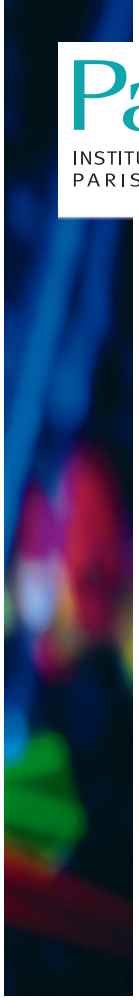
Reconstruction de phase par modèles de signaux : application à la séparation de sources audio

Paul Magron



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUT OF TECHNOLOGY

TELECOM ParisTech

2016-ENST-0078

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal & Images »

présentée et soutenue publiquement par

Paul MAGRON

le 2 décembre 2016

Reconstruction de phase par modèles de signaux : application à la séparation de sources audio

Directeur de thèse : **Roland BADEAU**
Co-encadrement de la thèse : **Bertrand DAVID**

Jury

M. Yannis STYLIANOU, Professeur, Université de Crète, Grèce	Examinateur
M. Philippe DEPALLE, Professeur, McGill University, QC, Canada	Rapporteur
M. Laurent GIRIN, Professeur, GIPSA-lab, Grenoble-INP, France	Rapporteur
M. Jonathan LE ROUX, Ingénieur de Recherche, MERL, MA, USA	Examinateur
M. Roland BADEAU, Maître de conférences, Télécom ParisTech, France	Directeur de thèse
M. Bertrand DAVID, Professeur, Télécom ParisTech, France	Co-encadrant de thèse

TÉLÉCOM ParisTech
École de l'Institut Mines-Télécom - Membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

Résumé

De nombreux traitements appliqués aux signaux audio travaillent sur une représentation Temps-Fréquence (TF) des données. Lorsque le résultat de ces algorithmes est un champ spectral d'amplitude, la question se pose, pour reconstituer un signal temporel, d'estimer le champ de phase correspondant. C'est par exemple le cas dans les applications de séparation de sources, qui estiment les spectrogrammes des sources individuelles à partir du mélange ; la méthode dite de filtrage de Wiener, largement utilisée en pratique, fournit des résultats satisfaisants mais est mise en défaut lorsque les sources se recouvrent dans le plan TF.

Cette thèse aborde le problème de la reconstruction de phase de signaux dans le domaine TF appliquée à la séparation de sources audio. Une étude préliminaire révèle la nécessité de mettre au point de nouvelles techniques de reconstruction de phase pour améliorer la qualité de la séparation de sources. Nous proposons de baser celles-ci sur des modèles de signaux. Notre approche consiste à exploiter des informations issues de modèles sous-jacents aux données comme les mélanges de sinusoïdes. La prise en compte de ces informations permet de préserver certaines propriétés intéressantes, comme la continuité temporelle ou la précision des attaques. Nous intégrons ces contraintes dans des modèles de mélanges pour la séparation de sources, où la phase du mélange est exploitée. Les amplitudes des sources pourront être supposées connues, ou bien estimées conjointement dans un modèle inspiré de la factorisation en matrices non-négatives complexe. Enfin, un modèle probabiliste de sources à phase non-uniforme est mis au point. Il permet d'exploiter les à priori provenant de la modélisation de signaux et de tenir compte d'une incertitude sur ceux-ci.

Ces méthodes sont testées sur de nombreuses bases de données de signaux de musique réalistes. Leurs performances, en termes de qualité des signaux estimés et de temps de calcul, sont supérieures à celles des méthodes traditionnelles. En particulier, nous observons une diminution des interférences entre sources estimées, et une réduction des artéfacts dans les basses fréquences, ce qui confirme l'intérêt des modèles de signaux pour la reconstruction de phase.

Abstract

A variety of audio signal processing techniques act on a Time-Frequency (TF) representation of the data. When the result of those algorithms is a magnitude spectrum, it is necessary to reconstruct the corresponding phase field in order to resynthesize time-domain signals. For instance, in the source separation framework the spectrograms of the individual sources are estimated from the mixture ; the widely used Wiener filtering technique then provides satisfactory results, but its performance decreases when the sources overlap in the TF domain.

This thesis addresses the problem of phase reconstruction in the TF domain for audio source separation. From a preliminary study we highlight the need for novel phase recovery methods. We therefore introduce new phase reconstruction techniques that are based on music signal modeling : our approach consists in exploiting phase information that originates from signal models such as mixtures of sinusoids. Taking those constraints into account enables us to preserve desirable properties such as temporal continuity or transient precision. We integrate these into several mixture models where the mixture phase is exploited ; the magnitudes of the sources are either assumed to be known, or jointly estimated in a complex nonnegative matrix factorization framework. Finally we design a phase-dependent probabilistic mixture model that accounts for model-based phase priors.

