait de guitare traité par une pédale wah-wah). La valeur de la β -divergence est représentée après chaque itération (la valeur initiale, avant la première itération, n'est pas représentée car elle est beaucoup plus importante que les valeurs suivantes). Les figures montrent une décroissance monotone de la fonction de coût et une convergence apparente. Dans la figure 3.8(a), toutes les initialisations conduisent à la même valeur finale de la fonction de coût et l'aspect global de l'évolution est très similaire pour toutes les initialisations. En revanche, dans la figure 3.8(b), toutes les initialisations ne conduisent pas à la même valeur de la fonction de coût, ce qui montre qu'une initialisation multiple peut être utile.

3.4 Conclusion

Dans ce chapitre, nous avons proposé un nouvel algorithme itératif qui constitue une extension de la NMF basée sur un modèle source/filtre. Nous avons montré que cette

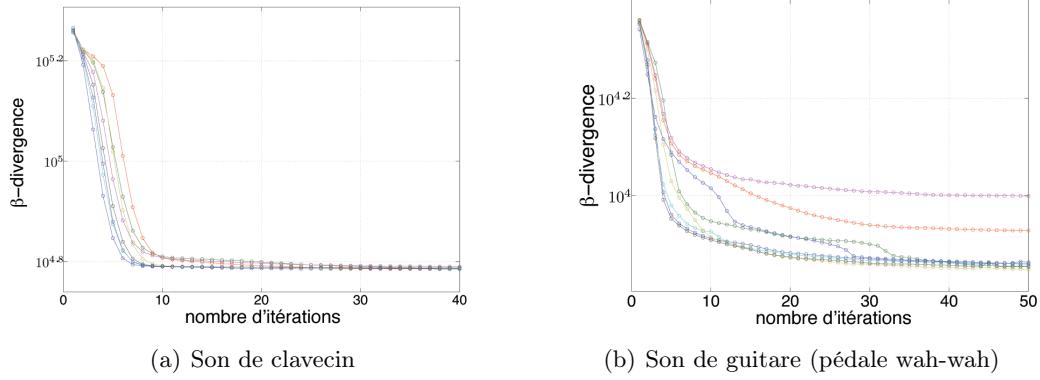


FIG. 3.8 – Evolution de la fonction de coût (β -divergence avec $\beta = 0.5$) au cours des itérations (décomposition des extraits des sections 3.3.3 et 3.3.4).

représentation est particulièrement adaptée pour décomposer efficacement et avec sens des objets sonores non-stationnaires contenant d'importantes variations spectrales.

Il pourrait être intéressant d'introduire des contraintes de continuité entre les filtres d'une trame à l'autre en suivant par exemple l'approche de [Virtanen, 2007, Bertin *et al.*, 2010].

Par ailleurs, cette décomposition ne permet pas de prendre en compte des variations de fréquence fondamentale (comme on peut en trouver dans le vibrato). Les variations de fréquence fondamentale font l'objet du chapitre suivant.

Chapitre 4

Modélisation des variations de fréquence fondamentale

Ce chapitre est consacré à l'étude des variations de fréquence fondamentale au cours du temps. Deux types de décomposition ont été proposés : le premier type de décomposition consiste à construire des atomes harmoniques paramétriques pour lesquels la fréquence fondamentale est un paramètre. Le second type est une adaptation des méthodes de décomposition invariantes par translation fréquentielle : alors que ces dernières décomposent des spectrogrammes à Q-constant (résolution fréquentielle logarithmique), la méthode proposée permet de décomposer des spectrogrammes standard ce qui permet une utilisation aisée en séparation de sources.

4.1 Spectrogramme paramétrique

Dans cette section, nous présentons une nouvelle méthode de décomposition des spectrogrammes musicaux dérivée de la Factorisation en matrices non-négatives (**NMF**). Cette méthode utilise des atomes harmoniques dont la fréquence fondamentale peut varier dans le temps : ces atomes correspondent à des notes de musique. Ces atomes sont paramétriques et sont construits à l'aide des valeurs des paramètres qui sont appris à partir des données dans un cadre assez similaire à la **NMF**. Cette paramétrisation des spectrogrammes permet de modéliser avec précision certains effets musicaux (tels que le vibrato) qui sont difficilement analysables avec une **NMF**.

Ce travail a fait l'objet d'une publication dans [Hennequin *et al.*, 2010c].

4.1.1 Modèle

Dans le modèle proposé, les motifs fréquentiels deviennent paramétriques et peuvent varier dans le temps, à travers un paramètre θ_{rt} . L'équation de la **NMF** standard (2.1) page 23 est donc remplacée par :

$$[\mathbf{V}]_{ft} \approx [\hat{\mathbf{V}}]_{ft} = \sum_{r=1}^R w_{fr}^{\theta_{rt}} h_{rt}, \quad (4.1)$$

où θ_{rt} est le paramètre associé à l'atome r à l'instant t . Ce paramètre peut être interprété comme l'« état » de l'atome à l'instant considéré : l'atome r est synthétisé à partir de la valeur de ce paramètre à chaque instant t . La dépendance temporelle de ce paramètre va permettre de modéliser des phénomènes non-stationnaires tels que le vibrato.

Le paramètre choisi ici est la fréquence fondamentale instantanée de l'atome : $\theta_{rt} = f_0^{rt}$. Chaque atome va être construit comme un peigne harmonique paramétré par cette fréquence fondamentale.

4.1.1.1 Atome harmonique paramétrique

L'expression de l'atome paramétrique est donc la suivante :

$$w_{fr}^{f_0^{rt}} = \sum_{k=1}^{n_h(f_0^{rt})} a_k g(f - kf_0^{rt}). \quad (4.2)$$

Cet atome correspond dans le domaine temporel à un son périodique stationnaire fenêtré, c'est-à-dire une somme de sinusoïdes fenêtrée. La transformée de Fourier d'un signal périodique de fréquence fondamentale f_0 dont les amplitudes des harmoniques sont notées a_k est une somme de distributions de Dirac centrées respectivement en kf_0 d'amplitude a_k (avec $k \in \mathbb{Z}^*$). La transformée de Fourier de ce même signal fenêtré est alors le produit de convolution de la somme d'impulsions précédente avec la transformée de Fourier de la fenêtre. Comme l'atome construit doit être non-négatif, nous avons pris le module au carré de cette transformée de Fourier. Afin de rendre les calculs plus simples, le module au carré de la somme des harmoniques est remplacé par la somme des modules au carré de chaque harmonique : g est le module au carré de la transformée de Fourier de la fenêtre utilisée dans le calcul de la Transformée de Fourier à Court Terme (**TFCT**). Les interférences entre deux partiels successifs sont donc négligées ; cette approximation est justifiée pour des fréquences fondamentales suffisamment grandes (ou de manière équivalente, pour des fenêtres d'analyse suffisamment longues). Le choix du module au carré plutôt que directement du module permet à g de rester une fonction dérivable, ce qui permet d'utiliser des algorithmes de minimisation standard.

Un exemple d'un tel atome paramétrique est représenté dans la figure 4.1.

Le nombre d'harmoniques est noté $n_h(f_0^{rt})$. Les amplitudes a_k de chaque harmonique sont ici supposées identiques pour tous les atomes et sont apprises de façon non supervisée. Il est aussi possible de considérer des jeux d'amplitudes différents pour chaque atome, cependant, ce choix augmente de façon importante les problèmes d'ambiguïté d'octave (de douzième, de double octave...).

Dans le modèle proposé, on fait les hypothèses suivantes :

- La partie harmonique des notes est supposée stationnaire au sein d'une trame d'analyse (ainsi la transformée de Fourier d'un harmonique isolé est la transformée de Fourier de la fenêtre d'analyse centrée sur la fréquence de l'harmonique) ;
- Les interférences entre les harmoniques sont supposées négligeables (cette hypothèse est valable pour des fréquences fondamentales suffisamment élevées) ;
- Les interférences entre les fréquences négatives (non prises en compte) et positives sont également supposées négligeables ;

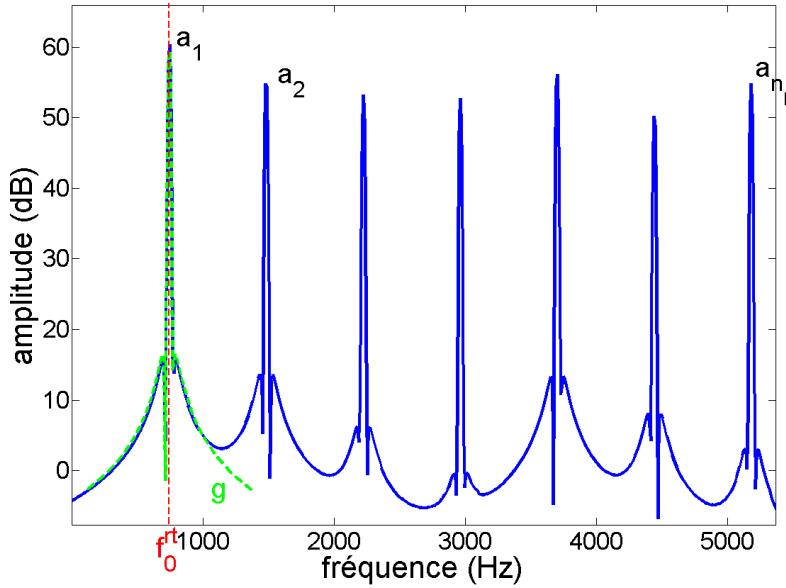


FIG. 4.1 – Atome paramétrique $w_{f_r}^{f_r t}$. Le carré du module de la transformée de Fourier de la fenêtre d'analyse g est représenté en vert pointillé.

- Le repliement spectral introduit par l'échantillonnage du signal est supposé faible (ce qui permet d'utiliser l'expression analytique de la transformée de Fourier continue de la fenêtre d'analyse) ;
- Les hypothèses classiques de la **NMF** sur la sommation des composantes positives sont faites.

Les deux premières hypothèses sont les hypothèses les plus fortes et en pratique la deuxième hypothèse entraîne les troisième et quatrième hypothèses.

4.1.1.2 Expression de g

La fonction g peut prendre différentes formes suivant la fenêtre d'analyse choisie. Nous donnons ici son expression ainsi que l'expression de sa dérivée pour une fenêtre gaussienne et pour une fenêtre de type cosinus (dont les fenêtres de Hann et de Hamming sont des cas particuliers).

Dans cette section, nous utilisons la définition suivante de la transformée de Fourier d'un signal continu x :

$$\hat{X}(f) = \int_{-\infty}^{+\infty} x(t) e^{-i2\pi ft} dt.$$

Fenêtre gaussienne : La fenêtre gaussienne est définie dans le domaine temporel par l'expression :

$$h(t) = e^{-\frac{t^2}{\sigma^2}},$$

où σ caractérise la largeur de la fenêtre.

La transformée de Fourier de cette fenêtre est donnée par :

$$\hat{H}(f) = \frac{e^{-\sigma^2\pi^2f^2}}{\sqrt{2}\sigma^2}.$$

Par conséquent, l'expression de g pour ce type de fenêtre est :

$$g(f) = |\hat{H}(f)|^2 = \frac{e^{-2\sigma^2\pi^2f^2}}{2\sigma^4}.$$

La dérivée de g est alors :

$$g'(f) = -\frac{2\pi^2 f e^{-2\sigma^2\pi^2f^2}}{\sigma^2}.$$

On peut remarquer que pour tout $f \in \mathbb{R}$ $\frac{g'(f)}{f} \leq 0$, ce qui constituera une propriété essentielle pour établir simplement les règles de mise à jour multiplicatives de f_0 dans l'algorithme d'estimation des paramètres..

La fenêtre gaussienne a de bonnes propriétés fréquentielles (elle possède un unique lobe et elle sature l'inégalité d'Heisenberg) mais reste rarement utilisée du fait notamment de son support infini (la tronquer lui fait perdre ses bonnes propriétés).

Fenêtre « cosinus » : Les fenêtres de type « cosinus » sont définies dans le domaine temporel par :

$$h(t) = (\alpha - \beta \cos(2\pi \frac{t}{T})) \mathbb{1}_{[0,T]}(t),$$

où T est la longueur de la fenêtre, et $\alpha + \beta = 1$ (le maximum de la fonction est égal à 1).

Ce type de fenêtre englobe :

- la fenêtre de Hann (parfois également appelée fenêtre de Hanning), pour $\alpha = \beta = 0.5$.
- la fenêtre de Hamming, pour $\alpha = 0.54$ et $\beta = 0.46$.

La transformée de Fourier de cette fenêtre est donnée par :

$$\hat{H}(f) = \frac{ie^{-i2\pi Tf}(-1 + e^{i2\pi Tf})(T^2f^2(\beta - \alpha) + \alpha)}{2\pi(T^2f^3 - f)}.$$

Par conséquent, pour ce type de fenêtre, l'expression de g est donné par :

$$g(f) = |\hat{H}(f)|^2 = \frac{1}{4\pi^2} (2 - 2 \cos(2\pi Tf)) \frac{(T^2f^2(\beta - \alpha) + \alpha)^2}{f^2(T^2f^2 - 1)^2}.$$

g peut être \mathcal{C}^1 -prolongée de façon triviale en 0 et $\pm T$ en posant $g(0) = \alpha^2T^2$ et $g(\pm T) = \frac{\beta^2T^2}{4}$. La dérivée de g est alors donnée par :

$$g'(f) = \frac{1}{4\pi^2} \left[(2 - 2 \cos(2\pi Tf)) \frac{2(f^2T^2(\beta - \alpha) + \alpha)}{f^2(f^2T^2 - 1)^2} \left(2fT^2(\beta - \alpha) - \frac{2fT^2(f^2T^2(\beta - \alpha) + \alpha)}{f^2T^2 - 1} \right. \right. \\ \left. \left. - \frac{f^2T^2(\beta - \alpha) + \alpha}{f} \right) + 4\pi T \sin(2\pi Tf) \frac{(T^2f^2(\beta - \alpha) + \alpha)^2}{f^2(T^2f^2 - 1)^2} \right].$$

Pour une fenêtre de Hann ou de Hamming, le lobe principal correspond aux fréquences $f \in [-\frac{2}{T}, \frac{2}{T}]$. Pour ces fréquences, on constate que $\frac{g'(f)}{f} \leq 0$ ce qui, une fois de plus, sera utilisé pour établir les règles de mise à jour de f_0 dans l'algorithme d'estimation des paramètres.

4.1.1.3 Fonction de coût et nombre d'atomes

Comme pour une **NMF**, la fonction de coût à minimiser est une divergence entre le spectrogramme original et le spectrogramme reconstruit :

$$\mathcal{C}(\Theta, \mathbf{H}, \mathbf{a}) = D(\mathbf{V} | \hat{\mathbf{V}}) = \sum_{ft} d(v_{ft} | \hat{v}_{ft}), \quad (4.3)$$

où $\Theta = (\theta_{rt})_{r \in \llbracket 1, R \rrbracket, t \in \llbracket 1, T \rrbracket}$, $\mathbf{H} = (h_{rt})_{r \in \llbracket 1, R \rrbracket, t \in \llbracket 1, T \rrbracket}$, et $\mathbf{a} = (a_k)_{k \in \llbracket 1, K \rrbracket}$.

Nous utilisons ici une β -divergence. On va alors chercher à minimiser la fonction de coût par rapport à h_{rt} , f_0^{rt} et a_k pour $r \in \llbracket 1, R \rrbracket$, $t \in \llbracket 1, T \rrbracket$ et $k \in \llbracket 1, K \rrbracket$.

Cependant, la fonction de coût est fortement non-convexe par rapport à f_0^{rt} : ce phénomène est illustré dans la figure 4.2 pour des valeurs fixées de r et t . On peut observer de nombreux minima locaux : les 2 principaux minima correspondent à des notes effectivement jouées, et plusieurs minima apparaissent au niveau de l'octave, la sous-octave, la douzième... de chacune de ces notes (cette figure peut être interprétée comme l'opposé d'une fonction similaire à un produit spectral). Par conséquent, une optimisation globale semble vouée à l'échec : en effet, dans l'exemple de la figure 4.2, un atome dont la fréquence fondamentale serait initialisée à 800Hz convergerait certainement vers 880Hz. Afin d'éviter ce problème, un atome est introduit pour chaque degré de la gamme chromatique, ce qui permet de remplacer le problème d'optimisation globale par plusieurs sous-problèmes d'optimisation locale (qu'on espère être localement convexes). La fréquence fondamentale de l'atome r peut alors être finement estimée autour de $f_0^{rt} \approx f_0^{\text{ref}} 2^{\frac{r-1}{12}}$, où f_0^{ref} est la fréquence fondamentale de l'atome le plus bas. Les atomes qui ont une activation trop basse pourront être considérés comme absents et pourront être supprimés.

4.1.1.4 Atomes de **NMF** standard pour modéliser les parties non-harmoniques

Dans un spectrogramme musical, les événements percussifs (générés par des instruments percussifs ou au niveau des attaques de certains instruments harmoniques) sont très mal représentés par des atomes harmoniques et ne sont donc pas pris en compte correctement dans notre modèle. Afin d'inclure ce type d'événement, un second terme est ajouté au spectrogramme paramétrique de l'équation (4.1) : quelques atomes de **NMF** standard sont ajoutés à la décomposition. L'équation (4.1) est donc remplacée par :

$$v_{ft} \approx \hat{v}_{ft} = \sum_{r=1}^R w_{fr}^{\theta_{rt}} h_{rt} + \sum_{r=1}^{R'} w'_{fr} h'_{rt}. \quad (4.4)$$

Par conséquent les atomes de la matrice \mathbf{W}' ne varient pas dans le temps et sont censés modéliser les parties non harmoniques. R' doit rester assez faible (de l'ordre de 1 ou 2) afin que ces atomes libres ne représentent pas des événements harmoniques.

4.1.2 Algorithme

La minimisation de $\mathcal{C}(\Theta, \mathbf{H}, \mathbf{a})$ peut être faite au moyen d'un algorithme de descente multiplicatif similaire à ceux généralement utilisés en **NMF** : pour les fréquences fondamentales, le choix de mise à jour multiplicative est motivé par leur positivité et par leur distribution naturelle sur une échelle logarithmique.

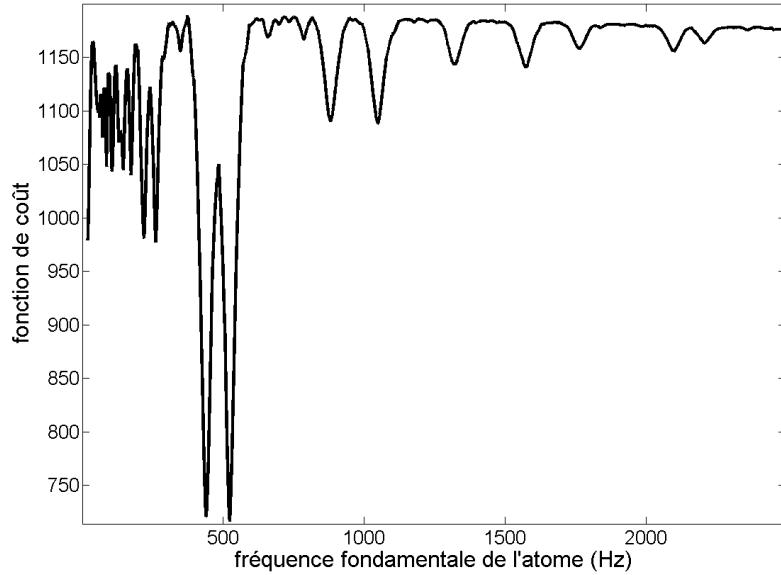


FIG. 4.2 – Fonction de coût exprimée en fonction de la fréquence fondamentale f_0^{rt} de l’atome r : le spectre analysé est un mélange de deux spectres harmoniques de fréquence fondamentale 440Hz et 523Hz.

Nous utilisons à nouveau la règle heuristique qui consiste à écrire la dérivée partielle de la fonction de coût par rapport à un paramètre comme une différence de deux termes positifs, le rapport de ces deux termes étant utilisé pour la mise à jour de ce paramètre (*cf. section 2.5.2.1*).

La dérivée partielle de la fonction de coût (4.3) par rapport à un paramètre λ est donnée par :

$$\frac{\partial \mathcal{C}}{\partial \lambda} = \sum_{ft} \frac{\partial d(v_{ft}, \hat{v}_{ft})}{\partial \lambda} = \sum_{ft} \frac{\partial \hat{v}_{ft}}{\partial \lambda} \frac{\partial d}{\partial y}(v_{ft}, \hat{v}_{ft}),$$

où $\frac{\partial d}{\partial y}$ désigne la dérivée partielle de d par rapport à son second argument. En utilisant une β -divergence, cette dérivée partielle s’écrit :

$$\frac{\partial d}{\partial y}(v_{ft}, \hat{v}_{ft}) = \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}).$$

Ainsi,

$$\frac{\partial \mathcal{C}}{\partial \lambda} = \sum_{ft} \frac{\partial \hat{v}_{ft}}{\partial \lambda} \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}).$$

4.1.2.1 Mise à jour de f_0

Afin d’obtenir la règle de mise à jour du paramètre Θ , on a besoin de la dérivée partielle de la fonction de coût par rapport à $\theta_{r_0 t_0}$:

$$\frac{\partial \mathcal{C}}{\partial \theta_{r_0 t_0}} = \sum_{ft} \frac{\partial \hat{v}_{ft}}{\partial \theta_{r_0 t_0}} \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}).$$

La dérivée partielle du spectrogramme paramétrique \hat{v}_{ft} (défini dans l'équation (4.1)) par rapport à $\theta_{r_0 t_0}$ est donnée par :

$$\frac{\partial \hat{v}_{ft}}{\partial \theta_{r_0 t_0}} = \delta_{tt_0} h_{r_0 t_0} \frac{\partial w_{fr_0}^{\theta_{r_0 t_0}}}{\partial \theta_{r_0 t_0}}.$$

Lorsque Θ est la fréquence fondamentale de chaque atome à chaque instant, la dérivée partielle de l'atome par rapport à sa fréquence fondamentale est obtenue à partir de l'équation (4.2) :

$$\frac{\partial w_{fr_0}}{\partial f_0^{r_0 t_0}} = - \sum_{k=1}^{n_h} a_k k g'(f - k f_0^{r_0 t_0}).$$

La dérivée partielle de la fonction de coût par rapport à $f_0^{r_0 t_0}$ vaut donc :

$$\frac{\partial \mathcal{C}}{\partial f_0^{r_0 t_0}} = - \sum_f \sum_{k=1}^{n_h} h_{r_0 t_0} a_k k g'(f - k f_0^{r_0 t_0}) (\hat{v}_{ft_0}^{\beta-1} - \hat{v}_{ft_0}^{\beta-2} v_{ft_0}).$$

Quand g a un unique lobe (par exemple lorsqu'on utilise une fenêtre d'analyse gaussienne dans le calcul de la **TFCT**), la remarque suivante n'est pas un problème. Cependant, il est courant que g ait plusieurs lobes, et qu'ainsi g' change de signe en de nombreux points. Afin de faciliter les calculs, seul le support du lobe principal de g (dénoté Λ) sera gardé dans l'expression de la dérivée :

$$g'(f) \approx g'_a(f) = g'(f) \mathbb{1}_\Lambda(t). \quad (4.5)$$

En supposant que le lobe principal a un unique maximum local — ce qui est vérifié pour les fenêtres de Hann et de Hamming (*cf.* figure 4.3) — la fonction $f \mapsto -g'_a(f)/f$ est alors non-négative (voir figure 4.3) et on peut écrire :

$$g'_a(f - k f_0^{r_0 t_0}) = -(f - k f_0^{r_0 t_0}) P(f - k f_0^{r_0 t_0}), \quad (4.6)$$

où P est une fonction positive continue.

En utilisant l'approximation (4.5), la dérivée partielle de la fonction de coût par rapport à $f_0^{r_0 t_0}$ peut être écrite simplement comme la différence de deux termes positifs :

$$\frac{\partial \mathcal{C}}{\partial f_0^{r_0 t_0}} \approx \mathcal{G}_{r_0 t_0} - \mathcal{F}_{r_0 t_0}, \quad (4.7)$$

avec

$$\mathcal{G}_{r_0 t_0} = \sum_{f,k} h_{r_0 t_0} a_k k P(f - k f_0^{r_0 t_0}) \hat{v}_{ft_0}^{\beta-2} (f \hat{v}_{ft_0} + k f_0^{r_0 t_0} v_{ft_0})$$

et

$$\mathcal{F}_{r_0 t_0} = \sum_{f,k} h_{r_0 t_0} a_k k P(f - k f_0^{r_0 t_0}) \hat{v}_{ft_0}^{\beta-2} (k f_0^{r_0 t_0} \hat{v}_{ft_0} + f v_{ft_0}).$$

On en déduit la règle de mise à jour de f_0 :

$$f_0^{r_0 t_0} \leftarrow f_0^{r_0 t_0} \frac{\mathcal{F}_{r_0 t_0}}{\mathcal{G}_{r_0 t_0}}. \quad (4.8)$$

²Afin de pouvoir représenter toutes les courbes sur le même dessin l'axe des ordonnées a été modifié d'une courbe à l'autre et est donc arbitraire.

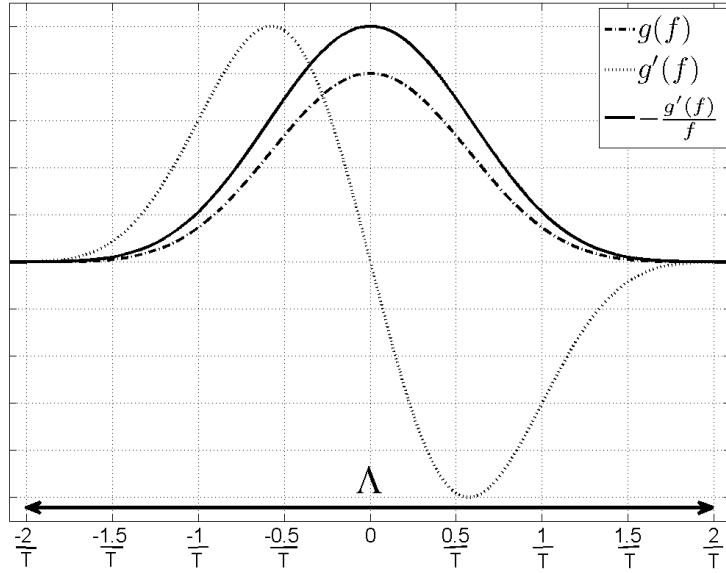


FIG. 4.3 – Lobe principal de g , dérivée de g et positivité de $P(f) = -\frac{g'(f)}{f}$ sur $\Lambda = [-\frac{2}{T}, \frac{2}{T}]$ pour une fenêtre de Hamming de longueur T .²

Le spectrogramme va alors être décomposé en utilisant un unique atome par demi-ton. Cependant, la règle de mise à jour des f_0 ne garantit pas que f_0^{rt} reste dans la bande de fréquence fondamentale qui lui a été attribuée ($f_0^{rt} \in [f_0^{\text{ref}} 2^{\frac{r-1}{12} - \frac{1}{24}}, f_0^{\text{ref}} 2^{\frac{r-1}{12} + \frac{1}{24}}]$) et ne glisse pas dans la bande d'un autre demi-ton. Par conséquent, l'évolution de f_0^{rt} doit être restreinte à cette bande : dans notre algorithme, quand une fréquence fondamentale quitte cette bande, on considère que l'atome associé ne doit pas être actif à ce moment et on impose alors une activation nulle.