Those methods are tested on a variety of realistic music signals. They compare favorably or outperform traditional source separation techniques in terms of signal reconstruction quality and computational cost. In particular, we observe a decrease in interferences between the estimated sources and a reduction of artifacts in the low-frequency components, which confirms the benefit of signal model-based phase reconstruction methods.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

TF	Temps-Fréquence
TFCT	Transformée de Fourier à Court Terme
FFT	Transformée de Fourier rapide (de l'anglais *Fast Fourier Transform*)
NMF	Factorisation en matrices non-négatives (de l'anglais *Nonnegative Matrix Factorization*)
CNMF	Factorisation en matrices non-négatives complexes (de l'anglais *Complex Nonnegative Matrix Factorization*)
HRNMF	NMF à Haute Résolution (de l'anglais *High Resolution NMF*)
MMSE	Minimum au sens des moindres carrés (de l'anglais *Minimum Mean Square Error*)
AR	Autorégressif
KL	Kullback-Leibler
IS	Itakura-Saito
QIFFT	Interpolation quadratique de FFT (de l'anglais *Quadratic Interpolated FFT*)
GL	Griffin et Lim
MM	Majoration-Minimisation
ME	Majoration-Égalisation
EM	Espérance-Maximisation
VEM	Espérance-Maximisation variationnel
SDR	Rapport signal sur distorsion (de l'anglais *Signal to Distortion Ratio*)
SIR	Rapport signal sur interférences (de l'anglais *Signal to Interference Ratio*)
SAR	Rapport signal sur artéfacts (de l'anglais *Signal to Artifact Ratio*)

Notations

x, x_k	Signaux temporels
X, X_k	Matrices (domaine Temps-Fréquence)
$\|.\|$, $\angle(.)$	Modules et arguments complexes (scalaires ou matrices)
$\bar{.}$	Conjugaison complexe (scalaires ou matrices)
$A(f,t)$	Coefficient d'indice (f,t) de la matrice A
A^T	Matrice transposée de A
A^H	Matrice transposée-conjuguée de A
AB	Produit matriciel conventionnel : $(AB)(f,t) = \sum_r A(f,r)B(r,t)$
$A \odot B$	Produit matriciel terme-à-terme
$\frac{A}{B}$	Division matricielle terme-à-terme
$A^{\odot B}$	Puissance matricielle terme-à-terme
$\|\|A\|\|_p$	norme p : $(\sum_{f,t} \|A(f,t)\|^p)^{\frac{1}{p}}$

$\sim$	Suit une loi
$\mathcal{N}$	Loi normale (réelle ou complexe)
$\mathcal{VM}$	Loi de Von Mises
$\mathcal{L}$	Loi de Lévy
$\mathcal{P}\alpha\mathcal{S}$	Loi Positive α-stable

$\mathcal{F}$	Composition des opérateurs TFCT inverse et TFCT
$[\![a,b]\!]$	ensemble des entiers compris entre a et b
$\overset{c}{=}$	Égalité à une constante additive près

Chapitre 1

Introduction

Sommaire

1.1 Contexte général

Le traitement des signaux audio a connu ces dernières décennies un essor considérable, et trouve aujourd'hui des applications dans de très nombreux domaines. En effet, du rehaussement de la parole dans les télécommunications à des appareils médicaux tels que les prothèses auditives, les applications de ce champ de recherche sont extrêmement variées.

Une branche de ce domaine est consacrée à la manipulation des signaux musicaux. Systèmes de recommandation comme *Spotify* [1], reconnaissance de morceaux de musique comme la célèbre application *Shazam* [2], logiciels qui éditent une partition à partir d'un enregistrement de musique : tous ces outils ont en commun de travailler sur ce même matériau brut. Les applications de la recherche en traitement du signal musical sont donc à destination du grand public, mais aussi des musiciens et des mélomanes, ainsi qu'à des industriels du cinéma, de la télévision et de la musique.

Un thème principal de ces recherches consiste en l'extraction automatique d'informations de nature musicale (on parle alors de MIR pour *Music Information Retrieval*). Il peut s'agir d'informations comme le rythme, la tonalité, les instruments et les accords présents dans un morceau, ou encore le genre musical. De telles données peuvent ensuite être utilisées pour mettre au point des systèmes de recommandation, des transcripteurs automatiques de musique ou bien des outils de détection de reprise ou de plagiat.

Une autre application du traitement du signal musical est la séparation de sources, qui vise à reproduire automatiquement la faculté de l'oreille humaine à se focaliser uniquement sur un instrument et à dissocier les différents flux musicaux présents dans un morceau. Un système de séparation de sources prend un morceau de musique en entrée, et fournit en sortie un ensemble de pistes assignées chacune à un instrument. En quelque sorte, il s'agit de l'opération inverse du mixage qui est effectué lors de l'enregistrement du morceau. Une telle séparation a de nombreuses applications en musique : débruitage des signaux, mixage augmenté et respatialisation (typiquement sur de vieux enregistrements), ou encore génération automatique d'accompagnement (karaoké).

Ces deux thèmes de recherche sont étroitement liés : la séparation de sources aide à la transcription automatique ou à la reconnaissance d'instruments (il est toujours plus simple d'opérer sur des sources isolées plutôt que sur des mélanges), et les informations de nature musicale comme le rythme ou le contenu mélodique aident à l'identification et à la séparation des pistes instrumentales. On parle alors de séparation de sources *informée* lorsque des informations additionnelles sur le contenu musical sont prises en compte, à l'inverse de la séparation *aveugle*.

1.2 Reconstruction de phase

1.2.1 Représentation Temps-Fréquence

De nombreuses méthodes de traitement du signal agissent dans le domaine Temps-Fréquence (TF), c'est-à-dire sur une représentation des signaux musicaux qui rend compte aussi bien du contenu fréquentiel du signal que de son évolution temporelle. Intuitivement, on peut voir une partition de musique comme une représentation TF : les notes sont organisées en fonction de leur hauteur selon un axe vertical (information fréquentielle) et en fonction de leur ordre d'apparition selon un axe horizontal (information temporelle). De telles représentations de signaux musicaux sont généralement *parcimonieuses*, c'est-à-dire qu'un grand nombre de

1. https://www.spotify.com/fr/
2. http://www.shazam.com

leurs coefficients ont une valeur faible ou nulle, et que l'énergie est concentrée dans un nombre réduit de points TF. Cette propriété de parcimonie est fréquemment exploitée en traitement du signal musical Abdallah et Plumbley (2006).

Dans ce manuscrit, nous utilisons la Transformée de Fourier à Court Terme (TFCT), qui est présentée dans l'Annexe A. La TFCT d'un signal $x \in \mathbb{R}^N$ est une matrice à valeurs complexes $X \in \mathbb{C}^{F \times T}$, où F désigne le nombre de canaux fréquentiels de la transformée, et T le nombre de trames temporelles. Le module et l'argument de X sont respectivement appelés *spectrogrammes d'amplitude*[3] et de *phase* de X.

1.2.2 Position du problème

Dans de nombreuses applications, on considère que l'amplitude contient toute l'information utile et on néglige l'importance de la phase. Par exemple, en traitement de la parole, on a durant plus de deux décennies considéré que la phase n'apportait aucune information supplémentaire à l'amplitude en termes d'intelligibilité Wang et Lim (1982). De nombreux traitements dont le but est d'extraire de l'information musicale, telle que le rythme, la tonalité ou le genre, agissent donc sur une quantité positive obtenue à partir de la TFCT, telle que son spectrogramme d'amplitude $|X|$ ou de puissance $|X|^{\odot 2}$ (où $.^{\odot}$ désigne la puissance matricielle terme à terme).

Néanmoins, des applications comme le débruitage Mowlaee et al. (2012), la séparation de sources Wang et Plumbley (2005), ou encore la modification de hauteur Laroche et Dolson (1999), visent à resynthétiser des signaux dans le domaine temporel. On effectue donc un certain nombre d'opérations sur une TFCT X afin d'en construire une nouvelle Y (ou plusieurs, dans le cas de la séparation de sources), puis on applique la TFCT inverse à Y pour synthétiser un signal temporel. Cette matrice Y doit donc comporter une information d'amplitude et une information de phase. Lorsque l'on n'effectue des traitements que sur une quantité positive extraite de X, l'information de phase est généralement perdue et il devient nécessaire de reconstruire celle de Y.

En traitement du signal audio, une attention particulière est donc portée à la phase de la TFCT dans des applications telles que l'étirement temporel Laroche et Dolson (1999), le rehaussement Mowlaee et al. (2016) ou encore la synthèse Stylianou (2001) de la parole. Le problème de la reconstruction de phase de TFCT dépasse par ailleurs le cadre de l'audio : c'est en effet une tâche qui concerne de nombreux domaines de la physique appliquée Jaganathan et al. (2015); Eldar et al. (2016) tels que l'optique Gerchberg et Saxton (1972), la cristallographie ou encore l'imagerie médicale.

Le travail que nous présentons ici s'inscrit dans ce cadre. Les méthodes usuelles de reconstruction de phase ne conduisent en effet pas toujours à des résultats de qualité. Aussi, nous proposons de nouvelles techniques basées sur les propriétés spécifiques des signaux musicaux, qui permettent une réduction des artéfacts dans les signaux estimés.

1.3 Application à la séparation de sources

La reconstruction de phase est un problème important en séparation de sources, qui est l'application principale de cette thèse.

3. En toute rigueur, le terme "spectrogramme" désigne une densité spectrale de puissance, c'est-à-dire le carré du module de X.

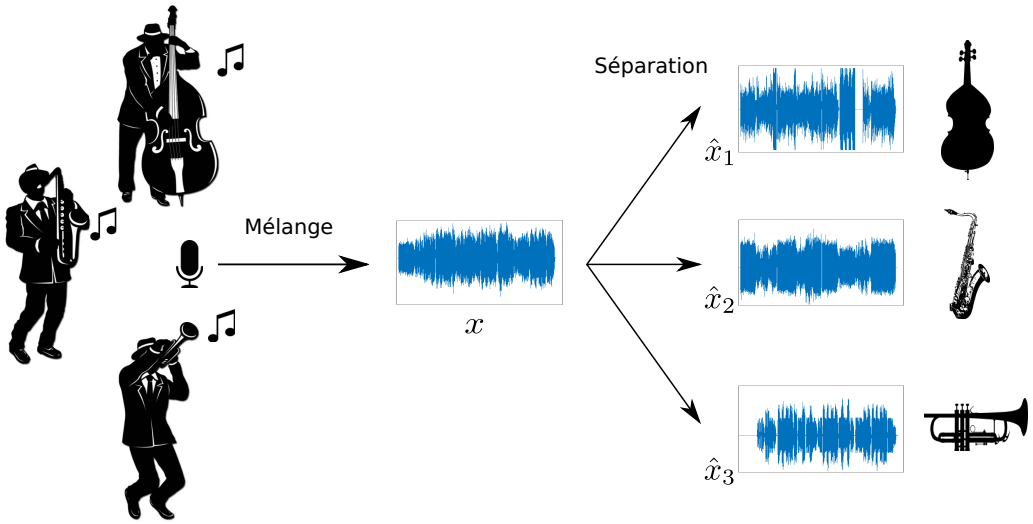


FIGURE 1.1 – Illustration du problème de séparation de sources : à partir d'un morceau de musique, on cherche à estimer chaque piste instrumentale isolée.