4.1.2.2 Mise à jour de \mathbf{H}

La règle de mise à jour de \mathbf{H} est obtenue de façon très similaire à celle des activations d'une NMF standard en calculant la dérivée de la fonction de coût par rapport à $h_{r_0 t_0}$:

$$\frac{\partial \mathcal{C}}{\partial h_{r_0 t_0}} = \sum_{ft} \frac{\partial \hat{v}_{ft}}{\partial h_{r_0 t_0}} \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}) = \sum_f w_{fr_0}^{\theta_{r_0 t_0}} \hat{v}_{ft_0}^{\beta-2} (\hat{v}_{ft_0} - v_{ft_0}) = \mathcal{P}_{r_0 t_0} - \mathcal{M}_{r_0 t_0}, \quad (4.9)$$

où $\mathcal{P}_{r_0 t_0} = \sum_f w_{fr_0}^{\theta_{r_0 t_0}} \hat{v}_{ft_0}^{\beta-1}$ et $\mathcal{M}_{r_0 t_0} = \sum_f w_{fr_0}^{\theta_{r_0 t_0}} \hat{v}_{ft_0}^{\beta-2} v_{ft_0}$ sont des termes positifs.

La règle de mise à jour de $h_{r_0 t_0}$ est alors donnée par :

$$h_{r_0 t_0} \leftarrow h_{r_0 t_0} \frac{\mathcal{M}_{r_0 t_0}}{\mathcal{P}_{r_0 t_0}}. \quad (4.10)$$

4.1.2.3 Mise à jour de \mathbf{a}

La règle de mise à jour de \mathbf{a} est obtenue de manière analogue, en calculant la dérivée de la fonction de coût par rapport à a_k :

$$\frac{\partial \mathcal{C}}{\partial a_k} = \sum_{ft} \frac{\partial \hat{v}_{ft}}{\partial a_k} \hat{v}_{ft}^{\beta-2} (\hat{v}_{ft} - v_{ft}). \quad (4.11)$$

La dérivée partielle de \hat{v}_{ft} par rapport à a_k est :

$$\frac{\partial \hat{v}_{ft}}{\partial a_k} = \sum_{r=1}^R g(f - kf_0^{rt}) h_{rt} \mathbb{1}_{[1, n_h(f_0^{rt})]}(k). \quad (4.12)$$

En notant r_k la valeur maximum de r pour laquelle $k \in [1, n_h(f_0^{rt})]$, l'équation (4.12) devient :

$$\frac{\partial \hat{v}_{ft}}{\partial a_k} = \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt}.$$

Par conséquent, la dérivée partielle de la fonction de coût par rapport à a_k peut être naturellement exprimée comme la différence de deux termes positifs :

$$\frac{\partial \mathcal{C}}{\partial a_k} = \mathcal{P}_k - \mathcal{M}_k, \quad (4.13)$$

avec

$$\mathcal{P}_k = \sum_{ft} \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt} \hat{v}_{ft}^{\beta-1},$$

$$\mathcal{M}_k = \sum_{ft} \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt} \hat{v}_{ft}^{\beta-2} v_{ft}.$$

La règle de mise à jour de a_k est alors :

$$a_k \leftarrow a_k \frac{\mathcal{M}_k}{\mathcal{P}_k}. \quad (4.14)$$

4.1.2.4 Mise à jour de \mathbf{W}' et \mathbf{H}'

Les règles de mises à jour de $\mathbf{W}' = (w'_{fr})_{f \in \llbracket 1F \rrbracket, r \in \llbracket 1R' \rrbracket}$ et $\mathbf{H}' = (h'_{rt})_{r \in \llbracket 1R' \rrbracket, t \in \llbracket 1T \rrbracket}$ sont quasiment identiques à celle d'une **NMF** standard :

$$\mathbf{W}' \leftarrow \mathbf{W}' \odot \frac{(\hat{\mathbf{V}}^{\odot \beta-2} \odot \mathbf{V}) \mathbf{H}'^T}{\hat{\mathbf{V}}^{\odot \beta-1} \mathbf{H}'^T}, \quad (4.15)$$

$$\mathbf{H}' \leftarrow \mathbf{H}' \odot \frac{\mathbf{W}'^T (\hat{\mathbf{V}}^{\odot \beta-2} \odot \mathbf{V})}{\mathbf{W}'^T \hat{\mathbf{V}}^{\odot \beta-1}}. \quad (4.16)$$

4.1.2.5 Contraintes

Des termes de contraintes peuvent être ajoutés à la fonction de coût afin de favoriser certaines propriétés de la décomposition, comme pour une NMF (*cf.* section 2.6.2). Les règles de mises à jour sont obtenues de la même façon que dans les sections précédentes en reprenant l'approche simple d'établissement des règles multiplicatives (*cf.* section 2.5.2.1). La dérivée partielle du terme de contrainte C_p par rapport au paramètre θ à mettre à jour est exprimée comme une différence de deux termes positifs :

$$\frac{\partial C_p}{\partial \theta} = p_\theta^p - m_\theta^p.$$

La règle de mise à jour (2.13) page 44 du paramètre θ devient :

$$\theta \leftarrow \theta \frac{m_\theta + m_\theta^p}{p_\theta + p_\theta^p},$$

où p_θ et m_θ ont été définis dans l'équation (2.12) page 44.

Plusieurs types de contraintes ont été considérés pour la décomposition proposée :

Contrainte de parcimonie sur les colonnes de \mathbf{H} : comme proposée dans [Virtanen, 2007, Hoyer, 2004]. Cette contrainte vise à réduire le nombre d'atomes actifs simultanément.

Contrainte de décorrélation entre les activations d'un atome et de son octave : (et éventuellement de sa douzième, sa double octave...) inspirée de celle proposée dans [Zhang et Fang, 2007], mais avec une expression différente. Cette contrainte vise à réduire les ambiguïtés d'octaves en pénalisant l'activation simultanée d'un atome et de son octave. Le terme de coût associé à cette contrainte (décorrélation des octaves) est :

$$C_\rho(\mathbf{H}) = \mu \sum_{r=1}^{R-12} \rho(h_{r,:}, h_{r+12,:}),$$

où $\rho(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$ et μ désigne le poids de la contrainte. On obtient alors pour cette contrainte :

$$M_{h_{rt}}^\rho = \mu \left(\mathbb{1}_{[0,R-12]}(r) \frac{\rho(h_{r,:}, h_{r+12,:})}{\|h_{r,:}\|_2^2} h_{rt} + \mathbb{1}_{[13,R]}(r) \frac{\rho(h_{r,:}, h_{r-12,:})}{\|h_{r,:}\|_2^2} h_{rt} \right)$$

et

$$P_{h_{r,t}}^\rho = \mu \left(\mathbb{1}_{[0,R-12]}(r) \frac{h_{r+12,t}}{\|h_{r,:}\|_2 \|h_{r+12,:}\|_2} + \mathbb{1}_{[13,R]}(r) \frac{h_{r-12,t}}{\|h_{r,:}\|_2 \|h_{r-12,:}\|_2} \right).$$

Contrainte de régularité spectrale sur les coefficients d'amplitude a_k des harmoniques : similaire à celle proposée dans [Virtanen, 2007] pour la régularité temporelle. On évite ainsi que les atomes soient trop irréguliers en terme d'enveloppe spectrale et ne modélisent des objets qu'ils ne sont pas censés modéliser (on peut par exemple penser au

cas où l'amplitude des harmoniques impairs serait nulle ce qui n'est pas souhaitable). Le terme de coût associé à cette contrainte est :

$$\mathcal{C}_{\text{reg}}(\mathbf{a}) = \frac{\mu}{\|\mathbf{a}\|_2^2} \sum_{k=2}^{n_h} (a_k - a_{k-1})^2,$$

où μ désigne le poids de la contrainte. On obtient alors pour cette contrainte :

$$M_{a_k}^{\text{reg}} = \mu \left(\frac{2}{\|\mathbf{a}\|_2^2} \mathbb{1}_{[1, n_h-1]}(k)(a_k + a_{k-1}) + \frac{2a_k}{\|\mathbf{a}\|_2^4} \sum_{k=1}^{n_h} (a_k - a_{k-1})^2 \right)$$

et

$$P_{a_k}^{\text{reg}} = 2 \frac{\mu}{\|\mathbf{a}\|_2^2} \mathbb{1}_{[2, n_h]}(k)(a_k + a_{k+1}).$$

Nous avons observé que les contraintes de décorrélation et de régularité spectrale pouvaient améliorer la décomposition fournie de façon significative.

4.1.2.6 Détails de l'implémentation

La méthode complète est détaillée dans l'Algorithme 4.1 page 93. Les contraintes présentées dans la section 4.1.2.5 n'y ont pas été introduites, mais il est facile de les prendre en compte : il suffit de modifier les règles de mise à jour (4.14) et (4.10) comme expliqué dans la section 4.1.2.5.

4.1.3 Exemple

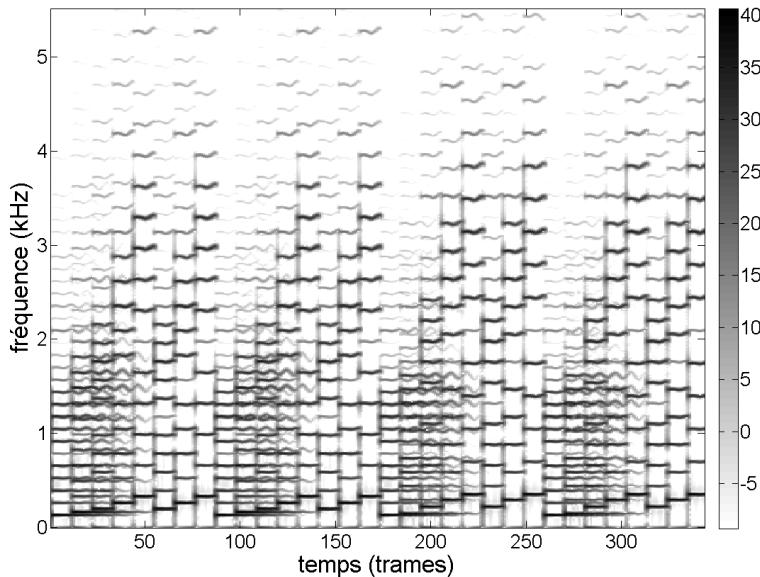


FIG. 4.4 – Spectrogramme original de l'extrait du premier prélude de Bach.

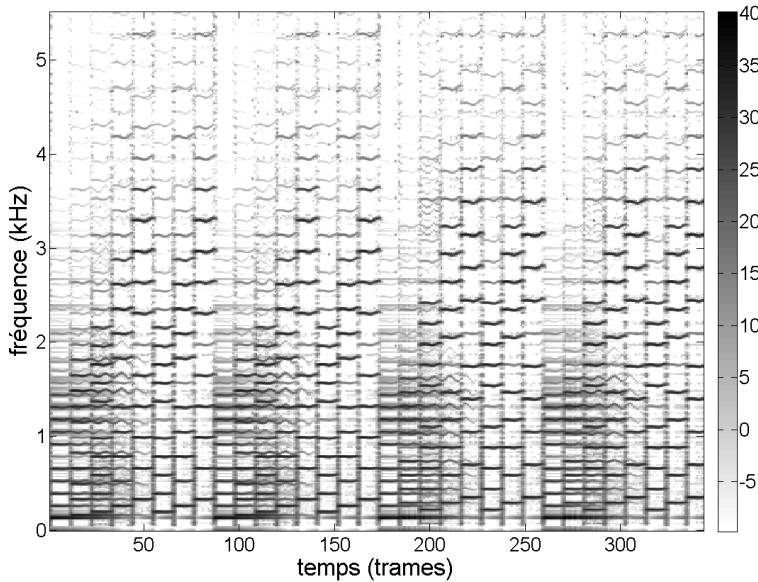


FIG. 4.5 – Spectrogramme reconstruit de l'extrait du premier prélude de Bach.

La figure 4.6 représente les activations h_{rt} de la décomposition du spectrogramme de puissance (représenté dans la figure 4.4) d'un extrait (quatre premières mesures) du premier prélude de Jean-Sébastien Bach joué par un synthétiseur. La fréquence d'échantillonnage de l'extrait est $f_s = 11025\text{Hz}$. Une fenêtre de Hamming de 1024 échantillons (93ms) avec 75% de recouvrement a été utilisée pour le calcul de la **TFCT**. La décomposition comporte 72 atomes (ce qui correspond à 6 octaves), les fréquences fondamentales se répartissant tous les demi-tons de 55Hz (*La0*) à 3322Hz (*Sol#6*). La divergence de Kullback-Leibler (**KL**) (β -divergence avec $\beta = 1$) a été utilisée. Dans un premier temps aucun terme de contrainte n'a été ajouté, et aucun atome de **NMF** (atome non harmonique) n'a été ajouté dans la décomposition. Toutes les notes ont été jouées par un synthétiseur avec un léger vibrato. La polyphonie maximum de ce morceau est de 3 notes.

Dans la figure 4.6, les notes du morceau apparaissent clairement avec d'importantes activations. On voit cependant apparaître de nombreuses répliques aux niveau des octaves, douzième... des notes jouées. De plus, au moment des attaques, de nombreux atomes sont actifs simultanément du fait qu'aucun atome harmonique n'est capable de représenter correctement une attaque.

Afin de réduire ce problème, la décomposition est recalculée en ajoutant un terme de contrainte de décorrélation et un terme de contrainte de régularité spectrale (*cf. section 4.1.2.5*) ainsi qu'un atome de **NMF** standard (atome non-paramétrique) afin de mieux prendre en compte les attaques (*cf. section 4.1.1.4*). Les activations obtenues avec cette nouvelle décomposition sont représentées dans la figure 4.7.

On peut constater que les répliques ont alors quasiment disparu ce qui montre que la contrainte de décorrélation a fait effet. De plus, le phénomène d'activation simultanée de plusieurs atomes harmoniques au moment des attaques a diminué, ce qui montre que l'atome de **NMF** standard ajouté peut modéliser les attaques.

Le spectrogramme reconstruit est alors représenté dans la figure 4.5 (avec ou sans contraintes, le spectrogramme reconstruit a une apparence très similaire, mais les éléments de la décomposition ont plus de sens en incluant des contraintes).

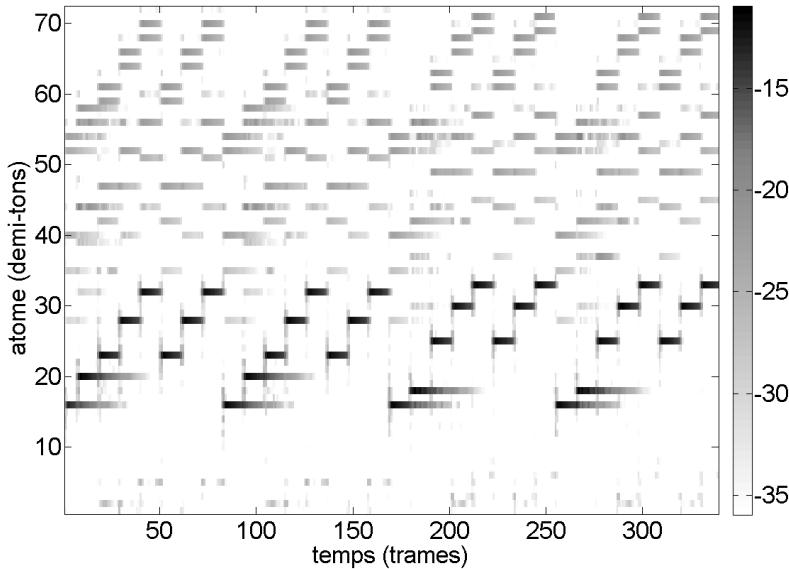


FIG. 4.6 – Activations de la décomposition du spectrogramme de l'extrait du premier prélude de Bach sans contraintes et sans atomes de NMF standard. L'échelle de couleurs est en dB.

Afin de visualiser les fréquences fondamentales estimées, on peut facilement construire une représentation temps/fréquence des activations qui fait apparaître ces fréquences fondamentales variables : pour chaque atome r , à chaque instant t , un pic étroit est généré dans le plan temps/fréquence au point (t, f_0^{rt}) avec une amplitude égale à l'activation h_{rt} (ce qui revient en fait à représenter les activations en les centrant sur la fréquence fondamentale estimée à chaque instant). Une telle représentation est donnée dans la figure 4.8 pour le même extrait. Cette représentation fait apparaître le vibrato généré par le synthétiseur.

L'exemple précédent était généré par un synthétiseur et donc particulièrement adapté à la décomposition proposée, notamment car le type de synthétiseur utilisé vérifie à peu près le modèle de transposition simplifiée utilisée (toutes les notes ont le même jeu d'amplitudes des harmoniques).

Lorsqu'on cherche à décomposer des spectrogrammes générés à partir d'instruments réels, la représentation fournie peut être moins claire et le réglage des contraintes pour diminuer les ambiguïtés d'octaves est particulièrement complexe. Ainsi pour le même extrait joué par un piano, on obtient les activations représentées dans la figure 4.9. Les difficultés rencontrées sont liées à plusieurs facteurs :

- Le modèle de transposition utilisé est loin d'être réaliste dans le cas du piano.
- Les notes du piano peuvent présenter une certaine inharmonicité qui n'est pas du tout considérée dans le modèle
- Les notes du piano contiennent des partiels qui ne font pas partie de la série (quasi)-

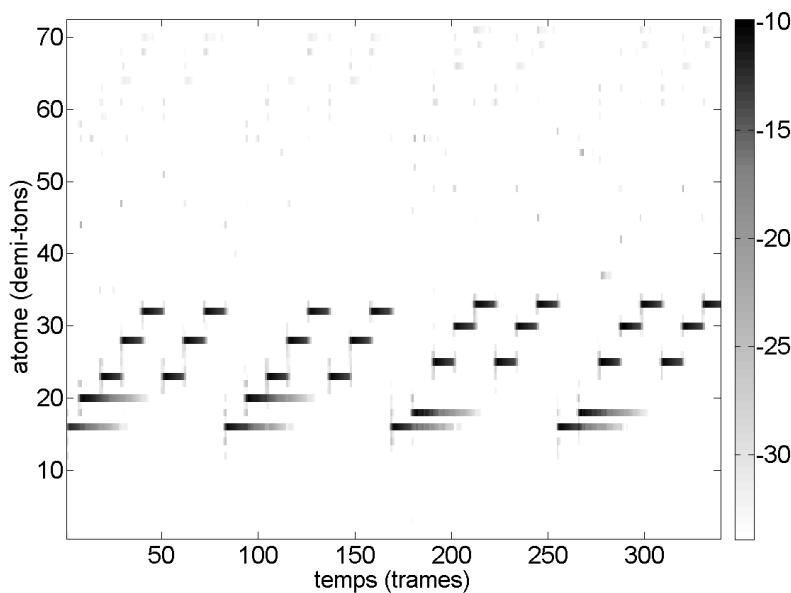


FIG. 4.7 – Activations de la décomposition du spectrogramme de l'extrait du premier prélude de Bach avec contrainte de décorrélation et de régularité spectrale et un atome non harmonique. L'échelle de couleur est en dB.

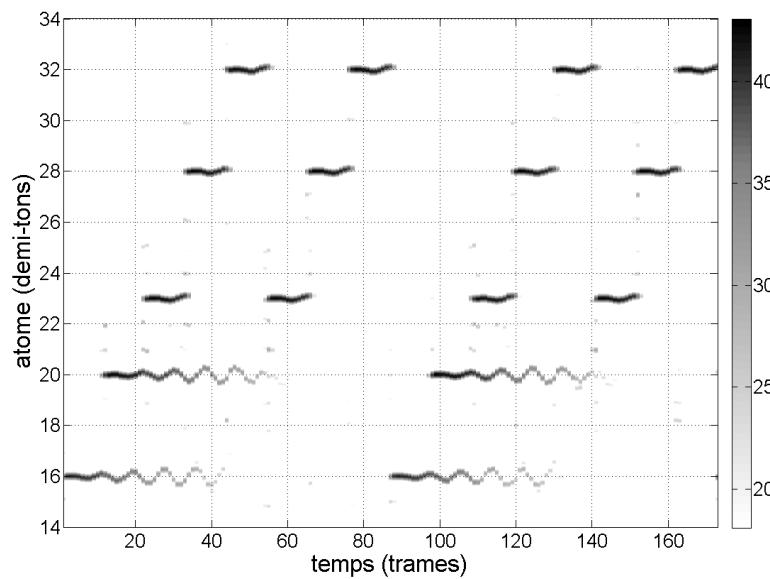


FIG. 4.8 – Représentation des activations incluant les fréquences fondamentales estimées (deux premières mesures du premier prélude de Bach).

harmonique.

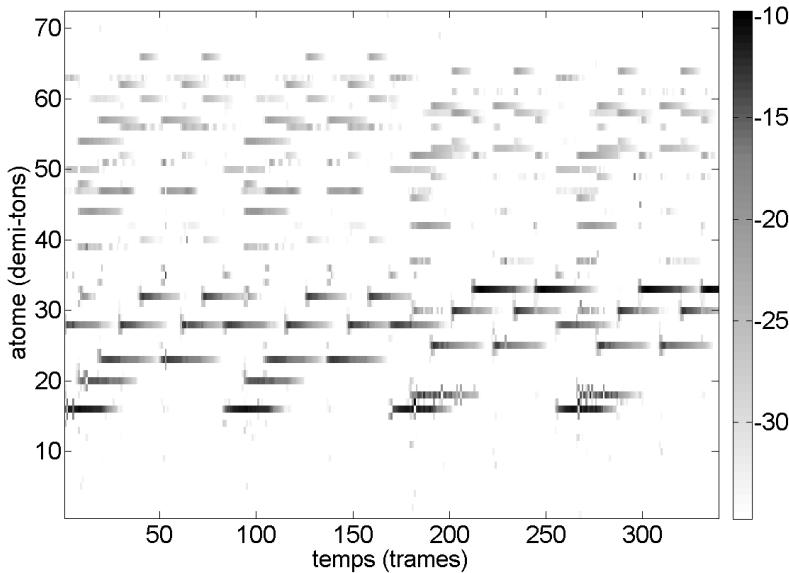


FIG. 4.9 – Activations h_{rt} de la décomposition du spectrogramme de l'extrait du premier prélude de Bach joué par un piano.

Remarque : L'ajout d'un paramètre d'inharmonicité semble nécessaire pour pouvoir décomposer correctement des sons de piano. Nous avons d'ailleurs tenter d'introduire dans le modèle un paramètre d'inharmonicité par atome mais sans succès : l'algorithme était en effet capable de déterminer l'inharmonicité pour une note isolée mais pour un signal polyphonique, les inharmonicités estimées étaient la plupart du temps erronées (ce qui s'explique sans doute par le fait que le paramètre d'inharmonicité est influencé en majeure partie par des décalages des partiels de haute fréquence et que ces partiels n'ont que peu d'énergie et se retrouvent généralement noyés dans une zone où la densité de partiels est importante). Il semble donc qu'il faille utiliser un modèle global d'inharmonicité sur toute la tessiture avec un petit nombre de paramètres (beaucoup moins qu'un paramètre par note) afin de pouvoir estimer correctement l'inharmonicité.

Il semble donc que ce type de décomposition souffre d'un problème de robustesse. Une des principales raison de ce manque de robustesse semble résider dans le fait que cette décomposition soit trop contrainte : la structure imposée aux atomes est en effet très forte.

Ce type de décomposition peut tout de même être utilisée dans le cadre de signaux réels comme le montre l'application de séparation de sources informée par la partition présentée dans la section 5.1 page 111.

4.1.4 Conclusion

Dans cette première partie de chapitre, nous avons présenté une nouvelle méthode de décomposition de spectrogrammes musicaux sur une base d'atomes harmoniques paramétriques correspondant chacun à une note de musique. Cette décomposition fournit une

bonne représentation des différentes notes jouées dans le spectrogramme. La décomposition étant paramétrique, elle est modulable et il est envisageable d'introduire de nouveaux paramètres dans le modèle comme un paramètre d'inharmonicité, des paramètres timbraux, des paramètres de variation rapide de la fréquence fondamentale (*chirp*).

La méthode de décomposition proposée présente cependant des problèmes de robustesse (pour décomposer des signaux réels) qui nous ont amené à réfléchir à des solutions alternatives. Ces problèmes semblent en grande partie liés à la très forte contrainte imposée aux atomes : l'utilisation d'atomes laissés totalement libres mais dont la structure harmonique est supposée à travers un modèle de transformation d'une note à l'autre a donc été envisagée. Ce nouveau modèle est présenté dans la section suivante.

Algorithme 4.1 : Décomposition paramétrique de spectrogrammes sur des atomes harmoniques variables

Données : \mathbf{V} (spectrogramme à décomposer), R (nombre d'atomes harmoniques), R' (nombre d'atomes non harmoniques), n_{iter} (nombre d'itérations), β (paramètre de la β -divergence)

Résultat : $\{f_0^{rt}\}_{r \in [1, R], t \in [1, T]}, \mathbf{H}, \mathbf{a}, \mathbf{W}', \mathbf{H}'$

Initialiser $\mathbf{H}, \mathbf{W}', \mathbf{H}'$ avec des valeurs aléatoires positives

Initialiser \mathbf{a} avec des 1

Initialiser f_0 avec les fréquences normalisées de la gamme chromatique :

$$f_0^{rt} = 2^{\frac{r-1}{12}} f_0^{\text{ref}}$$

pour $j = 1$ à n_{iter} **faire**

- Calculer l'atome paramétrique en suivant l'équation (4.2)
- Calculer $\hat{\mathbf{V}}$ en suivant l'équation (4.4)
- pour chaque** r et t **faire**

 - Mettre à jour f_0^{rt} en suivant l'équation (4.8)
 - si** $\left| 12 \log_2 \frac{f_0^{rt}}{f_0^{\text{ref}}} - (r-1) \right| > 1$ **alors**

 - | Imposer $h_{rt} = 0$

fin

fin

 Calculer l'atome paramétrique en suivant l'équation (4.2)

 Calculer $\hat{\mathbf{V}}$ en suivant l'équation (4.4)

pour chaque k **faire**

- |Mettre à jour a_k en suivant l'équation (4.14)

fin

 Calculer l'atome paramétrique en suivant l'équation (4.2)

 Calculer $\hat{\mathbf{V}}$ en suivant l'équation (4.4)

pour chaque r et t **faire**

- |Mettre à jour h_{rt} en suivant l'équation (4.10)

fin

 Calculer $\hat{\mathbf{V}}$ en suivant l'équation (4.4)

 Mettre à jour \mathbf{W}' en suivant l'équation (4.15)

 Calculer $\hat{\mathbf{V}}$ en suivant l'équation (4.4)

 Mettre à jour \mathbf{H}' en suivant l'équation (4.16)

fin

4.2 Transformation des atomes

Dans la décomposition présentée dans la section 4.1, les atomes peuvent être modifiés au cours du temps, via le paramètre de fréquence fondamentale. Cette transformation n'est possible qu'à travers la forme particulière donnée aux atomes paramétriques par rapport à la fréquence fondamentale. Cette contrainte est très forte et peut poser des problèmes pour modéliser des signaux réels comme le montre la section 4.1.3. Par conséquent, il semble plus pertinent de développer des méthodes dans lesquelles la transformation ne dépend pas d'une forme paramétrique de l'atome qui peut ainsi garder une forme totalement libre. Cette idée a déjà été exploitée par les méthodes de décomposition invariantes par translation fréquentielle pour des spectrogrammes à Q-constant.