1.3.1 Mélange de sources

Considérons un modèle de mélange de sources *linéaire instantané monocanal* : les observations sont constituées d'un seul signal de mélange (contrairement au cas multicanal, comme pour des mélanges stéréo), et on ne s'intéresse pas aux phénomènes convolutifs tels que la réverbération. Ainsi, pour chaque instant n, le signal de mélange $x(n)$ est égal à la somme de K signaux sources notés x_k :

$$x(n) = \sum_{k=1}^{K} x_k(n). \tag{1.1}$$

Le problème de la séparation de sources consiste à obtenir, $\forall k \in [\![1, K]\!]$ [4], un estimateur $\hat{x}_k$ des sources x_k, comme c'est illustré sur la figure 1.1. On parle de problème *sous-déterminé* car il y a plus d'inconnues (sources) que de données disponibles (mélanges observés).

D'autres modèles de mélange existent, notamment le modèle multicanal convolutif LEGLAIVE et al. (2016b); OZEROV et FÉVOTTE (2010). Dans ce cadre, les phénomènes de salle (tels que la réverbération) sont pris en compte. Considérer un mélange multicanal suppose l'accès à davantage d'observations, et implique de devoir inclure les paramètres de mixage dans le modèle.

En appliquant une transformation telle que la TFCT au modèle (1.1), on obtient le modèle de mélange suivant dans le domaine TF :

$$\forall (f,t),\ X(f,t) = \sum_{k=1}^{K} X_k(f,t). \tag{1.2}$$

On cherche donc une estimation $\hat{X}_k$ des TFCT X_k des différentes sources afin de pouvoir resynthétiser des signaux temporels $\hat{x}_k$. Cette approche est illustrée sur la figure 1.2.

4. Dans tout ce manuscrit, $[\![a, b]\!]$ désigne un intervalle d'entiers, c'est-à-dire l'ensemble des entiers compris entre a et b.

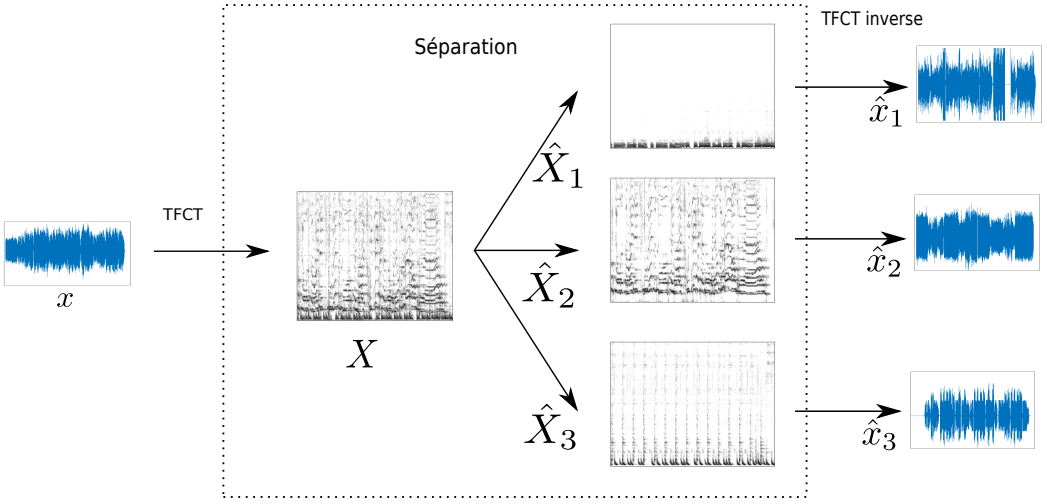


FIGURE 1.2 – Séparation de sources dans le domaine Temps-Fréquence.

1.3.2 Méthodes de séparation de sources

Pour estimer les composantes complexes $\hat{X}_k$, il est usuel d'appliquer un masque G_k à valeurs non-négatives à la TFCT du mélange :

$$\hat{X}_k = G_k \odot X, \tag{1.3}$$

où $\odot$ désigne la multiplication matricielle terme à terme. Le filtrage de Wiener, très employé dans ce contexte FÉVOTTE et al. (2009), consiste à fabriquer un tel masque à partir d'estimations des spectrogrammes de puissance des sources : la séparation de sources se ramène alors à un problème d'estimation de spectrogrammes d'amplitude ou de puissance.

De nombreuses méthodes de séparation de sources dans le domaine TF sont basées sur une hypothèse de redondance des signaux musicaux. En effet, ces signaux sont composés d'évènements (tels que les notes ou les frappes de batterie) qui se répètent au cours du temps, répétitions auxquelles notre oreille est sensible. L'idée qui consiste à exploiter cette redondance est à la base de diverses méthodes :

- L'analyse en composantes principales (PCA pour *Principal Component Analysis*) HUANG et al. (2012) ;
- L'analyse en composantes probabilistes latentes (PLCA pour *Probabilistic Latent Component Analysis*) SMARAGDIS et al. (2006) ;
- Les factorisations en matrices non-négatives (NMF pour *Nonnegative Matrix Factorization*) LEE et SEUNG (1999).

Nous nous sommes plus particulièrement intéressés aux approches NMF car celles-ci fournissent une factorisation des spectrogrammes audio qui fait sens : un des facteurs est un dictionnaire d'atomes spectraux et l'autre est une matrice d'activations temporelles. Les méthodes NMF fournissent donc une estimation des spectrogrammes d'amplitude des sources à partir desquels on peut calculer un masque G_k.

1.3.3 Verrous

L'application d'une méthode de masquage (1.3) telle que le filtrage de Wiener revient à donner la phase du mélange X à chacune des sources estimées $\hat{X}_k$. Or, l'hypothèse que les sources ont pour phase celle du mélange n'est pas vérifiée lorsque les sources se recouvrent

dans le domaine TF (ce qui est fréquent pour les signaux de musique). Même dans le cas où les spectrogrammes d'amplitude sont parfaitement estimés, ce filtrage introduit des interférences entre les sources estimées, et des artéfacts dans les basses fréquences Vincent (2010).

Des méthodes alternatives de reconstruction de phase reposent sur l'exploitation de propriétés mathématiques de la TFCT Sun et Smith (2012). C'est le cas des approches qui utilisent la *consistance* de la TFCT Griffin et Lim (1984); Le Roux et al. (2008c) et des méthodes combinant consistance et masquage Le Roux et Vincent (2013); Sturmel et Daudet (2013). Ces méthodes ne suppriment cependant pas complètement les artéfacts dans les signaux estimés. En outre, elles n'exploitent pas les propriétés spécifiques des signaux musicaux.

1.4 Approche et présentation du manuscrit

1.4.1 Approche retenue

Le premier objectif de cette thèse est d'évaluer l'impact de la phase sur la qualité des signaux reconstruits dans le cadre de la séparation de sources. Une étude comparative des principales techniques de reconstruction de phase dans les méthodes de séparation de sources basées sur la NMF a conduit à deux conclusions. Tout d'abord, les techniques couramment utilisées dans la littérature ne produisent pas toujours des signaux satisfaisants du point de vue de la qualité audio : on observe notamment des interférences entre sources estimées et des artéfacts dans les signaux reconstruits. Il est donc indispensable de s'intéresser à de nouvelles méthodes de reconstruction de phase, car celle-ci limite la qualité de la séparation de sources. En outre, les méthodes les plus prometteuses sont celles qui utilisent une structuration des représentations TF basée sur la modélisation des signaux audio.

Ces conclusions ont défini l'axe de recherche de cette thèse : la reconstruction de phase dans le domaine TF par modèles de signaux audio. Notre approche consiste à utiliser les propriétés qui découlent de modèles tels que les mélanges de sinusoïdes et d'impulsions, et les spécificités observées dans de nombreux signaux de musique (telle que la redondance des évènements audio). Ces propriétés fournissent des contraintes pour structurer la phase dans le domaine TF.

Ces méthodes sont utilisées pour la reconstruction *aveugle* de phase (par exemple en restauration audio), ou intégrées dans le cadre de la séparation de sources (où la phase du mélange peut être exploitée), qui est l'application principale de cette thèse. Nous considérons aussi bien le cas de la seule estimation des phases (les amplitudes sont supposées connues) que le cas de l'estimation conjointe des amplitudes et des phases, dans un modèle de type NMF complexe Kameoka et al. (2009).

Nous constatons expérimentalement que l'incorporation d'informations sur la phase basées sur les modèles de signaux améliore la qualité de la séparation de sources comparativement aux approches usuelles, qui n'utilisent pas de tels à prioris. En particulier, la contrainte de phase issue d'un modèle sinusoïdal réduit les artéfacts et les interférences entre sources, et le modèle basé sur la redondance des évènements audio améliore perceptivement la précision des attaques des sources.

1.4.2 Structure du document

Ce manuscrit est organisé comme suit. Tout d'abord, une première partie est consacrée à un état de l'art et à une étude comparative des méthodes usuelles de reconstruction de phase pour la séparation de sources.

— Le chapitre 2 présente ces méthodes ainsi qu'un certain nombre de rappels sur la NMF.

— Ces différentes méthodes sont comparées expérimentalement dans le chapitre 3 afin d'en identifier le potentiel et les limites.

Les conclusions de cette première partie ont défini l'orientation de notre travail pour la suite de la thèse : la reconstruction de phase par modèles de signaux, qui fait l'objet de la deuxième partie de ce manuscrit. Nous proposons notamment d'incorporer ces modèles de phase dans des méthodes de séparation de sources.

— Le chapitre 4 présente la méthode de déroulé linéaire des phases à partir de l'analyse de mélanges de sinusoïdes, et propose une première application qui est la restauration d'enregistrements corrompus par des craquements.

— Le chapitre 5 présente une technique itérative de séparation de sources qui utilise ce modèle de déroulé linéaire.

— Le chapitre 6 introduit un modèle de phase basé sur la répétition d'évènements audio pour la reconstruction des phases dans les trames d'attaque.

— Le chapitre 7 décrit un modèle de NMF complexe avec les contraintes de phase précédemment établies.

Dans une troisième partie, nous mettons au point des modèles probabilistes de sources, avec pour objectif de modéliser la phase de façon non-uniforme, et les amplitudes de façon robuste.

— Dans le chapitre 8, la phase est considérée comme étant une variable aléatoire de Von Mises, ce qui permet d'inclure un à priori basé sur le modèle de déroulé linéaire.

— Le chapitre 9 s'intéresse à la famille de lois positives α-stables qui modélisent de façon robuste des données non-négatives telles que les spectrogrammes d'amplitude.

Enfin, un résumé des contributions de cette thèse, ainsi que quelques perspectives de recherche future se trouvent dans le chapitre 10.