Dans cette section, nous présentons une nouvelle méthode de décomposition de spectrogrammes également basée sur cette idée, qui vise à transposer les décompositions invariantes par translation fréquentielle adaptées aux spectrogrammes à Q-constant (avec une résolution fréquentielle logarithmique) à des spectrogrammes standard issus d'une TFCT (avec une résolution fréquentielle linéaire). Cette technique a l'avantage de permettre une reconstruction facile des signaux latents par filtrage de Wiener, ce qui peut être utilisé par exemple dans des applications de séparation de sources.

Les décompositions invariantes par translation fréquentielle [Schmidt et Mørup, 2006, Mysore et Smaragdis, 2009, Smaragdis *et al.*, 2008] permettent de décomposer les spectrogrammes à Q-constant [Brown, 1991] avec un unique atome fréquentiel pour chaque instrument harmonique : avec une résolution fréquentielle logarithmique, une translation le long de l'axe des fréquences correspond en effet à une transposition. Ainsi chaque note d'un instrument peut être modélisée par un motif de référence translaté à la bonne fréquence fondamentale.

Malheureusement, la Transformée à Q-Constant (TQC) est difficilement utilisable dans les applications de séparation de source : bien qu'une inversion quasi-parfaite de la TQC ait été récemment proposée [Schörkhuber et Klapuri, 2010], la résolution variable d'une décomposition de spectrogramme à Q-constant complique le masquage temps/fréquence. Des tentatives d'utilisation de décompositions invariantes par translation fréquentielle en séparation de source ont été proposées [Fitzgerald *et al.*, 2005] mais la TQC n'y est pas inversée : une transformation de la résolution logarithmique en résolution linéaire est utilisée afin d'éviter l'inversion de la TQC. La séparation de sources utilisant directement la TQC reste donc un problème encore largement ouvert. Dans ce chapitre, nous présentons une nouvelle méthode de décomposition inspirée par les décompositions invariantes par translation fréquentielle, mais qui permet de décomposer les spectrogrammes standard obtenus par TFCT. Dans un spectrogramme standard, un changement de la fréquence fondamentale peut être approximé par une homothétie du motif spectral. Cette approximation reste valide pour de petites modifications de la fréquence fondamentale puisque :

- pour des instruments acoustiques réels, la répartition des amplitudes des harmoniques n'est pas la même d'une note à l'autre. Même si cette hypothèse est couramment utilisée dans les décompositions invariantes par translation fréquentielle, elle n'est exacte que pour quelques instruments électroniques ;
 - les partiels ne sont pas des Diracs dans le domaine fréquentiel et ont donc une certaine largeur (donnée par la taille et le type de la fenêtre d'analyse) qui est la même pour tous les partiels, mais une homothétie va élargir (ou affiner) ces partiels.
-

La décomposition proposée est appelée *Analyse probabiliste en composantes latentes (PLCA) invariante par homothétie fréquentielle*. Le modèle invariant par homothétie pose également d'autres problèmes dont ne souffrait pas le modèle invariant par translation : une homothétie sur un ensemble d'entiers ne donne des nombres entiers que pour des valeurs entières du facteur d'homothétie. On va donc modéliser la variable fréquentielle par une variable continue, qui sera intégrée localement pour obtenir une variable discrète.

Dans la section 4.2.1, nous rappelons le principe de la PLCA invariante par translation fréquentielle. Nous présentons ensuite, dans la section 4.2.2, le nouveau modèle invariant par homothétie et proposons un algorithme pour estimer les paramètres du modèle. Plusieurs exemples sont présentés dans la section 4.2.4.

Le travail présenté dans ce chapitre a fait l'objet des articles [Hennequin *et al.*, 2011c,d]. Précédemment, nous avions également abordé l'idée d'opérer directement une transformation sur des motifs spectraux dans le cadre d'une tentative de définition d'une mesure de similarité spectrale dans [Hennequin *et al.*, 2010b].

4.2.1 PLCA invariante par translation fréquentielle

Le modèle proposé est très largement inspiré de la PLCA invariante par translation [Smaragdis *et al.*, 2008] qui est un modèle de tirage issu de la PLCA (*cf.* section 2.4.2 page 39). Dans les décompositions de type PLCA, un spectrogramme non négatif \mathbf{V} est considéré comme un histogramme issu du tirage structuré d'une variable aléatoire fréquentielle f et d'une variable aléatoire temporelle t qui suivent une loi jointe $P(f, t)$. Suivant la forme donnée à $P(f, t)$, il est possible d'obtenir différentes décompositions.

Le modèle PLCA invariant par translation fréquentielle introduit comme le modèle PLCA classique une variable cachée « composante » z :

$$P(f, t) = \sum_{z=1}^Z P(z)P(f, t|z).$$

Ce modèle introduit une variable aléatoire supplémentaire de « transposition » τ et une variable de fréquence « de base » f' . f' et t sont indépendantes connaissant la variable composante z , et τ est indépendante de f' mais pas de t (connaissant z). f est obtenue par translation du motif de base : $f = f' + \tau$. $P(f, t)$ est obtenue comme un produit de convolution :

$$P(f, t) = \sum_{z=1}^Z P(z) \sum_{f'=1}^{F'} P_K(f'|z) P_I(f - f', t|z).$$

Démonstration. On a :

$$P(f, t|z) = P(f|z, t)P(t|z).$$

Comme $f = f' + \tau$, et que f' et τ sont indépendantes conditionnellement à z , la loi de f conditionnellement à z est le produit de convolution des lois de f' et τ et on a :

$$P(f|z, t) = \sum_{f'=1}^{F'} P(f'|z, t)P(f - f'|z, t) = \sum_{f'} P_K(f'|z)P(f - f'|z, t).$$

Donc :

$$P(f, t|z) = \sum_{f'=1}^{F'} P_K(f'|z)P(f - f'|z, t)P(t|z) = \sum_{f'} P_K(f'|z)P_I(f - f', t|z).$$

□

P_K est appelée la distribution noyau (K de l'anglais *Kernel*) : elle correspond aux différents motifs spectraux qui sont translatés par la distribution d'impulsions P_I .

Afin d'illustrer ce type de décomposition, le spectrogramme à Q-constant représenté dans la figure 4.10 est décomposé à l'aide d'une **PLCA** invariante par translation fréquentielle avec un unique motif ($Z = 1$). Il s'agit du spectrogramme des premières notes d'*Au clair de la lune* jouées par un synthétiseur (cet exemple est donné à titre purement illustratif, et par conséquent l'utilisation d'un synthétiseur simple permet de respecter parfaitement les hypothèses du modèle de transposition). La figure 4.10 met en lumière le fait qu'une transposition dans ce type de spectrogramme est bien modélisée par une translation.

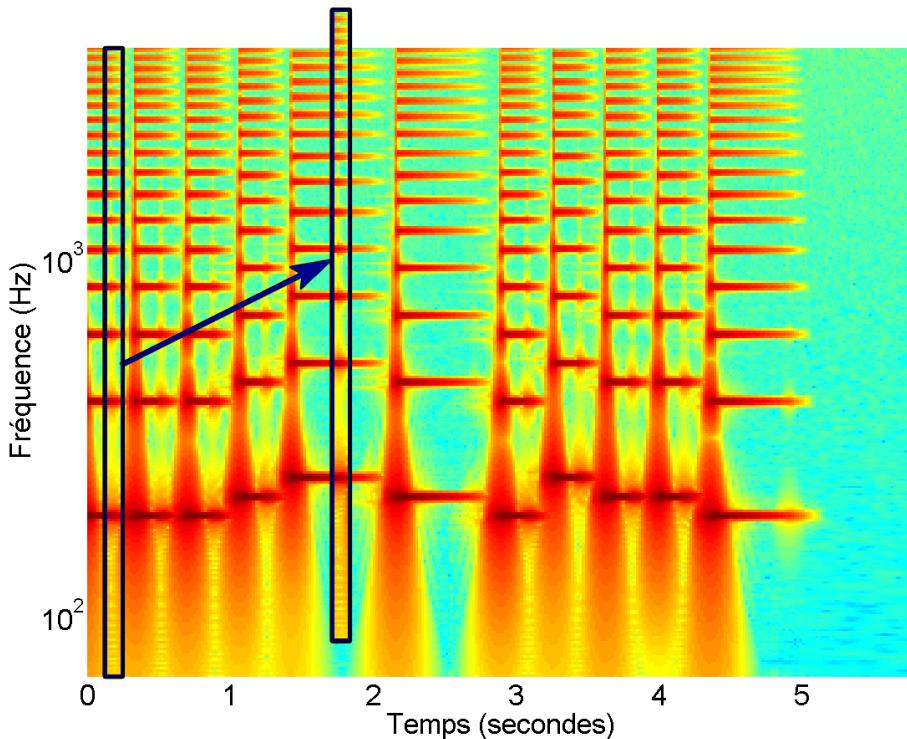


FIG. 4.10 – Spectrogramme à Q-constant des premières notes d'*Au clair de la lune* jouées par un synthétiseur. Illustration de l'équivalence translation/transposition.

Le motif fréquentiel extrait $P_K(:, 1)$ ainsi que la distribution d'impulsions $P_I(:, : | 1)$ obtenus sont représentés dans la figure 4.11 : on obtient bien un motif harmonique et la distribution d'impulsions décrit la ligne mélodique.

La figure 4.12 illustre la reconstruction du spectrogramme paramétrique par convolution du motif extrait avec la distribution d'impulsions.

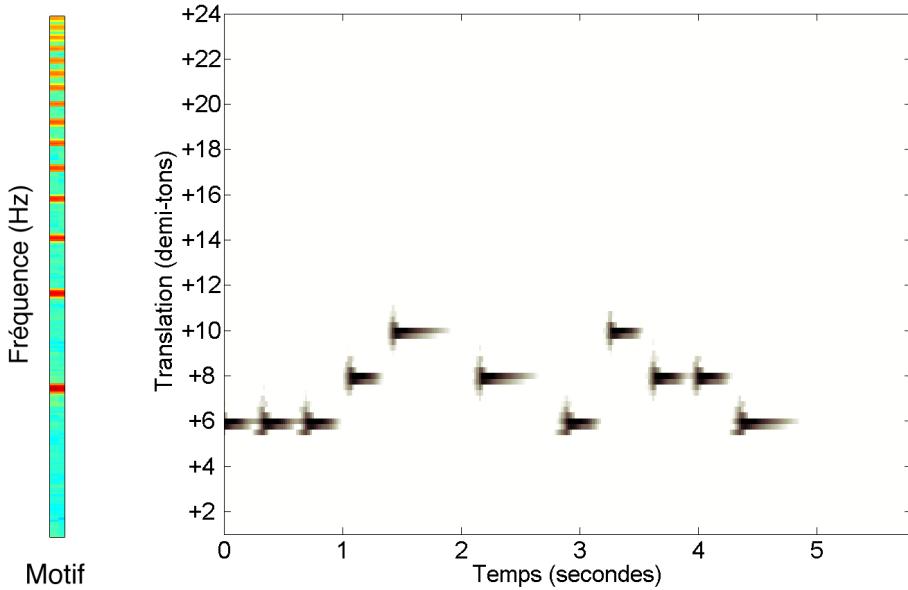


FIG. 4.11 – Décomposition obtenue par **PLCA** invariante par translation fréquentielle du spectrogramme de la figure 4.10 : à gauche, motif fréquentiel P_K extrait, à droite la distribution d’impulsions P_I .

4.2.2 Décomposition invariante par homothétie

Dans notre modèle (adapté aux spectrogrammes issus d’une **TFCT** et non plus aux spectrogrammes à Q-constant), la transposition n’est plus une translation mais est modélisée par une multiplication par un facteur scalaire $\lambda \in \mathbb{R}^+ \setminus \{0\}$, c’est-à-dire une homothétie du motif de base, comme l’illustre la figure 4.13. Il faut donc s’intéresser à la loi du produit de deux variables aléatoires indépendantes. Soit X une variable aléatoire discrète prenant ses valeurs dans $\{0, 1, 2, \dots, K\}$ et λ une variable aléatoire continue positive de densité p . La densité de λX est alors donnée par :

$$p_{\lambda X}(u) = \sum_{k=1}^K \frac{P(X = k)p(\frac{u}{k})}{k} + \delta(u)P(X = 0). \quad (4.17)$$

Dans notre modèle, on suppose que la variable aléatoire fréquentielle $f_c \in \mathbb{R}$ (le lien entre cette variable fréquentielle continue f_c et la variable discrète observée f sera clarifié plus tard) est obtenue en multipliant la fréquence de base $f' \in \{0, 1, \dots, F'\}$ (qui est indépendante de t conditionnellement à z) avec le facteur de transposition $\lambda \in \mathbb{R}^+ \setminus \{0\}$ (qui dépend de t mais pas de f' conditionnellement à z). En utilisant l’équation (4.17) avec $u = f_c$, $k = f'$ et $K = F'$, on obtient alors :

$$P(f_c, t | z) = \sum_{f'=1}^{F'} \frac{P_K(f'|z)}{f'} P_I\left(\frac{f_c}{f'}, t | z\right) + \delta(f_c)P_K(0|z).$$

Nous utilisons, comme dans la **PLCA** invariante par translation fréquentielle, la notation P_K pour la distribution noyau et P_I pour la distribution d’impulsions car ces distributions

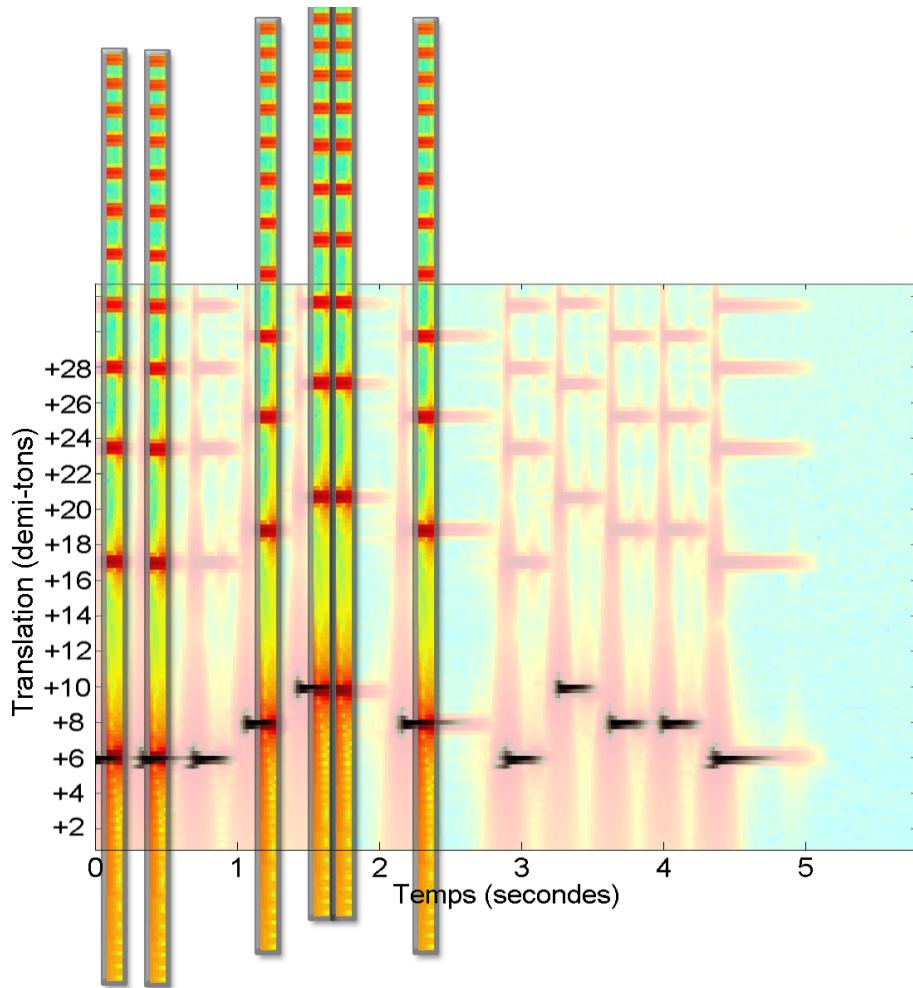


FIG. 4.12 – Reconstruction du spectrogramme de la figure 4.10 à partir de P_K et P_I .

jouent un rôle très comparable. Cependant, il est à noter qu'elles ne représentent pas tout à fait la même chose dans notre modèle.

Nous considérons ici que $P_K(0|z) = 0$ pour éviter la singularité de la fréquence nulle. Comme nous allons le voir plus tard, il peut toujours y avoir de l'énergie dans le canal fréquentiel 0 grâce à une homothétie de facteur inférieur à 1 du motif fréquentiel.

On obtient alors :

$$P(f_c, t) = \sum_{z=1}^Z P(z) \sum_{f'=1}^{F'} \frac{P_K(f'|z)}{f'} P_I\left(\frac{f_c}{f'}, t|z\right).$$

La variable aléatoire f_c est continue, mais la variable observée $f \in \{0, 1, \dots, F\}$ est discrète. On va donc supposer que $f = [f_c]$ et par conséquent :

$$P(f, t) = \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P(f_c, t) df_c.$$

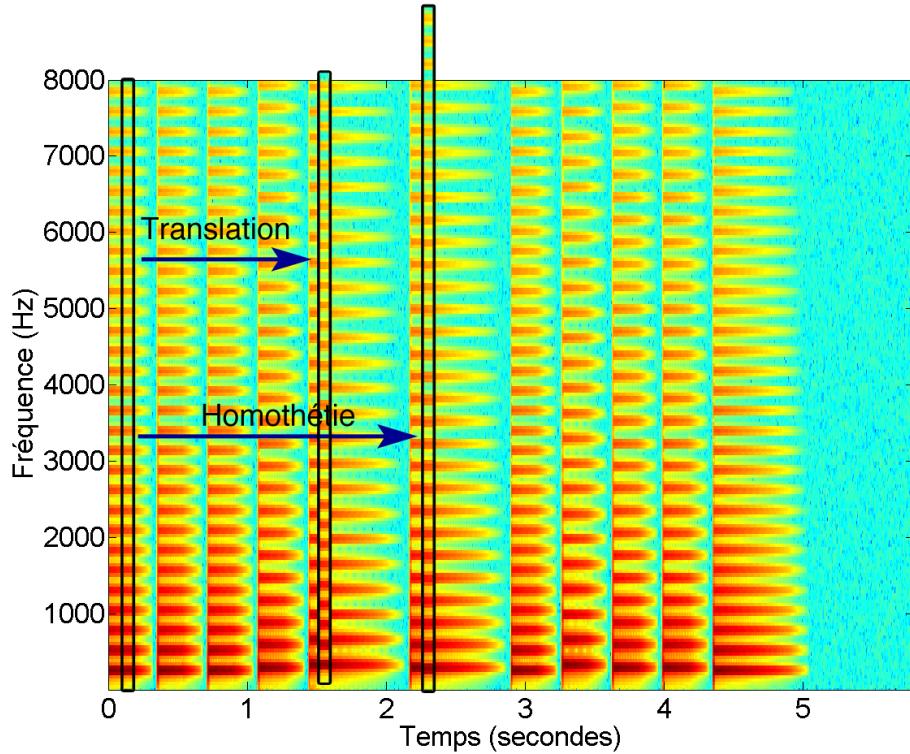


FIG. 4.13 – Spectrogramme classique (issu d'une **TFCT**) des premières notes d'*Au clair de la lune* jouées par un synthétiseur. Illustration de l'équivalence homothétie/transposition dans un tel spectrogramme (le motif translaté ne s'adapte pas correctement).

Par conséquent, $\forall f \in \{0, 1, \dots, F\}, \forall t \in \{1, \dots, T\}$:

$$P(f, t) = \sum_{z, f'} \frac{P(z) P_K(f'|z)}{f'} \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I \left(\frac{f_c}{f'}, t | z \right) df_c.$$

Le jeu de paramètres à estimer est donc : $\theta = \{P(z), P_K(f'|z), P_I(\lambda, t|z)\}$.

En pratique, il faut « discréteriser » P_I (qui est continue par rapport à λ) pour pouvoir estimer θ . On va ici paramétriser P_I en supposant que $\lambda \mapsto P_I(\lambda, t|z)$ est constante par morceaux pour tout t et z . On se donne une famille $\{\lambda_k\}_{k \in \{1, \dots, K\}}$ qui ne dépend pas de t et de z . On choisit ici $\lambda_k = 2^{\frac{k-k_0}{12n_{st}}}$: cette discréétisation exponentielle est en effet adaptée pour décrire des transpositions sur une échelle musicale correspondant à des subdivisions du ton (n_{st} correspond au nombre de subdivisions de chaque demi-ton). On suppose alors que P_I vérifie :

$$\forall \lambda \in [\lambda_k 2^{-\frac{1}{24n_{st}}}, \lambda_k 2^{\frac{1}{24n_{st}}}], \quad P_I(\lambda, t|z) = P_I(\lambda_k, t|z).$$

On suppose de plus que P_I est nulle en dehors de ces intervalles. Les valeurs de $P_I(\lambda_k, t|z)$ (pour tout k , t et z) suffisent alors à déterminer entièrement P_I .

On a alors :

$$\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I \left(\frac{f_c}{f'}, t | z \right) df_c = f' \sum_{k=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_k, t | z) \delta \lambda_k^{f,f'},$$

où $k_{\min}^{f,f'}$ est choisi tel que $\lambda_{k_{\min}^{f,f'}} 2^{-\frac{1}{24n_{st}}} < \frac{f-\frac{1}{2}}{f'} \leq \lambda_{k_{\min}^{f,f'}} 2^{\frac{1}{24n_{st}}}$ et $k_{\max}^{f,f'}$ est choisi tel que $\lambda_{k_{\max}^{f,f'}} 2^{-\frac{1}{24n_{st}}} \leq \frac{f+\frac{1}{2}}{f'} < \lambda_{k_{\max}^{f,f'}} 2^{\frac{1}{24n_{st}}}$ (avec les contraintes que $1 \leq k_{\min}^{f,f'} \leq K$ et $1 \leq k_{\max}^{f,f'} \leq K$) :

$$\begin{aligned} k_{\min} &= \left[k_0 - \frac{1}{2} + 12n_{st} \log_2 \left(\frac{f - \frac{1}{2}}{f'} \right) \right], \\ k_{\max} &= \left[k_0 + \frac{1}{2} + 12n_{st} \log_2 \left(\frac{f + \frac{1}{2}}{f'} \right) \right]. \end{aligned}$$

$\delta \lambda_k^{f,f'}$ est donné par $\delta \lambda_k^{f,f'} = \min(\lambda_k 2^{\frac{1}{24n_{st}}}, \frac{f+\frac{1}{2}}{f'}) - \max(\lambda_k 2^{-\frac{1}{24n_{st}}}, \frac{f-\frac{1}{2}}{f'})$. Lorsque $\delta \lambda_k^{f,f'}$ n'est pas limité par les contraintes sur f et f' , on notera $\delta \lambda_k = \lambda_k 2^{\frac{1}{24n_{st}}} - \lambda_k 2^{-\frac{1}{24n_{st}}}$.

Le jeu de paramètres à estimer a été transformé en : $\theta = \{P(z), P_K(f'|z), P_I(\lambda_k, t|z) | z \in \{1, \dots, Z\}, f' \in \{1, \dots, F'\}, k \in \{1, \dots, K\}, t \in \{1, \dots, T\}\}$.

Remarque : Il aurait aussi été envisageable de considérer une paramétrisation différente pour P_I par exemple une paramétrisation affine par morceaux de $\lambda \mapsto P_I(\lambda, t|z)$, ce qui permettrait de garder exactement le même jeu de paramètres mais en ayant une expression différente de $P_I(\lambda, t|z)$ en fonction des $P_I(\lambda_k, t|z)$. Pour rester dans un cadre général, la paramétrisation de P_I n'apparaîtra qu'à la fin des calculs, dans la recherche des règles de mise à jour.

4.2.3 Algorithme Espérance-Maximisation (EM)

Nous cherchons à estimer la valeur du paramètre θ qui maximise la log-vraisemblance des observations v_{ft} :

$$L((\bar{f}, \bar{t})|\theta) = \sum_{i \in I} \log P(f_i, t_i). \quad (4.18)$$

\bar{f} et \bar{t} correspondent à l'ensemble des tirages effectués sur f et t (les tirages étant indexés par $i \in I = \{1, \dots, N\}$ où N est le nombre total de tirages de f et t).

Comme le nombre de tirages qui ont conduit à la valeur (f, t) est v_{ft} , la log-vraisemblance peut être réécrite :

$$L((\bar{f}, \bar{t})|\theta) = \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log P(f, t).$$

On va chercher à maximiser cette log-vraisemblance à l'aide de l'algorithme Espérance-Maximisation (EM) avec pour variables latentes z et f' (il serait équivalent de considérer comme variables latentes z et λ puisque $f = \lambda f'$).

La log-vraisemblance complétée s'écrit :

$$\begin{aligned} L((\bar{f}, \bar{t}, z, f')|\theta) &= \sum_{i \in I} \log P(f_i, t_i, z, f') \\ &= \sum_{f=1}^F \sum_{t=1}^T v_{ft} \log P(f, t, z, f'). \end{aligned}$$

Or,

$$P(f, t, z, f') = \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P(f_c, t, z, f') df_c$$

et

$$P(f_c, t, z, f') = P(z) \frac{P_K(f'|z)}{f'} P_I(\frac{f_c}{f'}, t|z),$$

donc :

$$P(f, t, z, f') = P(z) \frac{P_K(f'|z)}{f'} \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I(\frac{f_c}{f'}, t|z) df_c.$$

On en déduit :

$$L((\bar{f}, \bar{t}, z, f')|\theta) = \sum_{f,t} v_{ft} \left\{ \log P(z) + \log P_K(f'|z) + \log \left(\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I(\frac{f_c}{f'}, t|z) df_c \right) \right\} + c,$$

où c est une constante qui ne dépend pas de θ .

L'espérance de la log-vraisemblance complétée par rapport à $z, f' | f, t, \theta^{(c)}$ (où $\theta^{(c)}$ est la valeur du paramètre courant) est donc :

$$\begin{aligned} Q(\theta | \theta^{(c)}) &= \mathbb{E}_{z, f' | f, t, \theta^{(c)}} (L((\bar{f}, \bar{t}, z, f')|\theta)) \\ &= \sum_{f', z, f, t} v_{ft} P(z, f' | f, t, \theta^{(c)}) \left\{ \log P(z) + \log P_K(f'|z) \right. \\ &\quad \left. + \log \left(\int P_I(\frac{f_c}{f'}, t|z) df_c \right) \right\} + c. \end{aligned} \quad (4.19)$$

On obtient une expression de $P(z, f' | f, t, \theta^{(c)})$ en fonction de $\theta^{(c)}$ grâce au théorème de Bayes (étape E) :

$$P(z, f' | f, t, \theta^{(c)}) = \frac{P(f, t, f' | z) P^{(c)}(z)}{P^{(c)}(f, t)} = \frac{P_K^{(c)}(f' | z) P^{(c)}(z) \int P_I^{(c)}(\frac{f_c}{f'}, t|z) df_c}{f' P^{(c)}(f, t)}. \quad (4.20)$$

L'exposant $(.)^{(c)}$ indique que les quantités sont calculées à partir du paramètre courant : $\theta^{(c)} = \{P^{(c)}(z), P_K^{(c)}(f' | z), P_I^{(c)}(\lambda, t | z)\}$.