Première partie

État de l'art

Chapitre 2

Reconstruction de phase dans les approches NMF

Sommaire

Ce chapitre dresse un état de l'art des méthodes de reconstruction de phase combinées aux approches NMF pour la séparation de sources audio dans le domaine TF. Nous présentons dans la section 2.1 les principales techniques de reconstruction de phase qui sont employées dans le domaine du traitement du signal audio, notamment dans le cadre de la séparation de sources. Étant donné que ces méthodes nécessitent l'estimation préalable d'un spectrogramme d'amplitude pour chaque source, nous effectuons dans la section 2.2 une présentation générale de la NMF. Puis, dans la section 2.3, nous introduisons les méthodes de NMF complexe et de NMF à haute résolution, dont le but est de procéder à l'estimation conjointe des amplitudes et des phases. La section 2.4 présente les principaux indicateurs utilisés pour quantifier la qualité de la séparation de sources. Enfin, nous résumons dans la section 2.5 les principaux verrous scientifiques de ces approches et motivons ce travail de thèse.

2.1 La reconstruction de phase

2.1.1 Importance de la phase

La question de l'importance perceptive de la phase est sujette à débat. Dans Wang et Lim (1982), les auteurs ont mesuré l'impact du spectrogramme et de la phase sur la qualité du rehaussement de la parole, et en ont déduit que la phase jouait un rôle mineur comparé au spectre d'amplitude. Le cadre expérimental était restreint (à des paramètres de longueur et type de fenêtre, rapport signal sur bruit et mesure d'évaluation précises), ce qui limitait la portée de ces conclusions. Dans Ephraim et Malah (1984), les auteurs ont montré qu'utiliser la phase du signal de parole bruitée conduisait à l'obtention d'un estimateur optimal au sens des moindres carrés (*cf.* section suivante) du signal de parole non bruité. Ainsi, durant de nombreuses années, la reconstruction de phase n'a pas été considérée comme un thème majeur d'investigation.

Les études plus récentes conduites dans Paliwal et Alsteris (2003, 2005); Shannon et Paliwal (2006); Alsteris et Paliwal (2006, 2007) mettent en lumière l'importance de la phase en matière d'intelligibilité des signaux de parole. Les auteurs montrent qu'un choix judicieux des paramètres de la transformée (taux de recouvrement, longueur de la fenêtre...) permet d'exploiter l'information de phase pour le débruitage de signaux de parole. Les études Paliwal et al. (2011); Gerkmann et al. (2012) montrent également l'impact de la phase sur la qualité globale de reconstruction de signaux de parole, et la nécessité de mettre au point de nouvelles méthodes pour sa reconstruction.

Dans Gaich et Mowlaee (2015); Koutsogiannaki et al. (2014), il est montré qu'une métrique utilisant l'information de phase rend mieux compte des observations subjectives en matière d'intelligibilité de la parole qu'une métrique ne tenant compte que de l'information d'amplitude. La technique de *randomisation* de phase Sugiyama et Miyahara (2013a), qui confère à la phase un caractère aléatoire dans les points TF correspondant à certain bruits (comme des craquements), améliore la qualité du débruitage de signaux par rapport à une approche basée sur la seule amplitude.

En termes de séparation de sources musicales, nous avons soulevé la question de l'importance de la phase par une nouvelle étude, qui fait l'objet du chapitre 3. Nous y montrons notamment que le choix de la méthode de reconstruction de phase dans une approche de séparation de sources basée sur la NMF peut significativement altérer les résultats. Cette conclusion fait écho à celles de précédentes études sur le sujet, comme Mowlaee et Martin (2012), où il est montré qu'un estimateur des sources utilisant une information de phase améliore la qualité de la séparation par rapport à un estimateur ne la prenant pas en compte.

2.1.2 Masque temps-fréquence

Dans le cas de mélanges de plusieurs sources, l'approche communément employée dans la littérature pour estimer les composantes complexes $\hat{X}_k$ consiste à appliquer un masque G_k à la TFCT du mélange X :

$$\hat{X}_k = G_k \odot X, \tag{2.1}$$

où $\odot$ désigne la multiplication terme à terme. On peut considérer un masque binaire : $G_k = \{0,1\}^{F\times T}$. La source complexe reconstruite est alors égale au mélange dans certains points TF, et est nulle dans les autres Yilmaz et Rickard (2004). Ce masquage est efficace lorsqu'il n'y a pas de recouvrement des sources dans le domaine TF. Sur des mélanges réalistes, il produit des artéfacts auditifs, la binarité du masque créant des discontinuités dans les signaux reconstruits.

En pratique, on utilise plutôt un masquage *doux* $G_k \in [0,1]^{F\times T}$. Le filtrage de Wiener Wiener (1949), fréquemment employé (voir par exemple Févotte et al. (2005)) consiste à utiliser le masque suivant, appelé *gain de Wiener* et calculé à partir d'estimations $\hat{V}_k^{\odot 2}$ des spectrogrammes de puissance des sources :

$$G_k = \frac{\hat{V}_k^{\odot 2}}{\sum_{l=1}^{K} \hat{V}_l^{\odot 2}}. \tag{2.2}$$

Il s'agit d'un estimateur MMSE (optimal au sens des moindres carrés, de l'anglais *Minimum Mean Square Error*). C'est par exemple montré dans Ephraim et Malah (1984) pour des processus aléatoires gaussiens. C'est pourquoi cette approche est depuis longtemps utilisée dans la littérature, et que l'on cherche à obtenir une estimation des spectrogrammes de puissance des sources. D'autres méthodes agissent sur les spectrogrammes d'amplitude, aussi certains estimateurs de sources utilisent des masques similaires à (2.2) construits à partir d'estimations des amplitudes $\hat{V}_k$ plutôt que des puissances $\hat{V}_k^{\odot 2}$ Virtanen (2007). Un cadre théorique est fourni dans Liutkus et Badeau (2015) pour justifier l'utilisation de spectrogrammes fractionnaires pour obtenir un estimateur des X_k (filtrage de Wiener généralisé), dans le cas où les sources sont des variables aléatoires α-stables Nolan (2015).