L'espérance complétée (4.19) doit être maximisée (étape M) par rapport à θ ($\theta^{(c)}$ restant fixe). Les probabilités constituant θ ($P(z)$, $P_K(f | z)$ et $P_I(\lambda, t | z)$) doivent se sommer à 1, par

conséquent la maximisation est contrainte. On utilise donc les multiplicateurs de Lagrange et on obtient le lagrangien ³ :

$$H(\theta|\theta^{(c)}) = \underbrace{\sum_{f,z,f',t} v_{ft} P(z, f'|f, t, \theta^{(c)}) \left\{ \log P(z) + \log P_K(f'|z) + \log \left(\int P_I\left(\frac{f_c}{f'}, t|z\right) df_c \right) \right\}}_{\text{Log-vraisemblance}} + \underbrace{\mu \left(1 - \sum_z P(z) \right) + \sum_z \rho_z \left(1 - \sum_{f'} P_K(f'|z) \right) + \sum_z \tau_z \left(1 - \sum_t \int_{\lambda_{\min}}^{\lambda_{\max}} P_I(\lambda, t|z) d\lambda \right)}_{\text{Contraintes}}, \quad (4.21)$$

où μ , ρ_z et τ_z sont des multiplicateurs de Lagrange.

4.2.3.1 Mise à jour de $P(z)$:

La dérivée partielle de $H(\theta|\theta^{(c)})$ (définie dans l'équation (4.21)) par rapport à $P(z)$ s'écrit :

$$\frac{\partial H(\theta|\theta^{(c)})}{\partial P(z)} = \sum_{f,f',t} v_{ft} \frac{P(z, f'|f, t, \theta^{(c)})}{P(z)} - \mu.$$

On doit avoir $\frac{\partial H(\theta|\theta^{(c)})}{\partial P(z)} = 0$, donc :

$$\sum_{f,f',t} v_{ft} P(z, f'|f, t, \theta^{(c)}) - \mu P(z) = 0. \quad (4.22)$$

En sommant sur z , on obtient :

$$\mu = \sum_{z,f,f',t} v_{ft} P(z, f'|f, t, \theta^{(c)}) = \sum_{f,t} v_{ft}.$$

Puis, à partir de (4.22), on obtient immédiatement la règle de mise à jour pour $P(z)$:

$$P(z) \leftarrow \frac{\sum_{f,t} v_{ft} P(z|f, t, \theta^{(c)})}{\sum_{f,t} v_{ft}}. \quad (4.23)$$

4.2.3.2 Mise à jour de $P_K(f'|z)$:

On obtient de manière tout à fait analogue la règle de mise à jour de $P_K(f'|z)$:

$$P_K(f'|z) \leftarrow \frac{\sum_{f,t} v_{ft} P(z, f'|f, t, \theta^{(c)})}{\sum_{f,t} v_{ft} P(z|f, t, \theta^{(c)})}. \quad (4.24)$$

³Les contraintes de non-négativité n'ont pas été incluses dans le calcul par souci de lisibilité. Ces contraintes sont en fait inactives comme le montre la forme de la solution finale qui garantit bien la positivité des paramètres.

4.2.3.3 Mise à jour de $P_I(\lambda_k, t|z)$:

La mise à jour de $P_I(\lambda_k, t|z)$ pose plus de problèmes. La dérivée partielle du lagrangien par rapport à $P_I(\lambda_k, t|z)$ est en effet :

$$\begin{aligned} \frac{\partial H(\theta|\theta^{(c)})}{\partial P_I(\lambda_k, t|z)} &= \sum_{f,f'} v_{ft} P(z, f'|f, t, \theta^{(c)}) \frac{\partial \log \left(\int P_I(\frac{f_c}{f'}, t|z) df_c \right)}{\partial P_I(\lambda_k, t|z)} - \tau_z \frac{\partial \int_{\lambda_{\min}}^{\lambda_{\max}} P_I(\lambda, t|z)}{\partial P_I(\lambda_k, t|z)} \\ &= \sum_{f,f'} v_{ft} \frac{P(z, f'|f, t, \theta^{(c)})}{\int P_I(\frac{f_c}{f'}, t|z) df_c} \frac{\partial \left(\int P_I(\frac{f_c}{f'}, t|z) df_c \right)}{\partial P_I(\lambda_k, t|z)} - \tau_z \frac{\partial \int_{\lambda_{\min}}^{\lambda_{\max}} P_I(\lambda, t|z)}{\partial P_I(\lambda_k, t|z)}. \end{aligned}$$

En utilisant la paramétrisation de P_I proposée (fonction constante par morceaux) :

$$\frac{\partial H}{\partial P_I} = \sum_{f,f'} v_{ft} \frac{P(z, f'|f, t, \theta^{(c)}) \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k)}{\sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'}} \delta \lambda_k^{f,f'} - \tau_z \delta \lambda_k,$$

on doit donc avoir :

$$\sum_{f,f'} v_{ft} \frac{P(z, f'|f, t, \theta^{(c)}) \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k)}{\sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'}} \delta \lambda_k^{f,f'} - \tau_z \delta \lambda_k = 0.$$

En multipliant par $P_I(\lambda_k, t|z)$, on obtient :

$$\sum_{f,f'} v_{ft} P(z, f'|f, t, \theta^{(c)}) \frac{P_I(\lambda_k, t|z) \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k)}{\sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'}} \delta \lambda_k^{f,f'} - \tau_z \delta \lambda_k P_I(\lambda_k, t|z) = 0.$$

En sommant l'expression précédente sur k et t , on obtient :

$$\tau_z = \sum_{k,t,f,f'} v_{ft} P(z, f'|f, t, \theta^{(c)}) \frac{P_I(\lambda_k, t|z) \delta \lambda_k^{f,f'} \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k)}{\sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'}}.$$

On considère alors la règle de mise à jour au point fixe suivante (à itérer plusieurs fois), en espérant que cette règle converge vers un zéro de $\frac{\partial H}{\partial P_I(\lambda_k, t|z)}$:

$$P_I(\lambda_k, t|z) \leftarrow \frac{1}{\tau_z \delta \lambda_k} \sum_{f,f'} v_{ft} P(z, f'|f, t, \theta^{(c)}) \frac{P_I(\lambda_k, t|z) \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k)}{\sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'}}, \quad (4.25)$$

la division par τ_z étant en pratique une normalisation.

Nous n'avons pas réussi à démontrer la convergence de P_I après plusieurs itérations de la règle de mise à jour (4.25) vers un zéro de $\frac{\partial H}{\partial P_I}$. Cependant, nous l'avons toujours observée en pratique en quelques itérations.

Comme $Q(\theta|\theta^{(c)})$ (définie dans l'équation (4.19)) et les contraintes sont \mathcal{C}^1 , que $Q(\theta|\theta^{(c)})$ est strictement concave, que les contraintes sont affines, et que les conditions de régularité sont satisfaites, un point stationnaire de $H(\theta|\theta^{(c)})$ est nécessairement le maximum global de $Q(\theta|\theta^{(c)})$ sous les contraintes de normalisation. Par conséquent, $Q(\theta|\theta^{(c)})$ est bien maximisé à chaque itération en utilisant les règles de mise à jour (4.23), (4.24) et (4.25) et l'algorithme Espérance/Maximisation (EM) fait bien croître la vraisemblance à chaque itération.

4.2.3.4 Mises à jour multiplicatives

Les mises à jour (4.23) et (4.24) peuvent être réécrites sous forme multiplicative, en utilisant l'expression (4.20) et en remplaçant le calcul du dénominateur de chacune des règles par une normalisation.

Mise à jour de $P(z)$:

$$\left\{ \begin{array}{l} P^{(i)}(z) \leftarrow P^{(c)}(z) \sum_{f,t,f'} \frac{v_{ft}}{P^{(c)}(f,t)} P_K^{(c)}(f'|z) \sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'} \\ P(z) \leftarrow \frac{P^{(i)}(z)}{\sum_{z'} P^{(i)}(z')} \quad (\text{normalisation}). \end{array} \right.$$

L'exposant $(.)^{(i)}$ indique une quantité intermédiaire dans le calcul (paramètre non normalisé).

Mise à jour de $P_K(f|z)$:

$$\left\{ \begin{array}{l} P_K^{(i)}(f'|z) \leftarrow P_K^{(c)}(f'|z) \sum_{f,t} \frac{v_{ft}}{P^{(c)}(f,t)} P^{(c)}(z) \sum_{k'=k_{\min}^{f,f'}}^{k_{\max}^{f,f'}} P_I(\lambda_{k'}, t|z) \delta \lambda_{k'}^{f,f'} \\ P_K(f'|z) \leftarrow \frac{P_K^{(i)}(f'|z)}{\sum_{f''} P_K^{(i)}(f''|z)} \quad (\text{normalisation}). \end{array} \right.$$

Mise à jour de $P_I(\lambda_k, t|z)$: On peut également réécrire la mise à jour (4.25) de P_I sous forme multiplicative :

$$\left\{ \begin{array}{l} P_I^{(i)}(\lambda_k, t|z) \leftarrow P_I(\lambda_k, t|z) P^{(c)}(z) \sum_{f'} P_K^{(c)}(f'|z) \sum_f \frac{v_{ft} \int P_I^{(c)}(\lambda, t|z) d\lambda \delta \lambda_k^{f,f'}}{P^{(c)}(f,t) \int P_I(\lambda, t|z) d\lambda} \mathbb{1}_{[k_{\min}^{f,f'}, k_{\max}^{f,f'}]}(k) \\ P_I(\lambda_k, t|z) \leftarrow \frac{P_I^{(i)}(\lambda_k, t|z)}{\delta \lambda_k \sum_{k',t'} P_I^{(i)}(\lambda_{k'}, t'|z)} \quad (\text{normalisation}). \end{array} \right.$$

avec :

$$\int P_I^{(c)}(\lambda, t|z)d\lambda = \sum_{k'} P_I^{(c)}(\lambda_{k'}, t|z)\delta\lambda_{k'}^{f,f'} \quad \text{et} \quad \int P_I(\lambda, t|z)d\lambda = \sum_{k'} P_I(\lambda_{k'}, t|z)\delta\lambda_{k'}^{f,f'}.$$

4.2.3.5 Complexité algorithmique

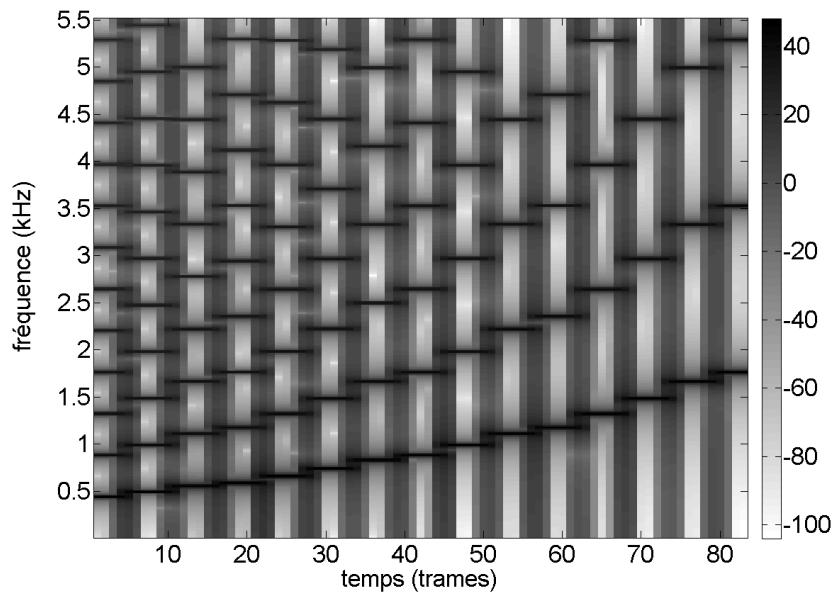
Le principal défaut de l'algorithme proposé dans cette section est son importante complexité algorithmique : contrairement à la **PLCA** invariante par translation fréquentielle, le calcul du spectrogramme paramétrique $P(f, t)$ ne peut être réalisé à l'aide d'un algorithme de convolution rapide. Ainsi le temps de calcul de la décomposition est assez important : de cent à mille fois le temps réel (ce facteur dépend de la taille choisie pour les paramètres, notamment de Z et de la finesse du pas de discréétisation de P_I) sur un ordinateur de bureau récent bi-processeur. Il est cependant possible de réduire considérablement ce temps de calcul en utilisant une **PLCA** invariante par translation pour initialiser l'algorithme : les distributions d'impulsions fournies par les deux algorithmes (invariant par translation et invariant par homothétie) sont en effet généralement très semblables et il est possible de transformer le motif P_K de la résolution logarithmique à la résolution linéaire. On dispose alors d'une « solution » grossière du modèle invariant par translation et la convergence de l'algorithme vers une solution fine ne se fait alors qu'en quelques itérations.

4.2.4 Exemples

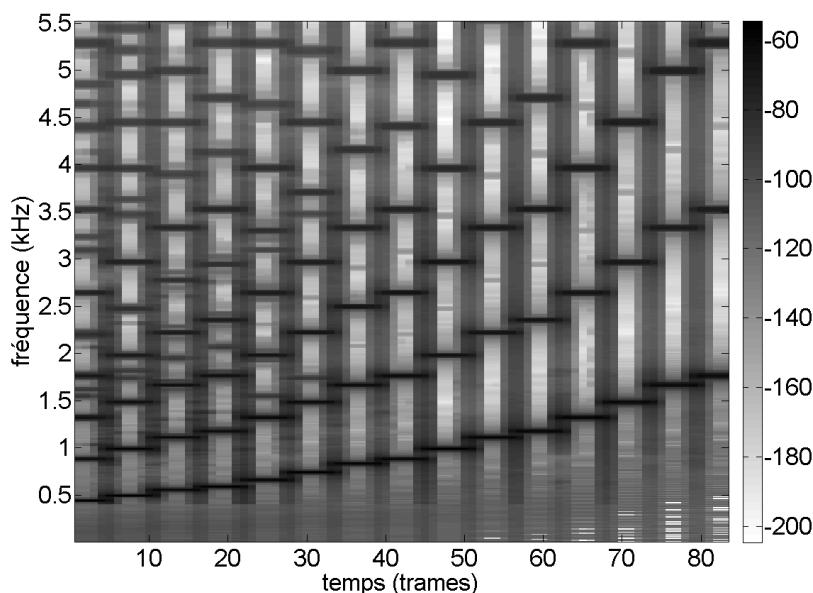
4.2.4.1 Exemple synthétique

Dans cette section, nous utilisons notre algorithme pour décomposer un spectrogramme synthétique très simple. Ce spectrogramme est obtenu par **TFCT** (fenêtre de Hann de 1024 échantillons, soit 93ms, avec 75% de recouvrement) d'une gamme de *La* majeur jouée sur 2 octaves (de *La4* à *La6*) par un synthétiseur, le son étant échantillonné à 11025Hz. Le spectrogramme original est représenté dans la figure 4.14(a). Le spectrogramme reconstruit à partir de la décomposition est présenté dans la figure 4.14(b) : on constate que le spectrogramme reconstruit est bien très similaire au spectrogramme original. La différence d'amplitude maximum entre le spectrogramme original et le spectrogramme reconstruit est due à la normalisation de $P(f, t)$ (le spectrogramme **V** n'est pas normalisé), mais la dynamique reste la même dans les deux spectrogrammes. Les harmoniques de haute fréquence du spectrogramme reconstruit sont légèrement plus larges que ceux du spectrogramme original.

La distribution noyau P_K obtenue est représentée dans la figure 4.16 : on constate que le motif factorisé est bien harmonique. On remarque également que les amplitudes deviennent très faibles pour les indices fréquentiels f grand. Ce phénomène est principalement lié au fait que notre modèle considère que les valeurs du spectrogramme hors de la zone de fréquences observée sont nulles alors que le spectrogramme modèle $P(f, t)$ peut prendre des valeurs non nulles hors de cette zone. On peut résoudre ce problème en utilisant l'approche décrite dans [Smaragdis *et al.*, 2010, 2009]. En pratique, pour des signaux d'instruments acoustiques réels, ce n'est pas un vrai problème, puisque le contenu harmonique de tels signaux est en grande partie noyé dans le bruit à partir de 10000Hz. Ainsi, un taux d'échantillonnage de 44100Hz ou plus permet de limiter fortement ce phénomène.



(a) Spectrogramme original



(b) Spectrogramme reconstruit

FIG. 4.14 – Spectrogramme original et spectrogramme reconstruit.

La distribution d’impulsions P_I obtenue est représentée dans la figure 4.15 : cette distribution prend clairement des valeurs très importantes à la position relative des notes effectivement jouées. On observe également quelques répliques des notes aux positions de contenu harmonique similaire (octave, douzième, double octave...). Au moment des attaques, P_I

prend des valeurs importantes pour de nombreuses valeurs du facteur d'homothétie λ : ce phénomène est dû à la forme plate du spectre des attaques qui est ici représenté par plusieurs motifs harmoniques dilatés ou compressés.

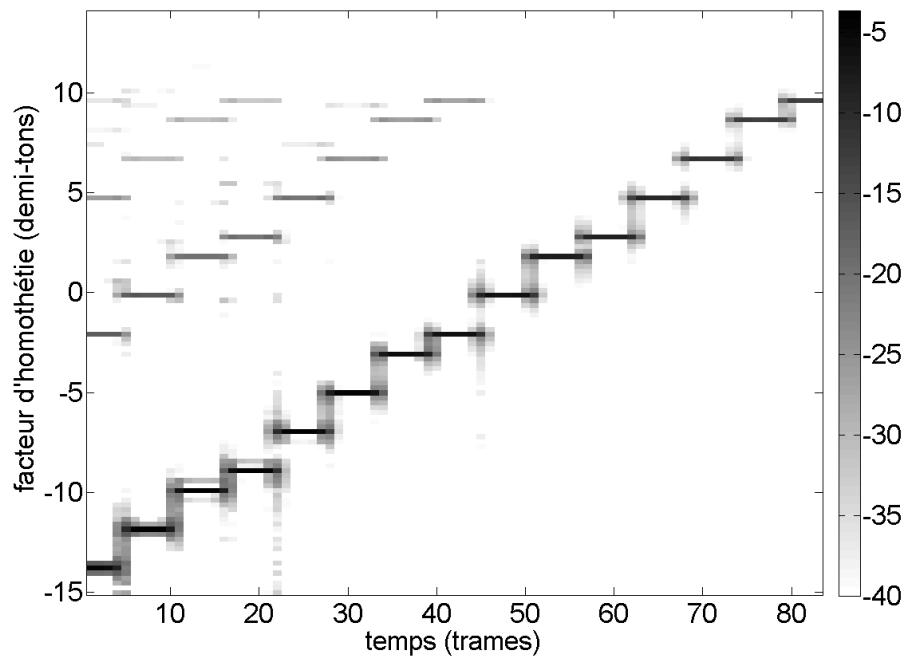


FIG. 4.15 – Distribution d’impulsions P_I de la gamme de *La* majeur.

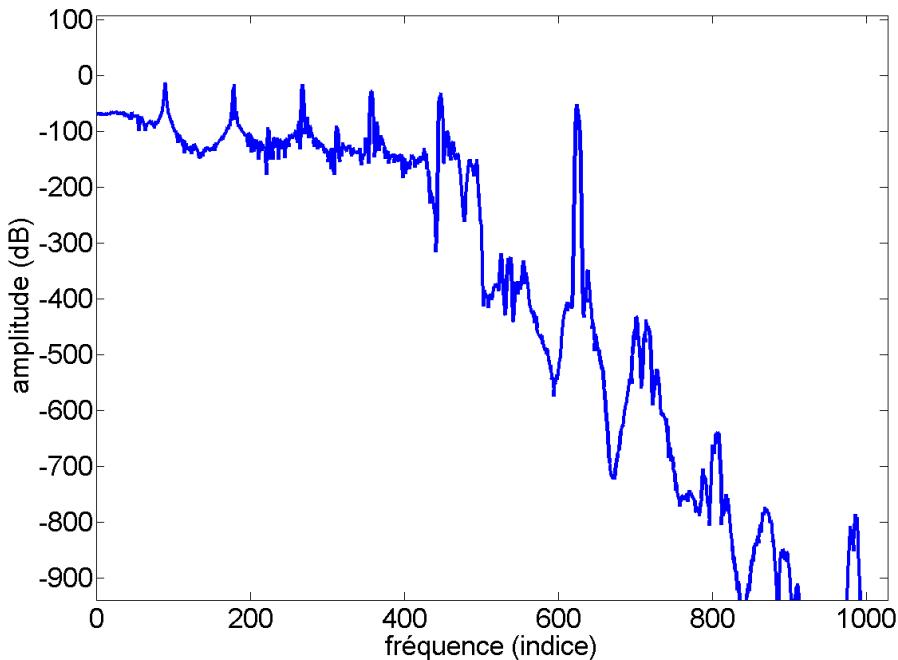


FIG. 4.16 – Distribution noyau P_K de la gamme de *La* majeur.

4.2.4.2 Enregistrement réel

Dans cette section, nous présentons la décomposition du spectrogramme des 10 premières secondes de la chanson *Because* des Beatles avec un unique motif ($Z = 1$). Le signal décomposé est une introduction polyphonique au clavecin enregistrée dans des conditions réelles. Le signal original stéréo a été transformé en un signal mono (en sommant les deux canaux) et sous-échantillonné de 44100Hz à 22050Hz. Le spectrogramme a été calculé à l'aide d'une **TFCT** en utilisant une fenêtre de Hann de 2048 échantillons et un recouvrement de 75%.

La distribution d'impulsions P_I obtenue est représentée dans la figure 4.17 : les notes réellement jouées sont matérialisées par des rectangles rouges dans la figure. On constate que dans tous les rectangles, P_I prend des valeurs importantes.

La distribution d'impulsions P_I obtenue est très similaire à la distribution d'impulsions qui peut être obtenue avec une décomposition invariante par translation fréquentielle. Cependant, comme notre décomposition est calculée sur un spectrogramme à résolution fréquentielle linéaire, celle-ci permet des applications de séparation des éléments constituants comme par exemple l'application décrite dans la section 5.2 page 122.

4.2.5 Conclusion

Nous avons proposé dans cette section une nouvelle méthode de décomposition des spectrogrammes musicaux. Cette décomposition est basée sur un petit nombre d'atomes fréquentiels qui peuvent être transformés à chaque instant : cette transformation, qui est

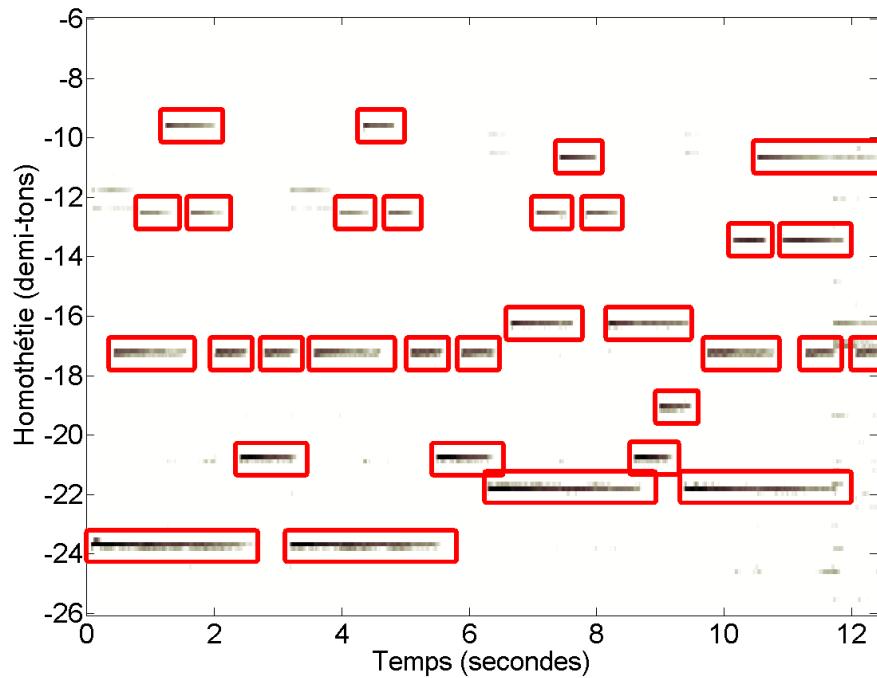


FIG. 4.17 – Distribution d’impulsions P_I de l’introduction de *Because*.

dans ce travail une homothétie, est censée modéliser une modification de fréquence fondamentale. Ainsi un unique motif peut modéliser toutes les notes d’un même instrument ainsi que des variations de fréquence fondamentale au sein d’une note. Nous verrons dans le chapitre 5 que ce type de décomposition peut permettre de manipuler les notes de façon indépendante dans un mélange polyphonique grâce à la représentation mi-niveau qu’elle fournit.

Chapitre 5

Applications et fusion des modèles

Dans ce chapitre, nous présentons des applications des décompositions présentées dans les chapitres précédents. La section 5.1 est consacrée à une application de séparation de source informée par la partition issue de la décomposition présentée dans la section 4.1. La section 5.2 décrit une méthode de modification de notes isolées dans un mélange polyphonique en utilisant la décomposition proposée dans la section 4.2. Enfin la section 5.3 présente une fusion des modèles présentés respectivement dans le chapitre 3 et dans la section 4.1 : ce type de décomposition hybride est présenté à titre indicatif car il ne permet pas, dans l'état actuel, la décomposition de signaux complexes. Il pourrait permettre de modéliser simultanément les variations d'enveloppe spectrale et les variations de fréquence fondamentale, ce qui permettrait par exemple une étude fine de signaux de voix chantée.

5.1 Séparation de sources informée par la partition

Dans cette section, nous présentons une nouvelle technique de séparation de sources mono-canal dans des mélanges musicaux, qui utilise l'information de la partition musicale du morceau. Cette information est utilisée pour initialiser et contraindre une généralisation multi-instrument de la décomposition proposée dans la section 4.1 page 77. Cette décomposition génère des masques temps/fréquence qui peuvent être utilisés pour séparer les différents instruments harmoniques présents dans le mélange à l'aide de filtrage de Wiener.

La séparation de sources sous-déterminée est un important domaine de recherche depuis plusieurs dizaines d'années. Dans les applications musicales, on cherche généralement à isoler d'un mélange le signal de chaque instrument.

La séparation de source dite aveugle a été d'abord étudiée dans le cas déterminé ou sur-déterminé avec des techniques telles que l'analyse en composantes indépendantes [Cardoso, 1998] puis plus tard dans le cas sous-déterminé avec par exemple l'utilisation de la Factorisation en matrices non-négatives (NMF) [Virtanen, 2007]. Un des principaux problèmes avec ce type de technique est la difficulté à regrouper les éléments factorisés associés à une même source. Ainsi de nombreux travaux ont cherché à introduire de l'information additionnelle afin d'améliorer les résultats de séparation. Plusieurs types d'information ont été considérés : dans [Smaragdis *et al.*, 2007], les différentes formes spectrales de chaque source sont apprises sur des sons isolés puis sont utilisées pour décomposer le signal de mélange. Dans [Parvaix *et al.*, 2010], les signaux sources eux-mêmes sont utilisés comme

information dans un schéma codeur/décodeur.