Notons que le masquage TF n'est pas une technique de reconstruction de phase, il s'agit d'une méthode d'estimation des composantes complexes à partir du mélange X qui implique que la phase de chaque source estimée est égale à celle du mélange.

Cette approche présente l'avantage d'être rapide, simple à mettre en oeuvre, et de donner de bons résultats lorsque les sources se recouvrent faiblement dans le domaine TF. Lorsque le recouvrement est plus important, la propriété d'additivité des spectrogrammes n'est plus vérifiée, et la phase du mélange n'est pas égale à celles des sources. Illustrons cette limite par un exemple simple. Considérons un mélange composé de deux signaux synthétiques qui sont des sommes de sinusoïdes amorties. Les sources sont observées successivement seules, puis activées simultanément. Leurs fréquences sont choisies de sorte à observer un phénomène de battements dans certains canaux lorsque les deux sources sont activées simultanément. Le signal est échantillonné à 11025 Hz et la TFCT du mélange est calculée avec une fenêtre de Hann de longueur 512 échantillons (soit 46 ms) et 75 % de recouvrement.

On suppose connus les spectrogrammes de puissance des deux sources et on applique le filtrage de Wiener afin de reconstruire les composantes complexes. La figure 2.1 illustre alors l'effet du filtrage de Wiener dans la bande de fréquences correspondant à 730 Hz. Cette figure montre l'incapacité du filtrage de Wiener à estimer convenablement une composante complexe à partir du mélange en cas de recouvrement. Dans ce cas, le phénomène de battements persiste dans les sources séparées.

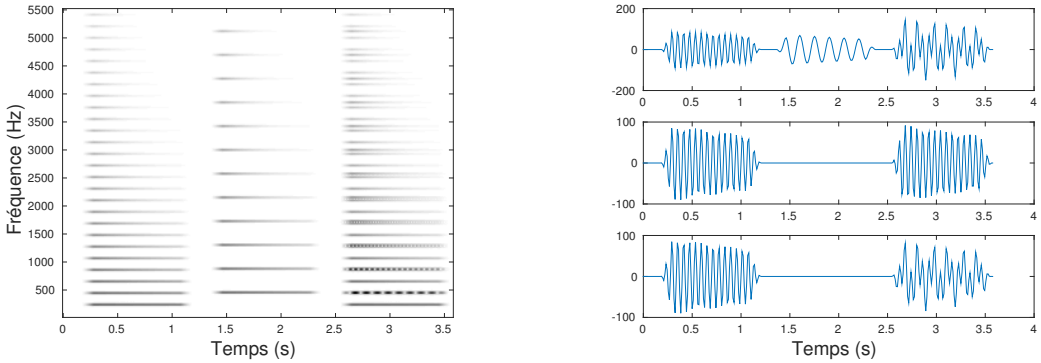


FIGURE 2.1 – Spectrogramme d'un mélange constitué de deux sources synthétiques (à gauche), et parties réelles de diverses composantes dans la bande de fréquences 730 Hz : mélange (en haut à droite), première source originale (au milieu à droite) et première source estimée par filtrage de Wiener (en bas à droite).

Le phénomène de recouvrement TF étant très fréquemment observé dans les mélanges de signaux musicaux (sources en relations harmoniques), il apparait nécessaire de trouver de nouvelles méthodes de reconstruction de phase pour l'estimation des composantes complexes dans le plan TF afin de synthétiser des signaux temporels de plus haute qualité. En outre, l'application du filtrage de Wiener dans ces points TF où les sources se recouvrent modifie les amplitudes de celles-ci, même si elles sont initialement supposées connues.

Notons enfin que le filtrage de Wiener peut également conduire à produire certains artéfacts dans les basses fréquences (notamment lorsque les signaux sont une basse et une batterie). Des méthodes de lissage de filtres de Wiener VINCENT (2010) peuvent alors être envisagées pour réduire ces artéfacts, mais cela ne supprime néanmoins pas les interférences entre sources.

2.1.3 Approches consistantes

La *consistance*, que nous définissons ci-après, désigne une propriété de la TFCT, indépendamment de la nature des signaux considérés. C'est en ce sens que nous l'entendrons dans le reste de ce manuscrit. Il existe des méthodes de reconstruction de phase qui sont basées sur la minimisation d'une fonction de coût qui pénalise l'*inconsistance* (ou, de façon équivalente, favorise la consistance).

Notion de consistance

Le concept de consistance LE ROUX et al. (2008c) est basé sur le fait que la TFCT n'est pas une transformation surjective de $\mathbb{R}^N$ dans $\mathbb{C}^{F\times T}$. En effet, toute matrice complexe n'est pas forcément la TFCT d'un signal réel. L'opérateur $\mathcal{F} = TFCT \circ TFCT^{-1}$ n'est pas la fonction identité dans $\mathbb{C}^{F\times T}$. On dit alors d'une matrice complexe qu'elle est *consistante* si elle est exactement la TFCT d'un signal [1]. Formellement, on définit alors l'espace des matrices consistantes comme étant l'ensemble image de l'opérateur de TFCT. L'application $\mathcal{F}$ est un projecteur sur le sous-espace des matrices consistantes.