Plus récemment, l'utilisation de la partition musicale (généralement sous la forme d'un fichier MIDI) comme information pour guider la séparation a été considérée dans plusieurs travaux : ce type d'information a l'avantage d'être une représentation très compacte de la musique (le fichier MIDI d'un morceau fait généralement quelques kilooctets contre plusieurs dizaines de mégaoctets pour la forme d'onde non compressée) ; de plus, une multitude de fichiers MIDI sont disponibles sur le Web. Dans [Woodruff *et al.*, 2006], une séparation de sources stéréo basée sur des indices spatiaux est renforcée par l'information de la partition. Dans une première étape, les points temps/fréquence ne contenant qu'une seule source (de façon très fortement prédominante) sont extraits et associés à leur source respective à partir de l'information spatiale seule. Dans une seconde étape, le nombre de sources présentes dans les points temps/fréquence restants est évalué en estimant la fréquence fondamentale de chaque source à l'aide des points temps/fréquence mono-source obtenus à la première étape. Les points proches d'un harmonique d'une source ont en effet de fortes chances de contenir de l'énergie de cette source. L'estimation des fréquences fondamentales des sources est renforcée en utilisant la partition (après association de chaque source obtenue dans la première étape à une piste de la partition, en utilisant la première estimation des fréquences fondamentales). Pour les points temps/fréquence ne contenant que deux sources, le problème reste inversible (deux canaux/deux sources) et la répartition dans chaque source se fait à partir de l'inversion de la matrice de mélange correspondante. Pour les autres points temps/fréquence, une troisième étape est nécessaire : l'importance relative de chaque source pour ces points temps/fréquence est estimée à partir d'un modèle temporel (l'amplitude des harmoniques correspondants est estimée en utilisant les amplitudes des trames précédentes et suivantes). Dans [Raphael, 2008], la mélodie principale est séparée de l'accompagnement en utilisant l'information de la partition. La séparation se fait à l'aide de deux masques binaires complémentaires (un pour le solo, un pour l'accompagnement) qui sont obtenus à l'aide d'une approche de type « classificateur » : à chaque point temps/fréquence, on cherche à associer une classe (solo ou accompagnement). La classification se fait à l'aide de descripteurs issus de la partition : distance horizontale (fréquence) aux harmoniques des notes jouées, distance verticale (temps)... Dans [Smaragdis et Mysore, 2009], Smaragdis propose d'utiliser une Analyse probabiliste en composantes latentes (**PLCA**) pour isoler des sons dans un mélange à partir d'une requête fredonnée. Cette requête imite un son cible que l'utilisateur cherche à extraire et sert d'information a priori dans la décomposition **PLCA** du mélange. Cette approche est réutilisée dans [Gansemann *et al.*, 2010b,a], en remplaçant la requête chantée par un fichier MIDI aligné au mélange et un synthétiseur : chaque piste du fichier MIDI est synthétisée et est alors utilisée comme a priori dans la **PLCA** du mélange. Dans [Every et Szymanski, 2004], des filtres harmoniques sont générés à partir de la partition : dans chaque fenêtre d'analyse, la fréquence fondamentale de chaque note active dans le fichier MIDI est estimée finement à partir des pics (maxima) f_{jp} du spectre de la trame les plus proches des harmoniques. Chaque maximum spectral est alors associé à un harmonique d'une ou plusieurs notes. S'il n'y a qu'une seule note, alors il suffit de filtrer le pic correspondant, sinon l'énergie est répartie entre les différentes notes en utilisant les amplitudes des harmoniques voisins (par exemple la moyenne des amplitudes de l'harmonique $j - 1$ et de l'harmonique $j + 1$).

Dans cette section, nous proposons une nouvelle approche de séparation de source informée par la partition, basée sur une extension multi-instrument du modèle paramétrique

de spectrogramme proposé dans la section 4.1. La partition utilisée comme information se présente sous la forme d'un fichier MIDI (qui est donc à un niveau symbolique moindre que celui d'une vraie partition). Le flux MIDI est supposé aligné temporellement sur le signal de mélange : nous ne nous intéressons pas dans ce travail aux thématiques d'alignement audio/MIDI qui font l'objet d'une littérature abondante [Dannenberg et Hu, 2003, Raphael, 2006, Joder et al., 2011]. L'information extraite de la partition est utilisée pour initialiser l'algorithme qui fournit la décomposition, cette initialisation contraignant la forme de la solution. L'algorithme optimise alors localement les paramètres (notamment les fréquences fondamentales de chaque atome dans chaque trame). La décomposition obtenue fournit un masque temps/fréquence pour chaque source : ces masques sont ensuite utilisés pour séparer les sources à partir du mélange par filtrage de Wiener. Ce travail est également décrit dans [Hennequin et al., 2011e].

Dans la section 5.1.1, une extension multi-instruments de la décomposition proposée dans la section 4.1 page 77 est présentée. Nous présentons alors le système de séparation informée par la partition dans la section 5.1.2. Une évaluation comparative des performances de cet algorithme est proposée dans la section 5.1.3.

5.1.1 Modèle paramétrique de spectrogramme de mélange

5.1.1.1 Modèle de spectrogramme source

Le modèle de spectrogramme utilisé pour chaque source du mélange est celui de la section 4.1 : le spectrogramme modèle de l'instrument isolé p (de la source p) est donc donné par l'équation (4.1) :

$$\begin{aligned} [\hat{\mathbf{V}}_p]_{ft} &= \sum_{r=1}^R w_{pfr}^{f_0^{prt}} h_{prt} \\ &= \sum_{r=1}^R \sum_{k=1}^{n_h(f_0^{rt})} a_k^p g(f - kf_0^{rt}) h_{prt}. \end{aligned} \tag{5.1}$$

Chaque source p a donc son propre jeu de paramètres f_0^{rt} , a_k^p et h_{prt} .

5.1.1.2 Modèle de spectrogramme de mélange

On suppose que le spectrogramme de puissance du mélange est simplement la somme des spectrogrammes de puissance des sources (ce qui correspond en fait à l'hypothèse courante faite en NMF qui consiste à considérer qu'il n'y a pas d'interférences entre les sources). Ainsi le modèle de spectrogramme du mélange est :

$$\mathbf{V}^{\text{mix}} \approx \hat{\mathbf{V}}^{\text{mix}} = \sum_{p=1}^P \hat{\mathbf{V}}_p, \tag{5.2}$$

où $\hat{\mathbf{V}}_p$ est le spectrogramme paramétrique de la source p (le nombre total de sources étant P) donné dans l'équation (5.1).

5.1.1.3 Modèle des éléments non harmoniques

Bien que cette partie du modèle ne soit pas utilisée dans la section 5.1.3 (section présentant les résultats) du fait que son apport ne soit pas significatif (les performances de séparation de sources ne sont pas améliorées par l'ajout de ce modèle), un travail a été fait pour prendre en compte les éléments non-harmoniques du son tels que le bruit « stationnaire » (bruit de souffle pour les instruments à vents), ainsi que les parties percussives (attaques).

Bruit localement « stationnaire » : On appelle « bruit stationnaire » (bien que ce bruit soit généralement non-stationnaire) tout ce qui n'est pas harmonique dans le spectrogramme mais dont l'activité est à peu près simultanée à la partie harmonique (typiquement le bruit de souffle pour les instruments à vent, ce qui peut rester dans le son du piano une fois qu'on a enlevé la partie harmonique et le bruit du marteau...).

On modélise ce bruit stationnaire par des atomes d'une NMF standard :

$$\hat{\mathbf{V}}_p^{\text{bruit}} = \mathbf{A}^p \mathbf{B}^p, \quad (5.3)$$

où \mathbf{A}^p et \mathbf{B}^p sont des matrices. \mathbf{A}^p est la matrice des atomes (colonnes) contenant les atomes de bruit « stationnaire » associés à la source k .

Lien avec les activations des atomes harmoniques : Pour éviter de prendre en compte dans ces atomes les parties non-stationnaires (attaques) et pour imposer que ces atomes représentent au mieux le bruit d'une unique source, il est nécessaire de forcer ces atomes de bruit stationnaire à être actifs en même temps que la partie harmonique. La technique adoptée est la suivante : on introduit un unique atome de bruit non-stationnaire par source (la matrice \mathbf{A}^p a donc une unique colonne), en supposant donc que l'enveloppe spectrale de ce bruit n'évolue pas au cours du temps. On impose alors que l'activation du bruit stationnaire soit égale à la somme des activations de toutes les notes.

La principale difficulté de ce type de modèle est d'éviter que les atomes « non-harmoniques » dont la forme est libre ne prennent en compte une partie de l'énergie des éléments harmoniques. De plus, il semble que les atomes NMF n'arrivent pas à prendre correctement en compte la partie non-harmonique associée à une source spécifique : on retrouve en effet dans la partie non-harmonique associée à une certaine source des éléments qui appartiennent à d'autres sources et la séparation peut être moins bonne que sans modèle de bruit.

Attaques : On appelle « attaque » tout ce qui n'est pas harmonique et qui est très fortement non-stationnaire. Les attaques sont modélisées par des motifs temps/fréquence et sont obtenues par une Déconvolution de facteurs de matrices non-négatives (NMFD) (*cf.* section 2.6.1.1 page 51) :

$$\mathbf{V}_p^{\text{attaques}} \approx \sum_{\tau=1}^L \mathbf{C}_{\tau}^p \overset{\leftarrow}{\mathbf{D}}_{\tau}^p, \quad (5.4)$$

où $\overset{\leftarrow}{\mathbf{D}}_{\tau}^p$ correspond à la matrice \mathbf{D}^p dont on a translaté tous les coefficients de τ indices vers la gauche : l'équation (5.4) correspond donc à une opération de convolution.

Lien avec les activations des atomes harmoniques : Les atomes NMFD ne présentent un intérêt que si les activations associées sont très parcimonieuses, l'idée étant que les motifs temps/fréquence factorisés doivent au mieux représenter les sons des attaques.

On associe donc un unique atome temps/fréquence à chaque source (on peut éventuellement ne pas en associer du tout aux sources qui ne sont pas censées avoir d'attaque) et on force les activations à être nulles partout sauf à l'instant des attaques (ces instants étant fournis par la partition).

Ce type d'atome temps/fréquence peut aussi permettre de modéliser les instruments uniquement percussifs.

Spectrogramme de mélange avec partie non harmonique : En prenant en compte la partie de bruit stationnaire et les attaques, le spectrogramme de source de l'équation (5.1) devient :

$$\hat{\mathbf{V}}_p = \hat{\mathbf{V}}_p^{\text{harmo}} + \hat{\mathbf{V}}_p^{\text{bruit}} + \hat{\mathbf{V}}_p^{\text{attaques}},$$

$$\text{avec } [\hat{\mathbf{V}}_p^{\text{harmo}}]_{ft} = \sum_{r=1}^R w_{pfr}^{f_0^{prt}} h_{prt}.$$

Le spectrogramme de mélange de l'équation (5.2) devient alors :

$$\mathbf{V}^{\text{mix}} \approx \hat{\mathbf{V}}^{\text{mix}} = \sum_{p=1}^P \left[\hat{\mathbf{V}}_p^{\text{harmo}} + \hat{\mathbf{V}}_p^{\text{bruit}} + \hat{\mathbf{V}}_p^{\text{attaques}} \right]. \quad (5.5)$$

Il est à noter qu'en utilisant un seul atome de bruit stationnaire par source, la partie bruit peut se réécrire simplement sous forme d'une NMF :

$$\hat{\mathbf{V}}^{\text{bruit}} = \sum_{k=1}^K \hat{\mathbf{V}}_k^{\text{bruit}} = \mathbf{E}\mathbf{F},$$

$$\text{avec } [\mathbf{E}]_{fk} = [\mathbf{A}^k]_{f1} \text{ et } [\mathbf{F}]_{kt} = [\mathbf{B}^k]_{1t}.$$

De même, lorsqu'on utilise un seul atome temps/fréquence d'attaque par source, on peut également réécrire le terme d'attaque sous forme d'une NMFD, en intervertissant de la même façon les indices d'atomes et de sources.

5.1.2 Système de séparation

Le modèle de spectrogramme présenté dans la section précédente est utilisé pour décomposer le spectrogramme de mélange à l'aide d'un algorithme à règle multiplicative qui vise à minimiser une β -divergence entre \mathbf{V}^{mix} et $\hat{\mathbf{V}}^{\text{mix}}$. Cet algorithme est très similaire à celui présenté dans la section 4.1.2. L'algorithme est initialisé en utilisant l'information de la partition. Comme nous le verrons cette initialisation particulière contraint également la forme de la solution. La partition, sous forme d'un flux MIDI supposé temporellement aligné avec le signal, permet d'initialiser la décomposition dans un voisinage vraisemblablement assez proche d'une décomposition optimale. Comme déjà évoqué précédemment, le travail présenté ici ne s'intéresse pas aux problématiques d'alignement.

5.1.2.1 Initialisation à l'aide de la partition

Le flux MIDI permet de générer un *piano roll* pour chaque source de même dimension que les activations des atomes harmoniques de cette source. Ce *piano roll* est utilisé pour initialiser les activations : pour tous les instants pour lesquels une note est active, l'activation de l'atome harmonique correspondant à cet instant est initialisée à 1 ; pour tous les instants pour lesquels une note n'est pas active, l'activation de l'atome harmonique correspondant à cet instant est initialisée à 0. Comme l'algorithme de décomposition utilisé est un algorithme multiplicatif, la forme de la solution est contrainte par cette initialisation : en effet les valeurs d'activation initialement nulles resteront nulles tout au long de l'algorithme. Ainsi on impose une forte contrainte à la solution. C'est pourquoi il est nécessaire d'élargir les initialisations à 1 légèrement avant le début de la note et après la fin de la note afin de prendre en compte d'éventuelles petites erreurs d'alignement ou bien d'éventuelles décroissances lentes de certaines notes (phase *Release* dans le modèle Attack Decay Sustain Release (**ADSR**) décrit par exemple dans [Roads, 1996, p.97-98]). Un exemple de masques d'initialisation des activations pour les différentes sources est présenté dans la figure 5.1.

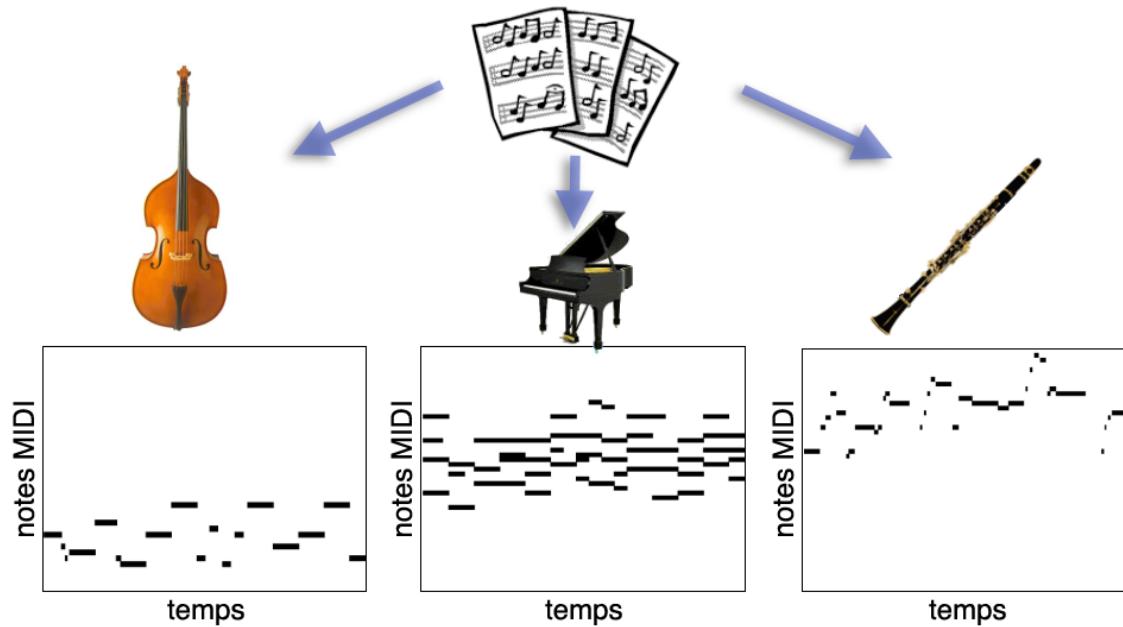


FIG. 5.1 – Masques d'activation pour les 3 instruments d'un morceau : les activations de chaque source sont initialisées avec un *piano roll* binaire (0 ou 1) de la piste MIDI correspondante.

5.1.2.2 Algorithme de décomposition

La décomposition est obtenue en minimisant une β -divergence entre \mathbf{V}^{mix} et $\hat{\mathbf{V}}^{\text{mix}}$ par rapport aux paramètres du modèle, c'est-à-dire pour chaque source p :

- la fréquence fondamentale de chaque atome r à chaque instant t : f_0^{prt} ,
- les amplitudes des harmoniques a_{pk} ,
- les activations de chaque atome à chaque instant h_{prt} .

La minimisation est faite grâce à un algorithme à règles de mise à jour multiplicatives qui sont successivement appliquées à chacun des paramètres précédents. Cet algorithme est très similaire à celui proposé dans la section 4.1.2 et les règles de mise à jour peuvent être facilement déduites du modèle mono-instrument. Les règles de mises à jour pour les paramètres de la source k sont données par :

$$\begin{aligned} f_0^{krt} &\leftarrow f_0^{krt} \frac{\mathcal{F}_{krt}}{\mathcal{G}_{krt}}, \\ h_{krt} &\leftarrow h_{krt} \frac{\mathcal{M}_{krt}}{\mathcal{P}_{krt}}, \\ a_{kp} &\leftarrow a_{kp} \frac{\mathcal{N}_{kp}}{\mathcal{Q}_{kp}}, \end{aligned}$$

avec :

$$\begin{aligned} \mathcal{G}_{krt} &= \sum_{f=1}^F \sum_{p=1}^{n_p} a_{kp} p P(f - pf_0^{krt}) (\hat{V}_{ft}^{\text{mix}})^{\beta-2} (f \hat{V}_{ft}^{\text{mix}} + pf_0^{krt} V_{ft}^{\text{mix}}), \\ \mathcal{F}_{krt} &= \sum_{f=1}^F \sum_{p=1}^{n_h} a_{kp} p P(f - pf_0^{krt}) (\hat{V}_{ft}^{\text{mix}})^{\beta-2} (pf_0^{krt} \hat{V}_{ft}^{\text{mix}} + f V_{ft}^{\text{mix}}), \\ \mathcal{P}_{krt} &= \sum_{f=1}^F w_{kfr}^{f^{krt}} (\hat{V}_{ft}^{\text{mix}})^{\beta-1}, \\ \mathcal{M}_{krt} &= \sum_{f=1}^F w_{kfr}^{f^{krt}} (\hat{V}_{ft}^{\text{mix}})^{\beta-2} V_{ft}^{\text{mix}}, \\ \mathcal{Q}_{kp} &= \sum_{f=1}^F \sum_{t=1}^T \sum_{r=1}^R g(f - pf_0^{krt}) h_{krt} (\hat{V}_{ft}^{\text{mix}})^{\beta-1}, \\ \mathcal{N}_{kp} &= \sum_{f=1}^F \sum_{t=1}^T \sum_{r=1}^R g(f - pf_0^{krt}) h_{krt} (\hat{V}_{ft}^{\text{mix}})^{\beta-2} V_{ft}^{\text{mix}}, \end{aligned}$$

où P est la fonction définie dans l'équation (4.6).

Cet algorithme a pour but d'estimer finement un masque temps/fréquence correspondant approximativement au spectrogramme de chaque source, à partir des masques gros-siers obtenus lors de l'initialisation à partir de la partition : la figure 5.2 illustre ce processus.

5.1.3 Résultats

Les performances de notre algorithme sont évaluées sur une base de données de fichiers MIDI que l'on synthétise à l'aide de deux banques de sons différentes. On obtient alors pour chaque fichier MIDI deux fichiers audio (un généré par banque de son) qui sont parfaitement alignés sur le MIDI. Nous comparons sur cette base de données les résultats de notre algorithme à celui basé sur la **PLCA** [Ganseman *et al.*, 2010b,a] : l'algorithme basé sur la **PLCA** commence par synthétiser la partition et les signaux audio obtenus pour chaque piste sont utilisés comme a priori dans la décomposition **PLCA** du mélange. Cet algorithme nécessite donc un synthétiseur et c'est pourquoi nous utilisons deux banques de sons différentes pour synthétiser nos fichiers MIDI : dans un premier temps, l'une est

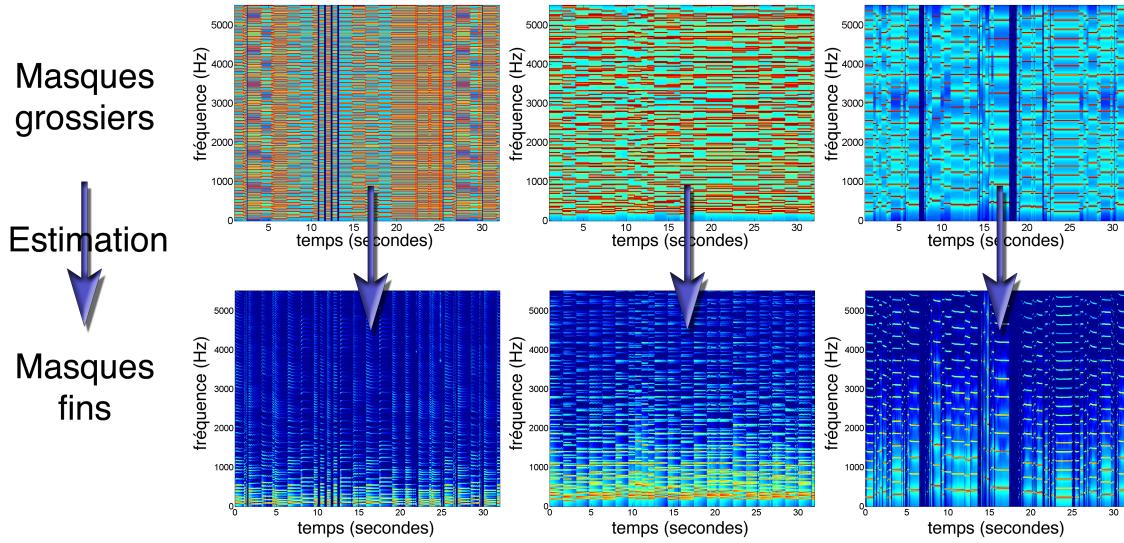


FIG. 5.2 – Estimation fine des masques temps/fréquence à partir des masques grossiers obtenus à l’initialisation.

utilisée pour synthétiser les mélanges à séparer, et l’autre est utilisée par l’algorithme de la **PLCA** pour obtenir les pistes séparées utilisées comme a priori, puis dans un deuxième temps, on répète l’expérience en inversant le rôle des deux banques de sons.

5.1.3.1 Description de la base de données

A notre connaissance, au moment où nous avons réalisé l’expérience, il n’existait pas de base de données publique contenant à la fois les instruments séparés et le mélange enregistrés dans des conditions réelles avec un fichier MIDI aligné pour plusieurs morceaux. Nous avons donc construit notre propre base de données en cherchant à reproduire au mieux les caractéristiques « réalistes » des morceaux de musique comme l’éventuel recouvrement des sources dans le domaine temps/fréquence (par opposition à la base de données construite aléatoirement proposée dans [Ganseman *et al.*, 2010a]).

Pour chaque morceau, les pistes étaient sommées entre elles pour obtenir le mélange. La base de données était constituée de 12 fichiers MIDI libres de droit de quatuor à corde de Bach, Beethoven et Boccherini. Ces fichiers MIDI ont été synthétisés à l’aide de deux différentes méthodes : la première méthode utilise des sons de notes isolées enregistrées à partir d’instruments de musique réels avec trois niveaux de vitesse différents. Les instants d’attaque sont alors synchronisés avec les messages MIDI **Note On** et les sons disparaissent au moment des **Note Off**. La liste d’instruments comprend du violon, de l’alto et du violoncelle. Cette première technique sera appelée « M1 » par la suite. La seconde technique est basée sur le logiciel TiMidity¹ utilisé avec une banque de sons généraliste et assez réaliste (Crisis General Midi 3.0²). Cette deuxième méthode sera désignée par « M2 ». M2 inclue un effet de réverbération (présent dans la banque de son) alors que M1 en est

¹<http://timidity.sourceforge.net/>

²<http://www.bismutnetwork.com/10Music/Crisis/Soundfont3.0.php>

dépourvu.

Dans ce travail, on n'évalue pas la dégradation des performances des algorithmes par rapport aux erreurs d'alignement. Les signaux de la base de données ont été rééchantillonnés à 11025Hz. Les signaux mélanges sont tous monophoniques. La base de données (fichiers MIDI, pistes séparées et mélanges) est disponible en ligne à l'adresse <http://perso.enst.fr/hennequi/database.zip>.

5.1.3.2 Expérience

Seules les 30 premières secondes de chaque morceau sont traitées. Les signaux de mélange sont séparés en utilisant l'information du fichier MIDI associé à l'aide de deux algorithmes :

- l'algorithme que nous avons présenté dans la section 5.1.2,
- l'algorithme basé sur la **PLCA** présenté dans [Ganseman *et al.*, 2010b,a].

Comme expliqué précédemment, l'algorithme basé sur la **PLCA** nécessite une phase d'apprentissage dans laquelle les pistes MIDI des différents fichiers sont synthétisées et utilisées comme *a priori* dans la décomposition du spectrogramme mélange. Afin de rendre la comparaison équitable entre les algorithmes, la méthode de synthèse utilisée pour générer les pistes séparées servant à l'apprentissage doit être différente de celle utilisée pour synthétiser les signaux de mélange. Comme les sons de notre base de données ont été synthétisés de deux façons différentes, nous en utilisons une pour l'apprentissage et une pour la séparation puis dans un second temps nous échangeons leurs rôles. Nous présentons aussi les résultats pour l'algorithme basé sur la **PLCA** lorsque la même méthode de synthèse est utilisée pour le test et l'apprentissage : ces résultats peuvent être interprétés comme une référence haute des algorithmes de **PLCA** et peuvent être quasiment considérés comme les résultats d'un oracle pour le filtrage de Wiener.

Pour les deux algorithmes, les signaux de mélange (ainsi que les signaux des pistes séparées pour l'algorithme basé sur la **PLCA**) ont été transformés en spectrogrammes en utilisant une Transformée de Fourier à Court Terme (**TFCT**) avec des fenêtres de Hann de 93ms, avec 75% de recouvrement.

La base de données contient par construction les sources séparées originales : on peut donc utiliser la boîte à outil **BSS_EVAL** [Vincent *et al.*, 2006] afin d'évaluer et de comparer les performances des algorithmes. Les performances sont donc évaluées en termes de Rapport signal à interférence (**SIR**), de Rapport signal à artefact (**SAR**) et de Rapport signal à distorsion (**SDR**), tous définis dans [Vincent *et al.*, 2006]. Les paramètres de l'algorithme basé sur la **PLCA** (nombre d'atomes, poids des *a priori*) ont été optimisés (sur un unique morceau) afin que l'algorithme donne les meilleurs résultats possibles.

5.1.3.3 Résultats

Les résultats sont représentés dans les figures 5.3 et 5.4 et détaillés dans les tableaux 5.1 et 5.2.

Dans les deux figures, les barres bleues correspondent aux rapports obtenus avec notre algorithme, les barres vertes aux rapports obtenus avec l'algorithme basé sur la **PLCA** en utilisant des méthodes de synthèse différentes pour l'apprentissage et l'évaluation, et les barres rouges aux rapports obtenus avec l'algorithme basé sur la **PLCA** mais cette fois-ci en utilisant la même méthode de synthèse pour l'apprentissage et l'évaluation (résultats de

référence). La figure 5.3 (et la table 5.1) correspond à l'expérience utilisant M1 comme méthode de synthèse pour les signaux d'évaluation et la figure 5.4 (et la table 5.2) correspond à l'expérience utilisant M2 comme méthode de synthèse pour les signaux d'évaluation. Les rapports donnés sont des moyennes sur tous les morceaux et toutes les sources.

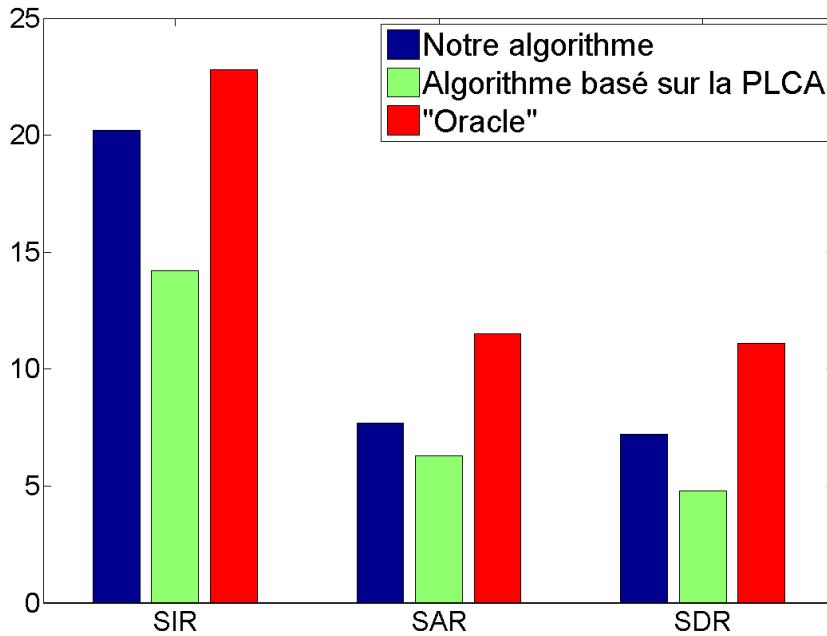


FIG. 5.3 – Résultats en utilisant M1 comme méthode de synthèse.

	SIR	SAR	SDR
Notre algorithme	20.2	7.7	7.2
Algo PLCA (apprentissage : M2)	14.2	6.3	4.8
Algo PLCA (apprentissage : M1)	22.8	11.5	11.1

TAB. 5.1 – Résultats en utilisant M1 comme méthode de synthèse.

Dans la figure 5.3, les valeurs des rapports (**SIR**, **SAR**, et **SDR**) sont meilleures pour notre algorithme que pour celui basé sur la **PLCA** de plus d'1dB en **SAR** et en **SDR** et de 6dB en **SIR**. De plus, ces valeurs sont assez proches de celles fournies par l'algorithme basé sur la **PLCA** en utilisant les mêmes signaux test et apprentissage (ici considéré comme un oracle). Dans la figure 5.4, les performances se détériorent pour les deux algorithmes, ce qui est probablement dû à l'importante réverbération produite par M2. Notre algorithme présente alors des résultats légèrement moins bons que celui basé sur la **PLCA**.

Plusieurs raisons peuvent expliquer que, bien que le modèle **PLCA** soit plus libre (notre modèle de spectrogramme paramétrique est bien plus contraint) et donc devrait être plus robuste, les performances de l'algorithme basé sur la **PLCA** ne sont pas meilleures :

- les échantillons isolés de violon et alto utilisés par la méthode M1 contiennent un

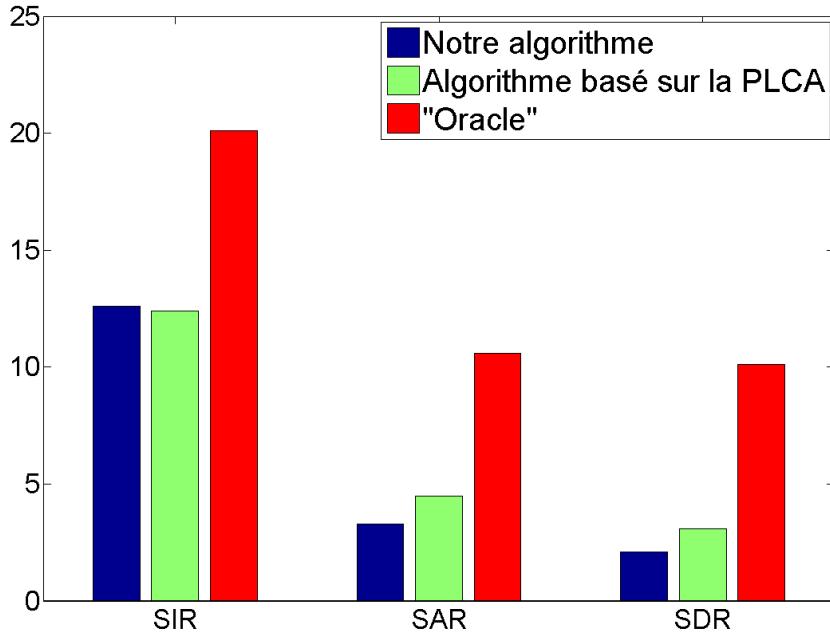


FIG. 5.4 – Résultats en utilisant M2 comme méthode de synthèse.

	SIR	SAR	SDR
Notre algorithme	12.6	3.3	2.1
Algo PLCA (apprentissage : M1)	12.4	4.5	3.1
Algo PLCA (apprentissage : M2)	20.1	10.6	10.1

TAB. 5.2 – Résultats en utilisant M2 comme méthode de synthèse.

vibrato important alors que dans les sons de M2, il n'y en a quasiment pas. Le fait que notre modèle modélise explicitement le vibrato explique probablement en partie les résultats de la figure 5.3 ;

- notre approche est moins supervisée que l'approche **PLCA** : en effet notre algorithme ne nécessite pas de synthétiser les pistes, et donc n'a pas besoin d'information sur les types d'instruments présents et n'utilise donc pas d'information sur le timbre des instruments ;
- les performances de l'algorithme basé sur la **PLCA** dépendent du synthétiseur utilisé pour générer les pistes séparées : par exemple, les signaux synthétisés avec M2 n'ont pas le même temps de relâchement que ceux de M1 et présentent de la réverbération, ce qui fournit vraisemblablement des a priori erronés pour les activations de la **PLCA**.

On peut donc considérer que notre méthode a des performances assez similaires à l'algorithme basé sur la **PLCA** tout en étant moins supervisée.

5.1.4 Conclusion

Nous avons présenté une nouvelle méthode de séparation de sources informée par la partition. Cette méthode est basée sur le modèle paramétrique de spectrogramme utilisant des atomes harmoniques variables présentés dans la section 4.1. Nous avons construit une base de données d'évaluation et notre méthode atteint des performances semblables à celle de l'algorithme basé sur la **PLCA** proposé dans [Ganseman *et al.*, 2010b] alors que notre méthode utilise moins d'informations (l'information sur le timbre des instruments n'est pas exploitée).

Plusieurs pistes pourraient être explorées pour prolonger ce travail :

- il serait possible d'utiliser d'autres critères de coût. En particulier, il serait intéressant d'introduire des critères de coût qui introduisent des poids pour renforcer les zones temps/fréquence où un seul instrument est actif, afin de favoriser l'apprentissage des amplitudes des harmoniques dans ces zones (et la fréquence fondamentale pour les instruments à fréquence fondamentale fixe).
- il serait intéressant de chercher à rendre le système plus robuste vis-à-vis d'erreurs importantes d'alignement : si l'alignement est très mauvais, les performances de notre système s'effondrent (c'est aussi le cas pour l'algorithme basé sur la **PLCA** si les a priori sur les activations sont forts).
- il pourrait être intéressant d'utiliser l'information de timbre utilisée par l'algorithme basé sur la **PLCA**, par exemple en introduisant un apprentissage supervisé des amplitudes des harmoniques de chaque instrument.

Cette application de séparation de sources informée a été publiée dans [Hennequin *et al.*, 2011e].

5.2 Modification de notes isolées dans un signal polyphonique

Il est possible de générer des masques temps/fréquence qui peuvent être directement utilisés pour séparer différentes composantes par filtrage de Wiener. Par conséquent, il est possible d'isoler des notes seules dans un signal polyphonique et de les modifier individuellement, par exemple en changeant la fréquence fondamentale.

L'utilisation de spectrogrammes issus d'une **TFCT** permet de reconstruire facilement le signal dans le domaine temporel. Notre décomposition permet d'isoler le spectrogramme d'une seule note, en ne reconstruisant qu'à partir d'une zone située autour d'une note dans la distribution d'impulsions P_I . Ainsi, on peut, à l'aide de filtrage de Wiener, reconstruire le son d'une note isolée, ce qui permet de traiter cette note indépendamment du reste du signal : il est par exemple possible de supprimer cette note, de modifier sa durée, de la décaler dans le temps, ou bien de la transformer en une autre note.

5.2.1 Méthode de séparation

5.2.1.1 Méthode

Comme montré dans la section 4.2 page 94, la décomposition invariante par homothétie temporelle permet d'obtenir, à partir d'un spectrogramme classique, une représentation multi-niveau qui permet de repérer aisément les notes présentes : les notes se traduisent en effet par des valeurs élevées dans la distribution d'impulsions P_I . On peut alors aisément repérer

l'ensemble de points (t, k) associé à une note : c'est en sélectionnant les indices associés à une note qu'on va être capable d'en isoler le son. On peut éventuellement sélectionner ces points de façon semi-automatique avec un outil graphique tel que l'outil « sélection baguette magique » qu'on trouve dans la plupart des logiciels de retouche d'images : cet outil fait une recherche récursive des points de couleur similaire dans le voisinage d'un point sélectionné (avec une certaine tolérance réglable).

Pour séparer un élément dont les indices dans la distribution d'impulsions sont $\mathcal{S} \in \{1, \dots, T\} \times \{1, \dots, K\}$, on peut alors couper la distribution $P(f, t)$ en deux parties complémentaires :

$$P(f, t) = P_{\mathcal{S}}(f, t) + P_{\complement \mathcal{S}}(f, t),$$

avec :

$$\begin{aligned} P_{\mathcal{S}}(f, t) &= \sum_{z, f'} P(z) P_K(f'|z) \sum_{k=k_{\min}^{f, f'}}^{k_{\max}^{f, f'}} P_I(\lambda_k, t|z) \mathbb{1}_{\mathcal{S}}(t, k) \delta \lambda_k^{f, f'}, \\ P_{\complement \mathcal{S}}(f, t) &= \sum_{z, f'} P(z) P_K(f'|z) \sum_{k=k_{\min}^{f, f'}}^{k_{\max}^{f, f'}} P_I(\lambda_k, t|z) (1 - \mathbb{1}_{\mathcal{S}}(t, k)) \delta \lambda_k^{f, f'}. \end{aligned}$$

$P_{\mathcal{S}}$ modélise le spectrogramme des éléments associés aux indices \mathcal{S} dans la distribution d'impulsions.

On obtient alors les « spectrogrammes » des éléments séparés par filtrage de Wiener :

$$\begin{aligned} [\mathbf{X}^{\mathcal{S}}]_{ft} &= \frac{P_{\mathcal{S}}(f, t)}{P(f, t)} [\mathbf{X}]_{ft}, \\ [\mathbf{X}^{\complement \mathcal{S}}]_{ft} &= \frac{P_{\complement \mathcal{S}}(f, t)}{P(f, t)} [\mathbf{X}]_{ft}. \end{aligned}$$

Les signaux temporels sont alors obtenus par **TFCT** inverse de $\mathbf{X}^{\mathcal{S}}$ et de $\mathbf{X}^{\complement \mathcal{S}}$.

5.2.1.2 Exemple

Pour illustrer cette technique, on reprend l'exemple décomposé à l'aide d'une **PLCA** invariante par homothétie fréquentielle avec un unique atome ($Z = 1$) présenté dans la section 4.2.4.2 page 108 : la distribution d'impulsions obtenue fait apparaître clairement des notes comme le montre la figure 5.5.

On peut alors facilement isoler les indices associés à une note : il suffit de sélectionner une zone noire dans la distribution d'impulsions. Dans la figure 5.6, on a séparé la distribution d'impulsions P_I en deux parties : dans la figure 5.6(a) est représentée la partie associée à la note qu'on veut isoler (la partie hachurée sur la figure est mise à zéro), et dans la figure 5.6(b), la partie complémentaire.

On peut alors reconstruire $P_{\mathcal{S}}(f, t)$ et $P_{\complement \mathcal{S}}(f, t)$ comme le montre la figure 5.7 : on constate bien que seuls les harmoniques de la note considérée apparaissent dans le spectrogramme modèle $P_{\mathcal{S}}(f, t)$ (*cf.* figure 5.7(c)) et que ces harmoniques ont bien disparu dans le spectrogramme modèle du reste $P_{\complement \mathcal{S}}(f, t)$ (*cf.* figure 5.7(d)) alors qu'ils apparaissent bien dans $P(f, t)$ (*cf.* figure 5.7(b)). On remarque cependant que l'attaque de la note isolée

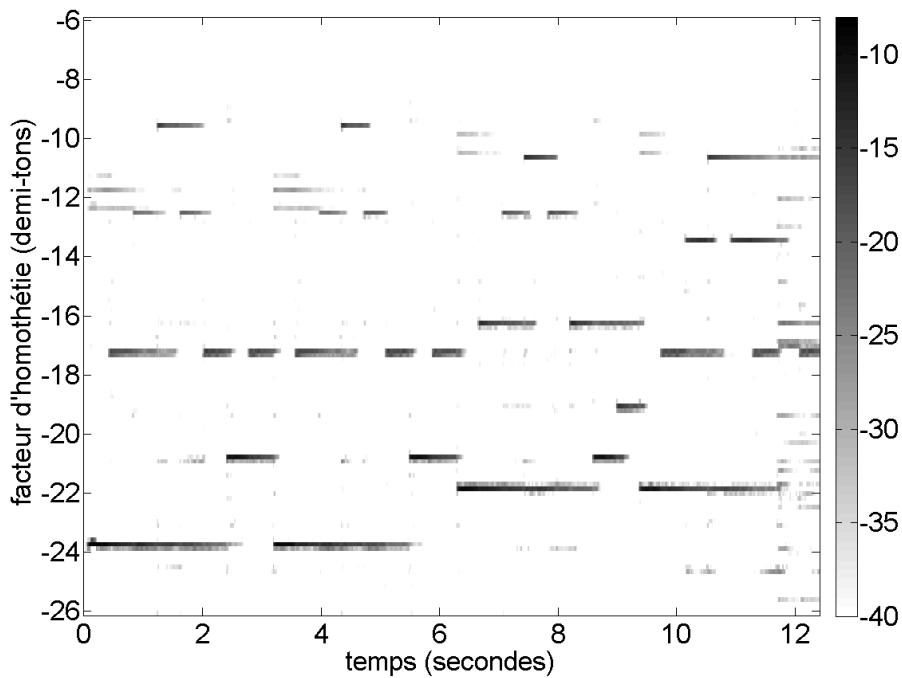


FIG. 5.5 – Distribution d’impulsions P_I de l’introduction de *Because* des Beatles.

(motif vertical en début de note) n’apparaît pas dans le spectrogramme modèle de la note isolée mais dans le spectrogramme modèle du reste : ceci est dû à l’absence de modélisation des éléments percussifs dans le modèle. Ainsi les éléments percussifs se retrouvent sous forme diffuse dans la distribution P_I et sont donc difficilement manipulables. L’introduction d’atomes de **PLCA** non modifiables (qu’on ne peut pas transformer par homothétie) serait une piste à explorer pour résoudre ce problème.

Par filtrage de Wiener, on peut alors récupérer d’une part le son de la note isolée, d’autre part, le reste du signal. On peut ainsi modifier le son de la note indépendamment du reste du signal.

5.2.2 Modifications

Plusieurs types de modifications simples peuvent être opérés sur une note isolée :

- il est possible par exemple d’en modifier le volume voire de les supprimer en appliquant seulement un gain sur le signal séparé ;
- il est également possible de modifier la fréquence fondamentale de la note isolée en utilisant un algorithme classique comme le vocodeur de phase [Flanagan et Golden, 1966] ou même PSOLA [Moulines et Charpentier, 1990, Moulines et Laroche, 1995] puisque le signal dont on veut modifier la fréquence fondamentale n’est pas polyphonique. Il est ainsi possible de modifier la fréquence fondamentale de toutes les occurrences d’une même note ce qui peut permettre de changer la tonalité d’un morceau par exemple de majeure à mineure ou réciproquement : ce type de trans-

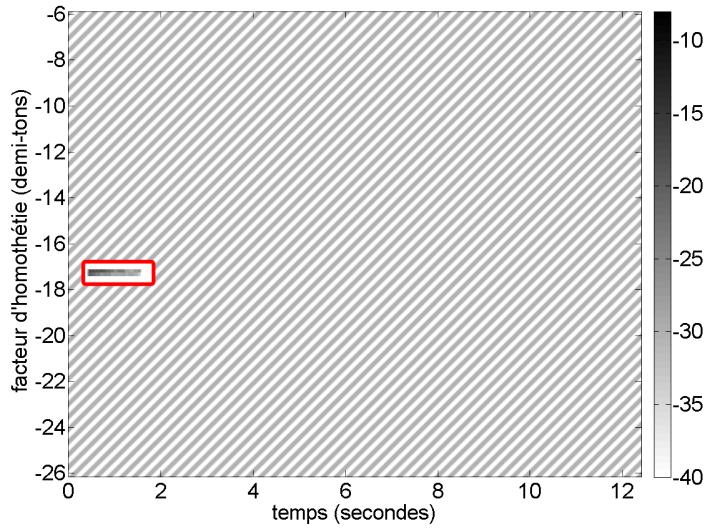
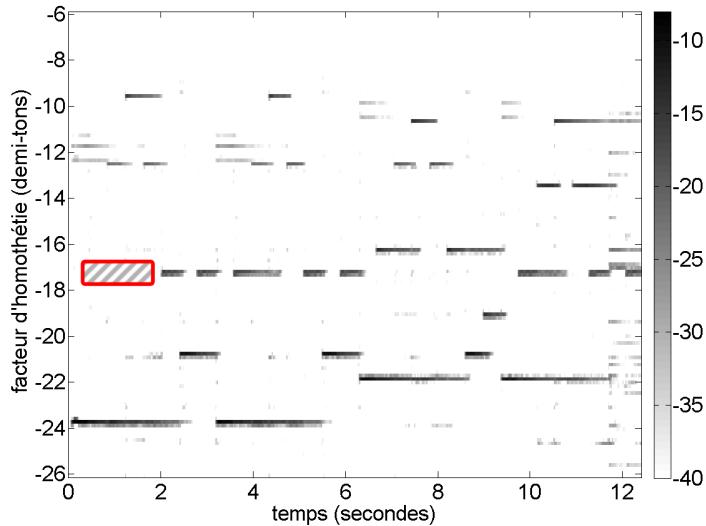
(a) P_I masqué : note isolée(b) P_I masqué : reste

FIG. 5.6 – Distribution d’impulsions P_I séparée en deux parties : la partie associée à la note à isoler 5.6(a) et la partie où cette note a été supprimée 5.6(b).

formation peut également être réalisé grâce au vocodeur à modulation (MODVOC) [Disch et Edler, 2008, 2009] ;

- il serait aussi théoriquement possible de modifier la position temporelle de la note isolée, en appliquant simplement un retard sur le signal : toutefois, il est à noter que notre système de séparation ne modélise pas les attaques et que le signal de note isolé ne contient généralement que la partie harmonique de la note. Pour réaliser cette modification proprement, il faudrait donc un modèle de son percussif qui permette

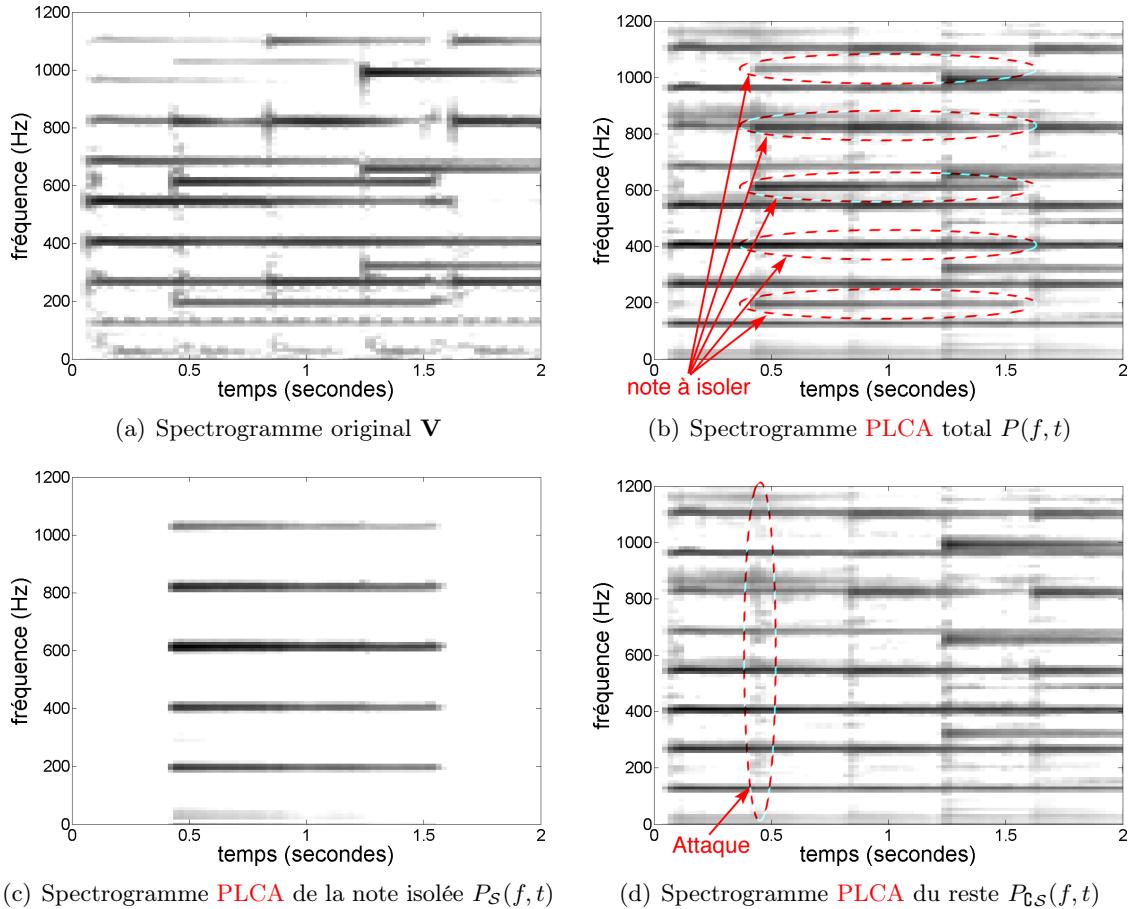


FIG. 5.7 – Zoom sur les basses fréquences des premières secondes du spectrogramme 5.7(a) original et des spectrogrammes modèles 5.7(b) (total), 5.7(c) (note isolée) et 5.7(d) (reste) de l'introduction de *Because*.

de séparer l'attaque avec la partie harmonique de la note.

En termes d'état de l'art sur ce type de modification de notes isolées dans un signal polyphonique, on peut citer le logiciel commercial *Melodyne* édité par Celemony³. Malheureusement, cette entreprise n'a pas du tout publié sur le fonctionnement de ce système.

Quelques exemples audio de cette application sont disponibles sur la page Web <http://perso.telecom-paristech.fr/hennequi/demoSIPPCA.html>.

5.3 Fusion des modèles paramétrique et source/filtre

Il est possible de fusionner le modèle de spectrogramme à atomes paramétriques (présenté dans la section 4.1 page 77) avec le modèle source/filtre (présenté dans le chapitre 3 page 77). Cependant, le nombre important d'atomes introduits par le modèle paramétrique rend l'estimation des filtres difficile en particulier pour l'utilisation de filtres Auto-Régressif(ve) (AR) : en effet, dans le modèle de la section 4.1, de nombreux atomes étaient

³<http://www.celemony.com/>

introduits même s'ils ne correspondaient pas à des éléments présents dans le mélange et étaient ensuite supprimés si leurs activations étaient trop faibles. La liberté supplémentaire apportée par exemple par un filtre AR permet aisément de faire correspondre un unique harmonique d'un atome censé être absent avec un partiel du spectrogramme original. Ce problème limite fortement l'utilisation du modèle présenté. Ce modèle n'est donc pas totalement abouti et est donné à titre indicatif : l'ajout de fortes contraintes de régularité semble nécessaire pour rendre le modèle robuste.

Nous présentons donc ici une utilisation simple d'une méthode de décomposition basée sur la fusion des deux modèles.

5.3.1 Modèle mixte

Le modèle mixte est simplement obtenu en fusionnant les équations (3.1) page 58 et (4.1) page 77. On obtient ainsi l'équation :

$$[\mathbf{V}]_{ft} \approx [\hat{\mathbf{V}}]_{ft} = \sum_{r=1}^R w_{fr}^{\theta_{rt}} h_{rt}(f). \quad (5.6)$$

Ainsi, dans ce modèle, les atomes peuvent varier dans le temps via le paramètre θ_{rt} (en l'occurrence la fréquence fondamentale de l'atome) et leur enveloppe spectrale peut être modifiée via la dépendance fréquentielle des activations.

Remarque : Il est à noter qu'une dépendance temporelle des atomes peut être interprétée de façon duale comme une dépendance fréquentielle des activations (si les activations agissent comme un opérateur sur les atomes). Ainsi, la manière dont ces dépendances sont introduites dans l'équation (5.6) constitue uniquement une façon de « penser » le modèle, et il serait tout à fait possible d'envisager une autre organisation de ces dépendances.

Nous reprenons la même structure paramétrique pour les atomes $w_{fr}^{\theta_{rt}}$ définie dans l'équation (4.2) :

$$w_{fr}^{f_0^{rt}} = \sum_{k=1}^{n_h(f_0^{rt})} a_k g(f - kf_0^{rt}).$$

Les atomes sont donc toujours des motifs harmoniques dont la fréquence fondamentale peut varier dans le temps et pour les raisons déjà évoquées dans la section 4.1.1.3 page 81, un atome est introduit par demi-ton de la gamme chromatique.

La paramétrisation Auto-Régressif(s) à Moyenne Ajustée (ARMA) de la dépendance fréquentielle des activations présentée dans l'équation (3.2) page 59 est également conservée :

$$h_{rt}^{\text{ARMA}}(f) = \sigma_{rt}^2 \frac{\left| \sum_{q=0}^Q b_{rt}^q e^{-i2\pi\nu_f q} \right|^2}{\left| \sum_{p=0}^P a_{rt}^p e^{-i2\pi\nu_f p} \right|^2}.$$

L'algorithme d'estimation utilisé est obtenu par les mêmes techniques que celles présentées dans les chapitres précédents.

5.3.2 Exemple de décomposition

Nous présentons dans cette section un exemple simple de décomposition d'un spectrogramme contenant des éléments fortement non-stationnaires présentant à la fois de fortes variations d'enveloppe spectrale et des variations de fréquence fondamentale.

Ce spectrogramme très simple, présenté dans la figure 5.8(a), est constitué de deux notes jouées par un synthétiseur. Il s'agit d'un synthétiseur soustractif (source/filtre) qui utilise un filtre résonant dont la fréquence de résonance varie rapidement au cours du temps (celle-ci augmente très rapidement au moment de l'attaque puis décroît lentement). Un léger vibrato est également introduit dans la forme d'onde. Les attaques des deux notes sont décalées, afin que les filtres qui agissent sur la forme d'onde de chacune des notes n'ait pas la même réponse fréquentielle au même instant. Les éléments du spectrogramme décomposé présentent ainsi simultanément des variations d'enveloppe spectrale et des variations de fréquence fondamentale.

On utilise le modèle de l'équation (5.6) pour décomposer ce spectrogramme avec 12 atomes (sur une octave), avec un filtre AR d'ordre 2 pour les activations temps/fréquence. On obtient alors le spectrogramme paramétrique présenté dans la figure 5.8(b) : ce spectrogramme préserve aussi bien les variations de fréquence fondamentale que les variations d'enveloppe spectrale.

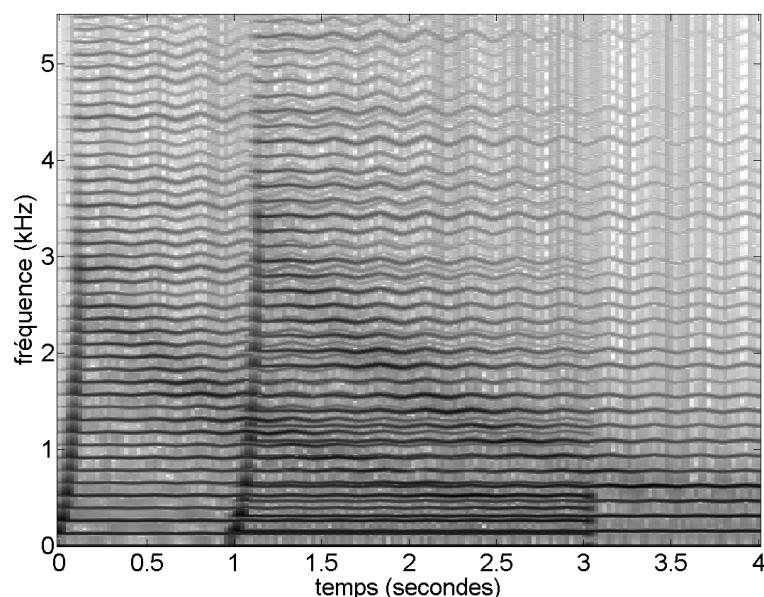
Une représentation des activations incluant les fréquences fondamentales estimées (représentation analogue à celle de la figure 4.8 page 90) est donnée dans la figure 5.9 : on voit clairement apparaître les deux notes présentes de manière prédominante, et le vibrato de chacune. La première note correspond à l'atome 5 et est active de 0s à 3s ; la seconde note correspond à l'atome 8 et est active de 1s à 4s.

Les activations temps/fréquence correspondant à chacune de ces notes sont représentées dans les figures 5.10(a) et 5.10(b). On voit clairement apparaître la résonance variable du filtre du synthétiseur dans chacune des notes.

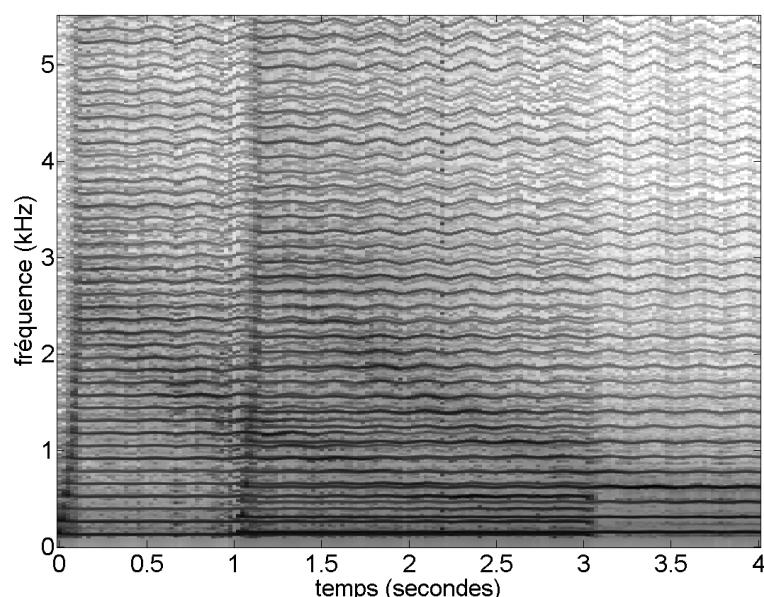
On constate que, dans les périodes d'inactivité des deux notes, l'algorithme estime parfois des filtres extrêmement résonants dont l'énergie (relativement faible) se concentre autour d'un unique partielle (qui correspond généralement à un élément associé à un autre atome). Cette estimation n'est évidemment pas souhaitable. Elle explique également, du moins en partie, les pics d'activations inattendus dans la figure 5.9.

Ce phénomène, en grande partie dû à la trop grande liberté laissée au modèle, est assez problématique et semble être actuellement la majeure limitation à l'utilisation du modèle mixte présenté dans cette section : il est en effet très difficile de contrôler le modèle et les résonances importantes au cours de l'algorithme. Le traitement de ces résonances non souhaitées n'est donc pour l'instant pas résolu. Une piste à étudier qui pourrait améliorer la robustesse du modèle consisterait à introduire des contraintes de régularité temporelle sur les filtres.

Nous avons également essayé d'appliquer l'algorithme hybride à des signaux de voix chantées qui contiennent à la fois d'importantes variations timbrales et des variations de fréquence fondamentale, mais les problèmes évoqués étaient particulièrement importants et rendaient la décomposition inutilisable.



(a) Spectrogramme original



(b) Spectrogramme reconstruit

FIG. 5.8 – Spectrogramme original et spectrogramme reconstruit des deux notes de synthétiseurs.

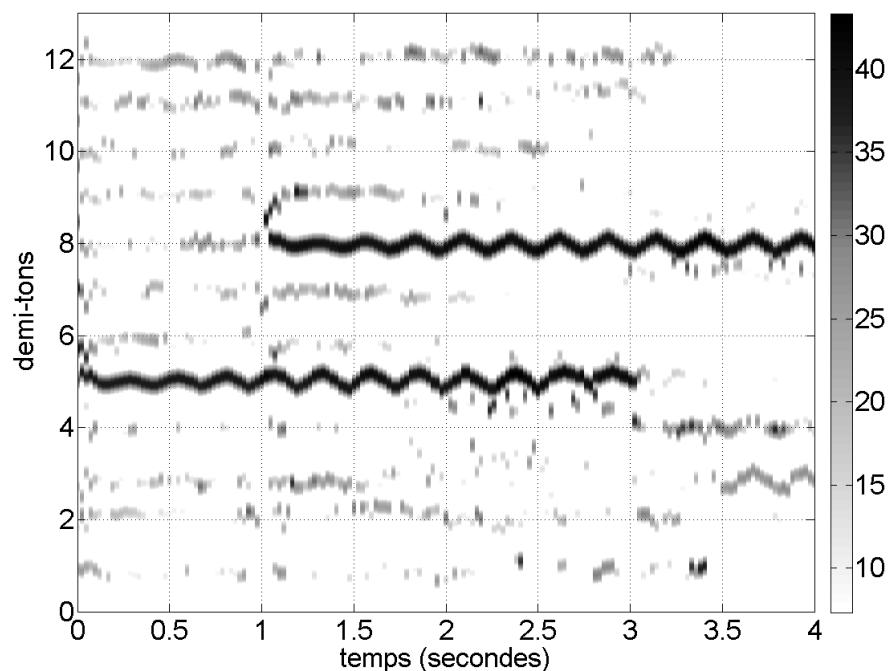
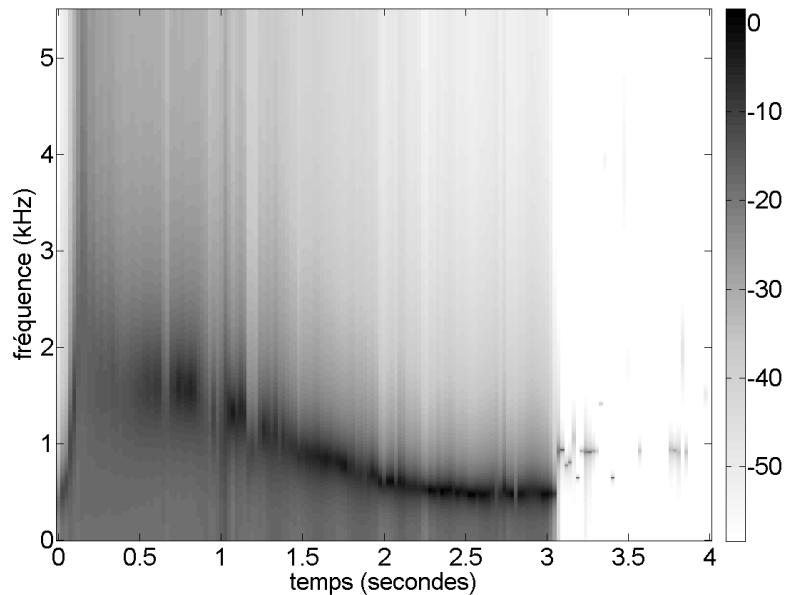
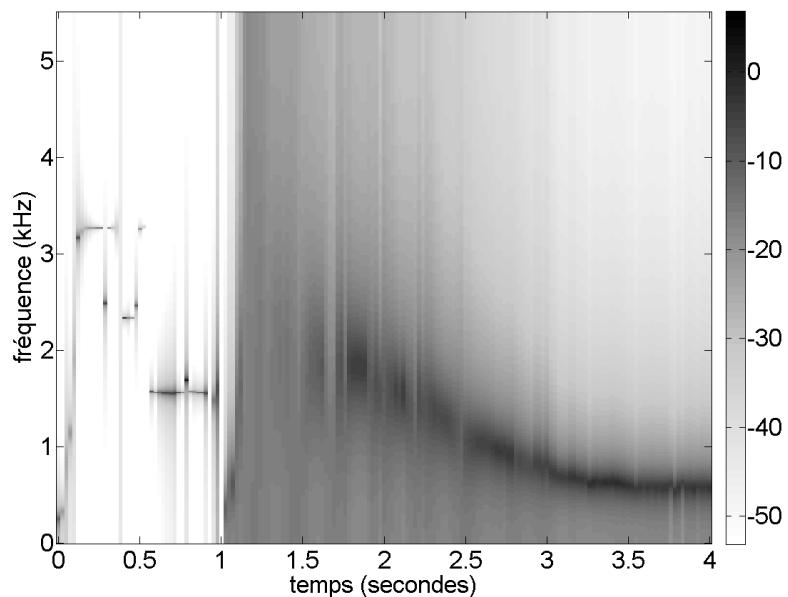


FIG. 5.9 – Représentation des activations incluant la fréquence fondamentale estimée à chaque instant en demi-tons relatifs.



(a) Activation temps/fréquence de l'atome 5



(b) Activation temps/fréquence de l'atome 8

FIG. 5.10 – Activation temps/fréquence $h_{rt}(f)$ des atomes 5 et 8.

Conclusion

Bilan

Dans cette thèse, nous avons proposé de nouvelles méthodes de décomposition des spectrogrammes musicaux. Ces méthodes sont issues des techniques de factorisation non-négative (Factorisation en matrices non-négatives (**NMF**) et ses dérivées comme l'Analyse probabiliste en composantes latentes (**PLCA**)) dans lesquelles nous avons introduit des modèles génératifs de spectrogrammes musicaux basés sur des modèles simples de synthèse sonore, tels que la synthèse source/filtre, la synthèse additive et la synthèse par table d'onde.

Ces nouvelles méthodes permettent notamment de prendre en compte certaines variations temporelles courantes au sein d'éléments sonores non-stationnaires, ce qui était problématique avec un simple modèle de décomposition linéaire comme la **NMF**. Les variations prises en compte sont de deux types : les variations d'enveloppe spectrale (courantes par exemple dans les sons d'instruments à cordes métalliques libres et dans tous les sons modulés par la bouche) et les variations de fréquence fondamentale (rencontrées par exemple dans des phénomènes tels que le vibrato ou la prosodie).

L'utilisation d'un modèle source/filtre permet ainsi de modéliser les variations d'enveloppe spectrale au cours du temps comme présenté dans le chapitre 3 et dans [Hennequin *et al.*, 2010a, 2011b].

L'introduction d'atomes harmoniques paramétriques à fréquence fondamentale variable [Hennequin *et al.*, 2010c] ainsi qu'un modèle de transformation des atomes [Hennequin *et al.*, 2011c,d] permettent de prendre en compte les variations de fréquence fondamentale, ce qui a été décrit dans le chapitre 4.

Bien que l'objet de cette thèse n'était pas de proposer une application précise, mais plutôt de fournir des méthodes permettant d'obtenir des représentations intermédiaires des signaux musicaux utilisables pour diverses applications, nous avons tout de même proposé quelques applications de ces méthodes dans le chapitre 5 : une application de séparation de sources informée par la partition musicale [Hennequin *et al.*, 2011e] ainsi qu'une application de transformation sélective du son dans un mélange polyphonique [Hennequin *et al.*, 2011d].

Même si des applications fonctionnelles ont été proposées, les modèles présentés restent perfectibles : ainsi l'algorithme proposé pour le modèle source/filtre (présenté dans le chapitre 3) présente des problèmes de stabilité numérique notamment lorsqu'il est utilisé avec un modèle Auto-Régressif(ve) (**AR**) et avec un nombre d'atomes trop important. Le modèle utilisant des atomes paramétriques (présenté dans la section 4.1) manque quant à lui de robustesse en termes de représentation pour des signaux complexes réels (même s'il

reste utilisable par exemple dans des applications de séparation de sources). L'algorithme de décomposition invariant par homothétie fréquentielle (présenté dans la section 4.2) reste coûteux en termes de temps de calcul (bien que des pistes d'amélioration aient déjà été évoquées). Enfin, le modèle hybride destiné à modéliser simultanément des variations de fréquence fondamentale et des variations d'enveloppe spectrale (présenté dans la section 5.3) n'est pas encore à un stade d'avancement qui permet de l'utiliser sur des signaux réels complexes.

Perspectives

Perfectibilité des modèles proposés

Les problèmes de robustesse liés à l'utilisation d'atomes paramétriques dont la forme est totalement contrainte ont été mis en lumière dans le chapitre 4 : l'utilisation d'atomes libres transformables (section 4.2) semble par exemple beaucoup plus adaptable à des signaux réels. C'est pourquoi il serait intéressant d'utiliser des modèles paramétriques plus souples, ou la forme de l'atome n'est pas totalement imposée mais est, par exemple, guidée par un atome synthétique : l'utilisation d'une loi de probabilité *a priori* paramétrique (par exemple une loi de Dirichlet comme proposé dans [Smaragdis *et al.*, 2006]) dont les paramètres seraient estimés en même temps que la décomposition semble une voie envisageable. Cette voie permettrait de garder une grande robustesse dans les cas de signaux réels qui peuvent s'écartier du modèle tout en gardant une structure paramétrique forte.

Il pourrait également intéresser d'élargir le modèle utilisant des atomes paramétriques (de la section 4.1) en introduisant d'autres paramètres afin de rendre les atomes plus généraux : on peut notamment penser à un paramètre d'inharmonicité qui permettrait de modéliser finement les sons de piano. Nous avons déjà tenté d'introduire un tel paramètre pour chaque atome sans succès, mais il semble que l'utilisation de modèles globaux des paramètres additionnels puisse permettre une estimation correcte : ainsi l'utilisation d'un modèle d'inharmonicité tel que présenté dans [Rigaud *et al.*, 2011], qui permet de réduire considérablement le nombre de paramètres, pourrait résoudre les problèmes rencontrés à ce sujet.

Pour éviter les problèmes de stabilité numérique rencontrés avec le modèle AR, plusieurs pistes semblent envisageables : une première piste consisterait à renforcer l'estimation des filtres en imposant une certaine continuité temporelle entre les filtres. L'utilisation d'une autre implémentation des filtres AR, par exemple une implémentation en treillis, pourrait également être envisagée : on chercherait alors à estimer les coefficients de réflexion, ce qui permettrait d'imposer des contraintes sur ceux-ci pour éviter d'obtenir des pôles trop proches du cercle unité. Il pourrait d'ailleurs être intéressant d'estimer directement les pôles.

Il semble également nécessaire d'approfondir le travail sur les modèles « hybrides » permettant de modéliser simultanément les variations de fréquence fondamentale et les variations d'enveloppe spectrale, le modèle présenté dans la section 5.3 page 126 n'étant

actuellement pas assez robuste pour être utilisé sur des signaux réels. Une piste de réflexion consisterait simplement à fusionner le modèle invariant par homothétie présenté dans la section 4.2 page 94 avec le modèle source/filtre (chapitre 3). Les modèles étant présentés dans des cadres un peu différents (**PLCA** pour le modèle invariant par homothétie et **NMF** déterministe pour le modèle source/filtre), ce travail nécessiterait de reprendre l'intégralité des calculs pour un des modèles. Cette piste semble prometteuse comme le montre [Fuentes *et al.*, 2011] qui propose un modèle hybride capable de modéliser les deux types de variations dans des spectrogrammes à Q-constant.

Il semble également nécessaire d'aborder un autre point crucial : certains des algorithmes présentés sont issus de méthodes heuristiques de minimisation (règles de mise à jour multiplicatives obtenues en écrivant le gradient de la fonction de coût comme une différence de deux termes positifs). Si, en pratique, ces méthodes peuvent donner de bons résultats, aucune preuve de convergence ou même de décroissance du critère n'a été apportée pour ces algorithmes. Il semble donc nécessaire de s'intéresser à des formalismes plus solides tels que celui des algorithmes Majoration/Minimisation (**MM**) (*cf.* section 2.5.2.2 page 45) afin de proposer des méthodes ayant de bonnes propriétés théoriques.

Enfin, les applications présentées sont dans le domaine de la séparation de sources (l'application de manipulation de notes isolées fonctionne par séparation des notes à modifier du reste du signal), mais il semble que les décompositions proposées puissent également servir de base pour d'autres applications. Ces décompositions fournissent en effet des représentations plus proches de la perception qu'un simple spectrogramme et pourraient ainsi être utilisées par exemple dans des tâches de transcription automatique.

Structuration temporelle

Bien que l'analyse de variations temporelles au sein d'éléments sonores soit un sujet central dans cette thèse, la structuration de ces variations (qui mènerait à une représentation de plus haut niveau) n'a pas vraiment été abordée, et il semble nécessaire de s'y attacher pour obtenir une réelle représentation « objets » des spectrogrammes musicaux. La robustesse des méthodes statistiques de décomposition de spectrogrammes basées sur les modèles de Markov cachés [Ozerov *et al.*, 2009, Nakano *et al.*, 2010b, Mysore *et al.*, 2010] en font une approche particulièrement intéressante pour traiter ce sujet. Cette approche est jusqu'alors assez complémentaire de celle incluant des modèles de signaux que nous avons choisie de suivre, mais il serait tout à fait possible de proposer une approche hybride, incluant à la fois un modèle « physique » des éléments décomposés et un modèle statistique structurant le déroulement temporel de ces événements. On pourrait ainsi envisager d'inclure des modèles physiques simples pour chaque état d'un objet sonore : on utiliserait un modèle d'attaque qui permettrait de générer un atome pour l'état « attaque », le modèle harmonique de la section 4.1 pourrait être réutilisé pour l'état « harmonique »...

Une telle structuration temporelle d'un modèle fortement paramétrique comme celui de la section 4.1 entraînerait d'importantes modifications au modèle même : ainsi, il semble par exemple qu'il serait nécessaire de prendre en compte l'élargissement des partiels de

haute fréquence dû à des variations rapides de la fréquence fondamentale, ou bien à une forte variation de leur amplitude.

Les modèles d'attaques et plus généralement de sons percussifs n'ont été abordés que succinctement (section 5.1.1.3 page 114) : ainsi, il semble nécessaire de s'intéresser plus en détail à ce type de sons qui représente une partie non négligeable du contenu des signaux musicaux. Si le problème majeur rencontré était la grande variété des sons d'attaques qui rendait difficile l'utilisation d'un modèle d'atome bien défini, il semble que la structuration à base de modèles de Markov cachés puisse jouer un rôle prépondérant dans la modélisation des attaques, comme suggéré dans [Nakano *et al.*, 2010b].

Pour rester dans l'idée de l'utilisation de modèles de synthèse sonore simples, la structuration temporelle pourrait également être introduite en incluant dans la décomposition des modèles temporels inspirés de ces méthodes de synthèse : par exemple, il pourrait être intéressant d'introduire le très classique modèle d'enveloppe temporelle Attack Decay Sustain Release (ADSR) [Roads, 1996, p.97-98] dans la décomposition.

Problèmes fondamentaux des décompositions

L'estimation de l'ordre des modèles semble également un problème sur lequel il faudrait se pencher : dans les méthodes proposées, le nombre d'atomes à utiliser est fixé et un nombre optimal d'atomes ne peut pas être estimé (de même, dans la décomposition source/filtre du chapitre 3, l'ordre des filtres est fixé). Pourtant le nombre d'atomes utilisés dans une décomposition est généralement un élément fondamental pour obtenir une bonne représentation du spectrogramme décomposé. Les méthodes bayesiennes non-paramétriques semblent fournir un cadre théorique solide pour traiter ce problème comme le suggère [Nakano *et al.*, 2011].

Enfin, le problème de l'évaluation d'une décomposition ou d'une représentation semble également fondamental. Lorsque la décomposition est utilisée dans le cadre d'une application bien définie, on peut généralement comparer les performances du système pour cette application : par exemple pour une application de séparation de sources, le Rapport signal à distorsion (SDR), le Rapport signal à artefact (SAR) et le Rapport signal à interférence (SIR) (*cf.* [Vincent *et al.*, 2006]) permettent une comparaison objective (même si éventuellement sujette à caution) entre les différents systèmes. Lorsqu'il s'agit de comparer directement des algorithmes de décomposition indépendamment d'une application potentielle, le problème est fondamentalement plus complexe : il est en effet difficile de définir ce qu'est une bonne représentation indépendamment d'une utilisation de cette représentation. Le problème de l'évaluation d'une méthode de décomposition reste ainsi un sujet ouvert particulièrement intéressant.

Bibliographie

Publications de l'auteur

Articles de revue

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Processing Letters*, 18(2):83–86, Février 2011a.

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : NMF with time-frequency activations to model nonstationary audio events. *IEEE Transactions on audio, speech, and language processing*, 19(4):744–753, Mai 2011b.

Articles de conférence

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : NMF with time-frequency activations to model non stationary audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, Mars 2010a.

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : Spectral similarity measure invariant to pitch shifting and amplitude scaling. In *Congrès Français d'Acoustique*, Lyon, France, Avril 2010b.

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *International Conference On Digital Audio Effects*, pages 246–253, Graz, Autriche, Septembre 2010c.

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : Scale-invariant probabilistic latent component analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Octobre 2011d.

Romain HENNEQUIN, Bertrand DAVID et Roland BADEAU : Score informed audio source separation using a parametric model of non-negative spectrogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, République Tchèque, Mai 2011e.

Rapport technique

Romain HENNEQUIN, Roland BADEAU et Bertrand DAVID : Scale-invariant probabilistic latent component analysis. Rapport technique, Telecom-ParisTech, Mars 2011c.

Références

- Samer A. ABDALLAH et Mark D. PLUMBLEY : Unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on neural Networks*, 17(1):179–196, Janvier 2006.
- Eric ALLAMANCHE, Thorsten KASTNER, Ralf WISTORF, Nicolas LEFEBVRE et Juergen HERRE : Music genre estimation from low level audio features. In *Audio Engineering Society Conference*, Juin 2004. URL <http://www.aes.org/e-lib/browse.cfm?elib=12820>.
- Roland BADEAU, Nancy BERTIN et Emmanuel VINCENT : Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, Décembre 2010.
- Roland BADEAU et Bertrand DAVID : Weighted maximum likelihood autoregressive and moving average spectrum modeling. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3761–3764, Las Vegas, Nevada, USA, Mars 2008.
- Roland BADEAU, Valentin EMIYA et Bertrand DAVID : Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3073–3076, Taipei, Taiwan, Avril 2009.
- Juan P. BELLO, Laurent DAUDET et Mark B. SANDLER : Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):2242–2251, Novembre 2006.
- Michael W. BERRY et Murray BROWNE : Email surveillance using nonnegative matrix factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, Février 2005.
- Nancy BERTIN : *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique*. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, Octobre 2009.
- Nancy BERTIN, Roland BADEAU et Gaël RICHARD : Blind signal decompositions for automatic transcription of polyphonic music : NMF and K-SVD on the benchmark. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I-65–I-68, Honolulu, Hawaii, USA, Avril 2007.
- Nancy BERTIN, Roland BADEAU et Emmanuel VINCENT : Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music

transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):538–549, Février 2010.

Judith C. BROWN : Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434, Janvier 1991.

Jean-François CARDOSO : Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, Octobre 1998.

Scott Shaobing CHEN, David L. DONOHO, MICHAEL et A. SAUNDERS : Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

Moody CHU, Fasma DIELE, Robert PLEMMONS et Stefania RAGNI : Optimality, computation, and interpretation of nonnegative matrix factorizations. Rapport non publié, disponible à l'adresse <http://www4.ncsu.edu/~mtchu/Research/Papers/nnmf.pdf>, 2004.

Andrzej CICHOCKI, Sun ichi AMARI, Rafal ZDUNEK, Raul KOMPASS et Gen Hori : Extended SMART algorithms for non-negative matrix factorization. In *International Conference on Artificial Intelligence and Soft Computing*, volume 4029, pages 548–562, Zakopane, Pologne, Juin 2006a.

Andrzej CICHOCKI, Rafal ZDUNEK et Sun-Ichi AMARI : Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. In *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 32–39, Charleston, SC, USA, Mars 2006b.

Andrzej CICHOCKI, Rafal ZDUNEK, Seungjin CHOI, Robert J. PLEMMONS et Shun-Ichi AMARI : Non-negative tensor factorization using alpha and beta divergences. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1393–1396, Honolulu, Hawaii, USA, Avril 2007.

Andrzej CICHOCKI, Rafal ZDUNEK et Shun ichi AMARI : New algorithms for nonnegative matrix factorization in applications to blind source separation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 621–625, Toulouse, France, Mai 2006c.

Andrzej CICHOCKI, Rafal ZDUNEK, Anh Huy PHAN et Shun-Ichi AMARI : *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, 2009.

Joel E. COHEN et Uriel G. ROTHBLUM : Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993. ISSN 0024-3795. URL <http://www.sciencedirect.com/science/article/pii/002437959390224C>.

Pierre COMON : Independent component analysis, a new concept. *Signal Processing special issue Higher-Order Statistics*, 36:287–314, Avril 1994.

- Arshia CONT : Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. In *IEEE International Conference in Acoustics and Speech Signal Processing (ICASSP)*. Toulouse, Mai 2006. URL <http://cosmal.ucsd.edu/arshia/papers/ICASSP06/>.
- Roger B. DANNENBERG et Ning HU : Polyphonic audio matching for score following and intelligent audio editors. In *International Computer Music Conference*, pages 27–34, San Francisco, CA, USA, 2003.
- Inderjit S. DHILLON et Suvrit SRA : Generalized nonnegative matrix approximations with Bregman divergences. In Y. WEISS, B. SCHÖLKOPF et J. PLATT, éditeurs : *Neural Information Processing Systems conference (NIPS)*, pages 283–290. MIT Press, Cambridge, MA, USA, Décembre 2006.
- Sascha DISCH et Bernd EDLER : An amplitude and frequency modulation vocoder for audio signal processing. In *Conference on Digital Audio Effects (DAFx)*, pages 257–263, Espoo, Finlande, Septembre 2008.
- Sascha DISCH et Bernd EDLER : Multiband perceptual modulation analysis, processing and synthesis of audio signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Avril 2009.
- David DONOHO et Victoria STODDEN : When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian THRUN, Lawrence SAUL et Bernhard SCHÖLKOPF, éditeurs : *Advances in Neural Information Processing Systems 16*, Cambridge, MA, USA, 2004. MIT Press.
- Jean-Louis DURRIEU, Bertrand DAVID et Gaël RICHARD : A musically motivated mid-level representation for pitch estimation and musical audio source separation. *Selected Topics in Signal Processing, IEEE Journal of*, PP(99):1, 2011. ISSN 1932-4553.
- Jean-Louis DURRIEU, Alexey OZEROV, Cédric FÉVOTTE, Gaël RICHARD et Bertrand DAVID : Main instrument separation from stereophonic audio signals using a source/filter model. In *European Signal Processing Conference (EUSIPCO)*, pages 15–19, Glasgow, Royaume-Uni, Août 2009a.
- Jean-Louis DURRIEU, Gaël RICHARD et Bertrand DAVID : Singer melody extraction in polyphonic signals using source separation methods. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 169–172, Las Vegas, Nevada, USA, Août 2008.
- Jean-Louis DURRIEU, Gaël RICHARD et Bertrand DAVID : An iterative approach to monaural musical mixture de-soloing. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 105–108, Taipei, Taiwan, Avril 2009b.
- Jean-Louis DURRIEU, Gaël RICHARD, Bertrand DAVID et Cédric FÉVOTTE : Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, Novembre 2010.

Shinto EGUCHI et Yutaka KANO : Robustifying maximum likelihood estimation. Rapport technique, Institute of Statistical Mathematics, Tokyo, Juin 2001.

Valentin EMIYA, Roland BADEAU et Bertrand DAVID : Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, Août 2010. ISSN 1558-7916.

Slim ESSID : *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. Thèse de doctorat, Université Pierre et Marie Curie, Décembre 2006.

Mark EVERY et John SZYMANSKI : A spectral-filtering approach to music signal separation. In *International Conference on Digital Audio Effects*, pages 197–200, Naples, Italie, Octobre 2004.

Jeffrey A. FESSLER et Alfred O. HERO : Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, Octobre 1994.

Derry FITZGERALD, Matt CRANITCH et Eugene COYLE : Non-negative tensor factorisation for sound source separation. In *Irish Signals and Systems Conference*, Dublin, Irlande, Septembre 2005.

Derry FITZGERALD, Matt CRANITCH et Eugene COYLE : Shifted non-negative matrix factorisation for sound source separation. In *IEEE conference on Statistics in Signal Processing*, pages 1132–1137, Bordeaux, France, Juillet 2005.

J. L. FLANAGAN et R. M. GOLDEN : Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, Novembre 1966.

George E. FORSYTHE, Michael A. MALCOLM et Cleve B. MOLER : *Computer Methods for Mathematical Computations*. Prentice-Hall, 1976.

Benoit FUENTES, Roland BADEAU et Gaël RICHARD : Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, République Tchèque, Mai 2011.

Cédric FÉVOTTE, Nancy BERTIN et Jean-Louis DURRIEU : Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 11(3):793–830, Mars 2009.

Cédric FÉVOTTE et Jérôme IDIER : Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Septembre 2011. URL <http://arxiv.org/abs/1010.1763>.

Joachim GANSEMAN, Paul SCHEUNDERS, Gautham J. MYSORE et Jonathan S. ABEL : Evaluation of a score-informed source separation system. In *International Society for Music Information Retrieval Conference*, Utrecht, Pays-Bas, Août 2010a.

- Joachim GANSEMAN, Paul SCHEUNDERS, Gautham J. MYSORE et Jonathan S. ABEL : Source separation by score synthesis. *In International Computer Music Conference*, New York, NY, USA, Juin 2010b.
- Cyril GOBINET, Eric PERRIN et Régis HUEZ : Application of non-negative matrix factorization to fluorescence spectroscopy. *In European Signal Processing Conference (EUSIPCO)*, Vienne, Autriche, Septembre 2004.
- Rémi GRIBONVAL et Emmanuel BACRY : Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, Janvier 2003.
- Harold HOTELLING : Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):471–441, Septembre 1933.
- Patrik O. HOYER : Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, Novembre 2004.
- Fumitada ITAKURA et Shuzo SAITO : Analysis synthesis telephony based on the maximum likelihood method. *In 6th International Congress on Acoustics*, pages C–17–C–20, Tokyo, Japon, 1968.
- Cyril JODER, Slim ESSID et Gaël RICHARD : A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2385–2397, 2011.
- Tadeusz KACZOREK : *Positive 1D and 2D Systems*. Springer, 2002.
- Raul KOMPASS : A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, Mars 2007.
- Solomon KULLBACK et Richard A. LEIBLER : On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mars 1951.
- Jonathan LE ROUX : *Exploitation de régularités dans les scènes acoustiques naturelles*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie, 2009.
- Jonathan LE ROUX, Alain de CHEVEIGNÉ et Lucas C. PARRA : Adaptive template matching with shift-invariant semi-NMF. *In Daphne KOLLER, Dale SCHUURMANS, Yoshua BENGIO et Léon BOTTOU, éditeurs : NIPS*, pages 921–928. MIT Press, 2008a.
- Jonathan LE ROUX, Hirokazu KAMEOKA, Nobutaka ONO, Alain de CHEVEIGNÉ et Shigeki SAGAYAMA : Computational auditory induction by missing-data non-negative matrix factorization. *In ITRW on Statistical and Perceptual Audio Processing*, Brisbane, Australie, Septembre 2008b.
- Jonathan LE ROUX, Hirokazu KAMEOKA, Nobutaka ONO et Shigeki SAGAYAMA : Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. *In Proc. 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 397–403, Graz, Autriche, Septembre 2010.

Daniel D. LEE et H. Sebastian SEUNG : Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Octobre 1999.

Daniel D. LEE et H. Sebastian SEUNG : Algorithms for non-negative matrix factorization. In MIT PRESS, éditeur : *Advances in Neural Information Processing Systems*, volume 13, pages 556–562, 2000.

Pierre LEVEAU : *Décompositions parcimonieuses structurées : application à la représentation objet de la musique*. Thèse de doctorat, Université Pierre et Marie Curie, 2007.

Pierre LEVEAU, Emmanuel VINCENT, Gaël RICHARD et Laurent DAUDET : Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1):116–128, Janvier 2008.

Chih-Jen LIN : Projected gradient methods for non-negative matrix factorization. Rapport technique, Department of Computer Science, National Taiwan University, 2007.

Shoji MAKINO, Shoko ARAKI, Ryo MUKAI et Hiroshi SAWADA : Audio source separation based on independent component analysis. In *International Symposium on Circuits and Systems*, pages V–668–V–671, Vancouver, BC, Canada, Mai 2004.

Stéphane MALLAT et Zhifeng ZHANG : Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, Décembre 1993.

Eric MOULINES et Francis CHARPENTIER : Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, Décembre 1990.

Eric MOULINES et Jean LAROCHE : Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2):175–205, Février 1995.

Gautham J. MYSORE et Paris SMARAGDIS : Relative pitch estimation of multiple instruments. In *International Conference on Acoustics, Speech and Signal Processing*, pages 313–316, Taipei, Taiwan, Avril 2009.

Gautham J. MYSORE, Paris SMARAGDIS et Bhiksha RAJ : Non-negative hidden Markov modeling of audio with application to source separation. In *Latent Variable Analysis and Signal Separation*, St-Malo, France, Septembre 2010.

Masahiro NAKANO, Hirokazu KAMEOKA, Jonathan LE ROUX, Yu KITANO, Nobutaka ONO et Shigeki SAGAYAMA : Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence. In *IEEE International Workshop on Machine Learning for Signal Processing*, Kittilä, Finlande, Août 2010a.

Masahiro NAKANO, Jonathan LE ROUX, Hirokazu KAMEOKA, Yu KITANO, Nobutaka ONO et Shigeki SAGAYAMA : Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms. In *Latent Variable Analysis and Signal Separation*, St-Malo, France, Septembre 2010b.

- Masahiro NAKANO, Jonathan Le ROUX, Hirokazu KAMEOKA, Nobutaka ONO et Shigeki SAGAYAMA : Infinite-state spectrum model for music signal analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1972–1975, Prague, République Tchèque, Mai 2011.
- Laurent OUDRE, Yves GRENIER et Cédric FÉVOTTE : Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2222–2233, Septembre 2009.
- Alexey OZEROV et Cédric FÉVOTTE : Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563, 2010.
- Alexey OZEROV, Cédric FÉVOTTE et Maurice CHARBIT : Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Octobre 2009.
- Pentti PAATERO et Unto TAPPER : Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 (2):111–126, 1994.
- Mathieu PARVAIX et Laurent GIRIN : Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721–1733, Août 2011.
- Mathieu PARVAIX, Laurent GIRIN et Jean-Marc BROSSIER : A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Transactions on audio, speech, and language processing*, 18(6):1464–1475, Août 2010.
- V. Paul PAUCA, Farial SHAHNAZ, Michael W. BERRY et Robert J. PLEMMONS : Text mining using non-negative matrix factorizations. In *SIAM international conference on data mining*, pages 452–456, Lake Buena Vista, Floride, USA, Janvier 2004.
- Jouni PAULUS et Tuomas VIRTANEN : Drum transcription with non-negative spectrogram factorization. In *European Signal Processing Conference (EUSIPCO)*, Antalya, Turquie, Septembre 2005.
- Christopher RAPHAEL : Aligning music audio with symbolic scores using a hybrid graphical model. *Machine learning*, 65(2-3):389–409, Décembre 2006.
- Christopher RAPHAEL : A classifier-based approach to score-guided source separation of musical audio. *Computer Music Journal*, 32(1):51–59, Printemps 2008.
- François RIGAUD, Bertrand DAVID et Laurent DAUDET : A parametric model of piano tuning. In *International Conference On Digital Audio Effects*, Paris, France, Septembre 2011.
- Curtis ROADS : *The computer music tutorial*. the MIT press, 1996.

Eric D. SCHEIRER : Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, 103(1):588–601, Janvier 1998.

Mikkel N. SCHMIDT et Hans LAURBERG : Nonnegative matrix factorization with gaussian process priors. *Computational Intelligence and Neuroscience*, 2008:1–10, 2008.

Mikkel N. SCHMIDT et Morten MØRUP : Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, volume 3889 de *Lecture Notes in Computer Science (LNCS)*, pages 700–707, Paris, France, Avril 2006. Springer.

Christian SCHÖRKHUBER et Anssi Klapuri : Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conference*, Barcelone, Espagne, Juillet 2010.

Madhusudana V. SHASHANKA, Bhiksha RAJ et Paris SMARAGDIS : Sparse overcomplete latent variable decomposition of counts data. In *Neural Information Processing Systems*, Vancouver, BC, Canada, Décembre 2007.

Madhusudana V. SHASHANKA, Bhiksha RAJ et Paris SMARAGDIS : Probabilistic latent variable models as non-negative factorizations. in special issue on advances in non-negative matrix and tensor factorization. *special issue on Advances in Non-negative Matrix and Tensor Factorization, Computational Intelligence and Neuroscience Journal*, 2008:ID 947438, 2008.

Paris SMARAGDIS : Non-negative matrix factor deconvolution ; extraction of multiple sound sources from monophonic inputs. In *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 494–499, Grenade, Espagne, Septembre 2004.

Paris SMARAGDIS et Judith C. BROWN : Non-negative matrix factorization for polyphonic music transcription. In *Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, NY, USA, Octobre 2003.

Paris SMARAGDIS et Gautham J. MYSORE : Separation by humming : User-guided sound extraction from monophonic mixtures. In *Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 69–72, New Paltz, NY, USA, Octobre 2009.

Paris SMARAGDIS, Bhiksha RAJ et Madhusudana SHASHANKA : Supervised and semi-supervised separation of sounds from single-channel mixtures. In *7th International Conference on Independent Component Analysis and Signal Separation*, Londres, Royaume-Uni, Septembre 2007.

Paris SMARAGDIS, Bhiksha RAJ et Madhusudana SHASHANKA : Sparse and shift-invariant feature extraction from non-negative data. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2069–2072, Las Vegas, Nevada, USA, Mars 2008.

Paris SMARAGDIS, Bhiksha RAJ et Madhusudana SHASHANKA : Missing data imputation for spectral audio signals. In *IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, Septembre 2009.

Paris SMARAGDIS, Bhiksha RAJ et Madhusudana SHASHANKA : Missing data imputation for time-frequency representations of audio signals. *Journal of Signal Processing Systems*, Août 2010.

Paris SMARAGDIS, Madhusudana V. SHASHANKA et Bhiksha RAJ : Latent Dirichlet decomposition for single channel speaker separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, Mai 2006.

Emmanuel VINCENT, Nancy BERTIN et Roland BADEAU : Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):528–537, Mars 2010.

Emmanuel VINCENT, Rémi GRIBONVAL et Cédric FÉVOTTE : Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, Juillet 2006.

Tuomas VIRTANEN : Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, Mars 2007.

Tuomas VIRTANEN, A. Taylan CEMGIL et Simon GODSILL : Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1825–1828, Las Vegas, Nevada, USA, Avril 2008.

Jan WEIL, Thomas SIKORA, Jean-Louis DURRIEU et Gaël RICHARD : Automatic generation of lead sheets from polyphonic music signals. In *International Society for Music Information Retrieval Conference*, Kobe, Japon, Octobre 2009.

John WOODRUFF, Bryan PARDO et Roger DANNENBERG : Remixing stereo music with score-informed source separation. In *International Conference on Music Information Retrieval*, Victoria, BC, Canada, Octobre 2006.

Rafal ZDUNEK et Andrzej CICHOCKI : Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8):1904–1916, Août 2007.

Ye ZHANG et Yong FANG : A NMF algorithm for blind separation of uncorrelated signals. In *International Conference on Wavelet Analysis and Pattern Recognition*, pages 999–1003, Pékin, Chine, Novembre 2007.

Table des figures

1.1	Spectrogramme de l'introduction de <i>Godfather Waltz</i> par Nino Rota. En bas : zoom sur des parties du spectrogramme, mettant en lumière l'existence de répétitions.	17
1.2	Spectrogramme d'une note (<i>Do4</i>) de violon contenant un important vibrato.	19
2.1	Factorisation en matrices non-négatives.	24
2.2	Représentation graphique de la divergence de Bregman associée à la fonction F entre les points x et y	27
2.3	Extension du cône polyédrique : les flèches bleues pleines correspondent aux vecteurs colonnes de \mathbf{W} . La zone quadrillée est l'espace positivement engendré par ces vecteurs. Les croix violettes correspondent aux vecteurs colonnes de \mathbf{V} (à proximité de l'espace quadrillé). La flèche rouge en pointillé correspond à un vecteur obtenu par extension du cône.	29
2.4	Factorisation en matrices non-négatives d'un spectrogramme.	32
2.5	Distribution de $U = \frac{ 1+\mu e^{i\phi} ^\alpha}{1+\mu^\alpha}$ pour $\mu = 1$ et $\alpha = 2$: la variable prend ses valeurs dans l'intervalle $[0, 2]$	34
2.6	$\mathbb{E}\{(U - 1)^2\}$ en fonction de α et μ . La courbe rouge représente le α qui minimise $\mathbb{E}\{(U - 1)^2\}$ pour chaque μ	35
2.7	$\mathbb{E}\{ U - 1 \}$ en fonction de α et μ . La courbe rouge représente le α qui minimise $\mathbb{E}\{ U - 1 \}$ pour chaque μ . Remarque : la courbe rouge devrait être, en théorie, symétrique sur le graphe (la distribution de U est en effet inchangée en remplaçant μ par $\frac{1}{\mu}$). La légère asymétrie dans le graphe est due aux erreurs d'estimation.	36
2.8	Règle de mise à jour dans un algorithme MM.	46
2.9	Evolution de la divergence d'Itakura-Saito (IS) pour deux exposants différents dans les règles multiplicatives : pour $\eta = 0.5$, le critère est décroissant (il existe une preuve théorique), pour $\eta = 1.2$, le critère n'est pas décroissant (ce n'est pas visible sur le graphe, mais la première itération fait croître le critère).	47
2.10	Evolution de la divergence de Kullback-Leibler (KL) au cours des itérations pour trois types d'algorithmes de NMF : algorithme à mises jour multiplicatives, algorithme de Newton à pas optimal, algorithme de gradient projeté.	50

2.11	Décomposition invariante par translation temporelle du spectrogramme V (en-bas à droite) contenant des éléments percussifs (boucle de batterie). U ₁ , U ₂ et U ₃ (en-bas à gauche) sont les trois atomes temps/fréquence, correspondant chacun à un élément de batterie (respectivement, grosse caisse, charleston et caisse claire). Les activations H (en-haut à droite) doivent être très parcimonieuses pour que la décomposition ait de l'intérêt : ici les activations prennent effectivement une forme très impulsive au moment des attaques de l'élément de batterie correspondant.	52
2.12	Décomposition d'un son de guimbarde à l'aide d'une NMF : en haut à gauche, spectrogramme original ; les autres spectrogrammes sont des spectrogrammes reconstruits à partir de la NMF de ce spectrogramme original en utilisant des nombres d'atomes <i>R</i> différents.	55
2.13	Décomposition du spectrogramme d'une note contenant du vibrato (spectrogramme original en haut à gauche) à l'aide d'une NMF : pour <i>R</i> = 1 atome (en haut à droite), <i>R</i> = 3 atomes (en bas à gauche) et <i>R</i> = 10 atomes (en bas à droite).	56
3.1	Spectrogramme de puissance original de l'extrait de guimbarde 3.1(a) et spectrogrammes reconstruits 3.1(b), 3.1(c) et 3.1(d).	69
3.2	Factorisation source/filtre (<i>R</i> = 1, <i>Q</i> = 0 et <i>P</i> = 2) du spectrogramme de puissance du son de guimbarde.	69
3.3	Spectrogramme de puissance original de l'extrait de didgeridoo 3.3(a) et spectrogrammes reconstruits 3.3(b), 3.3(c) et 3.3(d).	71
3.4	Spectrogramme original de puissance de l'extrait de clavecin 3.4(a) et spectrogrammes reconstruits 3.4(b), 3.4(c) et 3.4(d).	72
3.5	Factorisation source/filtre (<i>R</i> = 2, <i>Q</i> = 1 et <i>P</i> = 1) du spectrogramme de puissance de l'extrait de clavecin.	73
3.6	Spectrogramme de puissance original de l'extrait de guitare électrique traité par une pédale wah-wah 3.6(a) et spectrogrammes reconstruits 3.6(b), 3.6(c) et 3.6(d).	74
3.7	Factorisation source/filtre (<i>R</i> = 3 et <i>P</i> = 2) du spectrogramme de puissance du son de guitare électrique traité par une pédale wah-wah.	75
3.8	Evolution de la fonction de coût (β -divergence avec β = 0.5) au cours des itérations (décomposition des extraits des sections 3.3.3 et 3.3.4).	76
4.1	Atome paramétrique $w_{fr}^{f_0^{rt}}$. Le carré du module de la transformée de Fourier de la fenêtre d'analyse <i>g</i> est représenté en vert pointillé.	79
4.2	Fonction de coût exprimée en fonction de la fréquence fondamentale f_0^{rt} de l'atome <i>r</i> : le spectre analysé est un mélange de deux spectres harmoniques de fréquence fondamentale 440Hz et 523Hz.	82
4.3	Lobe principal de <i>g</i> , dérivée de <i>g</i> et positivité de $P(f) = -\frac{g'(f)}{f}$ sur $\Lambda = [-\frac{2}{T}, \frac{2}{T}]$ pour une fenêtre de Hamming de longueur <i>T</i> . ⁴	84
4.4	Spectrogramme original de l'extrait du premier prélude de Bach.	87
4.5	Spectrogramme reconstruit de l'extrait du premier prélude de Bach.	88

4.6 Activations de la décomposition du spectrogramme de l'extrait du premier prélude de Bach sans contraintes et sans atomes de NMF standard. L'échelle de couleurs est en dB.	89
4.7 Activations de la décomposition du spectrogramme de l'extrait du premier prélude de Bach avec contrainte de décorrélation et de régularité spectrale et un atome non harmonique. L'échelle de couleur est en dB.	90
4.8 Représentation des activations incluant les fréquences fondamentales estimées (deux premières mesures du premier prélude de Bach).	90
4.9 Activations h_{rt} de la décomposition du spectrogramme de l'extrait du premier prélude de Bach joué par un piano.	91
4.10 Spectrogramme à Q-constant des premières notes d' <i>Au clair de la lune</i> jouées par un synthétiseur. Illustration de l'équivalence translation/transposition. .	96
4.11 Décomposition obtenue par PLCA invariante par translation fréquentielle du spectrogramme de la figure 4.10 : à gauche, motif fréquentiel P_K extrait, à droite la distribution d'impulsions P_I	97
4.12 Reconstruction du spectrogramme de la figure 4.10 à partir de P_K et P_I . .	98
4.13 Spectrogramme classique (issu d'une Transformée de Fourier à Court Terme (TFCT)) des premières notes d' <i>Au clair de la lune</i> jouées par un synthétiseur. Illustration de l'équivalence homothétie/transposition dans un tel spectrogramme (le motif translaté ne s'adapte pas correctement).	99
4.14 Spectrogramme original et spectrogramme reconstruit.	106
4.15 Distribution d'impulsions P_I de la gamme de <i>La</i> majeur.	107
4.16 Distribution noyau P_K de la gamme de <i>La</i> majeur.	108
4.17 Distribution d'impulsions P_I de l'introduction de <i>Because</i>	109
 5.1 Masques d'activation pour les 3 instruments d'un morceau : les activations de chaque source sont initialisées avec un <i>piano roll</i> binaire (0 ou 1) de la piste MIDI correspondante.	116
5.2 Estimation fine des masques temps/fréquence à partir des masques grossiers obtenus à l'initialisation.	118
5.3 Résultats en utilisant M1 comme méthode de synthèse.	120
5.4 Résultats en utilisant M2 comme méthode de synthèse.	121
5.5 Distribution d'impulsions P_I de l'introduction de <i>Because</i> des Beatles. .	124
5.6 Distribution d'impulsions P_I séparée en deux parties : la partie associée à la note à isoler 5.6(a) et la partie où cette note a été supprimée 5.6(b). .	125
5.7 Zoom sur les basses fréquences des premières secondes du spectrogramme 5.7(a) original et des spectrogrammes modèles 5.7(b) (total), 5.7(c) (note isolée) et 5.7(d) (reste) de l'introduction de <i>Because</i>	126
5.8 Spectrogramme original et spectrogramme reconstruit des deux notes de synthétiseurs.	129
5.9 Représentation des activations incluant la fréquence fondamentale estimée à chaque instant en demi-tons relatifs.	130
5.10 Activation temps/fréquence $h_{rt}(f)$ des atomes 5 et 8.	131

Liste des tableaux

5.1 Résultats en utilisant M1 comme méthode de synthèse.	120
5.2 Résultats en utilisant M2 comme méthode de synthèse.	121

Remerciements

Je tiens à remercier avant tout mes directeurs de thèse, Bertrand David et Roland Badeau, qui ont su m'orienter intelligemment et m'accompagner dans mon travail en faisant preuve d'une grande disponibilité malgré des emplois du temps souvent surchargés : Bertrand pour son optimisme permanent et Roland pour sa grande rigueur scientifique.

Je remercie l'ensemble des membres du jury pour l'intérêt qu'ils ont porté à mes travaux : les rapporteurs Laurent Daudet et Bruno Torrésani, le président du jury Eric Moulines, ainsi que Paris Smaragdis et Arshia Cont.

Je remercie toute les membres présents et passés de l'équipe Audiosig de TÉLÉCOM ParisTech pour leur bonne humeur, leur compétence et leur sérieux qui ont permis de réaliser cette thèse dans d'excellentes conditions : Gaël Richard, Yves Grenier, Slim Essid, Cédric Févotte, Thomas Fillon, Benoit Fuentes, Manuel Moussallam, Rémi Foucard, Antoine Liutkus, Sébastien Fenet, François Rigaud, Cyril Joder, Mounira Maazaoui, Nicolas Lopez, Benoît Mathieu, Félicien Vallet, Sébastien Gulluni, Angelique Dremeau, Aymeric Masurelle, Gaël Ladreyt, Nicolas Moreau, Laurent Oudre, Mathieu Lagrange, Jean-Louis Durrieu, Mathieu Ramona, Nancy Bertin, Valentin Emiya, Ester Cierco, Kristoffer Sundøy, Robin Tournemenne. Un grand merci également au personnel administratif de TÉLÉCOM ParisTech qui fait un excellent travail toujours dans la bonne humeur, en particulier Laurence Zelmar, Fabrice Planche, Sophie-Charlotte Barrière, Florence Besnard, Fabienne Lassausaie, Frédéric Boulanger et Karima Andreani.

Un grand merci à tous les membres de ma famille pour leur soutien : ma mère Patricia, mon père Bernard, ma soeur Laura, mon grand-père Paul-Louis ainsi qu'à ma grand-mère Marinette qui nous a malheureusement quitté pendant cette thèse.

Je remercie également mes amis, en particulier Les membres de Tak'One, Romain, Paul et Cb (et puis aussi un peu Manu mais il a déjà été remercié plus haut, alors...) pour repousser sans cesse un peu plus loin les limites du steak, et tous ses fans (et moins fans) : La Nouille, Jeanne, La Blasse, Yaya et Pimpin, Amélie et La Drille, Claire et La Serge, Juliette et Sébastien, Mariam et Martin, Arnaud et Auriane, Sam et Julie, Sam et Julie (les deux autres), Léni et Sophie, La Pompe et Samer (non pas la mère de la Pompe, Samer quoi!), La Peyre, Lucie, Louis, Yannick, Malmeu, Duchatte, la Coul, Armand, Augustin, Jacou, Y... Je remercie également tous les ATIAM de ma promotion : Marc, Lise, Tifanie, Sarah, Emilien, Sophie, Gaëtan, Baptiste, Maxime, Gonçal, Julien.

Enfin, je remercie Caroline, qui me supporte depuis maintenant plus de cinq ans, pour ses nombreuses et vaines tentatives de comprendre le sujet de ma thèse mais surtout pour son soutien et son réconfort sans faille.

Et je remercie bien évidemment tous les autres, tous ceux qui de près ou de loin m'auront apporté leur aide ou leur soutien pendant ces trois ans de travail et tous ceux qui ont lu cette page sans y trouver leur nom.