La fonction d'*inconsistance* mesure l'écart entre une matrice complexe X et la TFCT de sa TFCT inverse. On définit la matrice d'inconsistance $I_X \in \mathbb{C}^{F\times T}$:

$$I_X = X - \mathcal{F}(X), \tag{2.3}$$

1. On dit également par extension qu'un spectrogramme (d'amplitude) est consistant s'il est égal au module d'une matrice consistante.

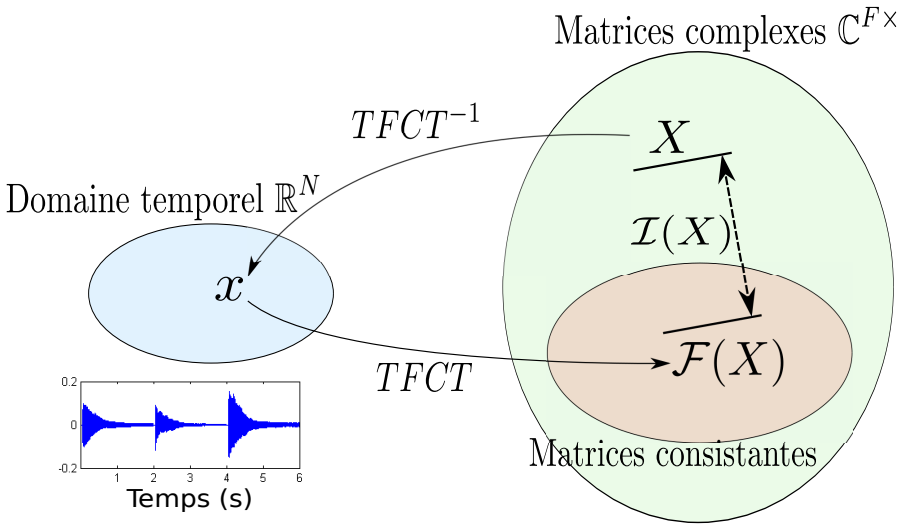


FIGURE 2.2 – Illustration de la notion de consistance : une matrice complexe quelconque n'est pas nécessairement égale à la TFCT de sa TFCT inverse. L'écart entre les deux est appelé inconsistance.

Algorithme 1 Griffin et Lim

Entrées :

Spectrogramme $V \in \mathbb{R}_+^{F\times T}$, phase initiale $\phi \in [0, 2\pi[^{F\times T}$, nombre d'itérations N_{it}.

Initialisation :

$\hat{X} = Ve^{i\phi}$

pour $it = 1$ à N_{it} **faire**

$Y = \mathcal{F}(\hat{X})$.

$\hat{X} = \dfrac{Y}{|Y|}V$.

fin pour

Sortie :

$\hat{X} \in \mathbb{C}^{F\times T}$

et la fonction d'inconsistance est donnée par le carré de la norme de Frobenius de cette matrice :

$$\mathcal{I}(X) = ||I_X||_2^2 = \sum_{f,t} |I_X(f,t)|^2. \tag{2.4}$$

La figure 2.2 illustre cette notion de consistance, qui est liée au caractère redondant de la TFCT, calculée en utilisant des fenêtres d'analyse successives qui se recouvrent dans le temps.

Algorithme de Griffin et Lim

Principe général Des approches itératives ont été mises au point NAWAB et al. (1983) afin de produire, à partir d'un spectrogramme d'amplitude donné, une matrice complexe qui soit la plus consistante possible. L'algorithme de Griffin et Lim (GL) GRIFFIN et LIM (1984) consiste à itérer l'opérateur $\mathcal{F}$ en forçant à chaque itération le module de la matrice obtenue à être égal à une valeur objectif V, comme c'est détaillé dans l'Algorithme 1.

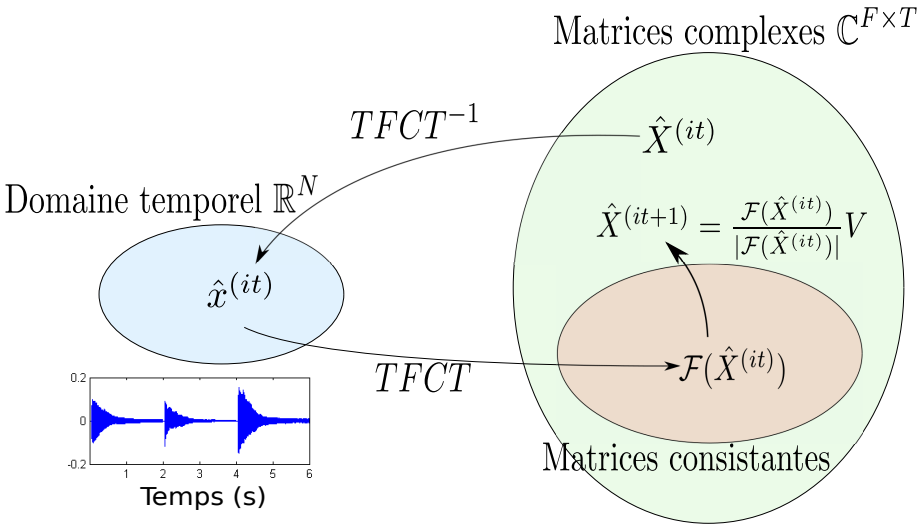


FIGURE 2.3 – Principe de l'algorithme GL : à chaque itération, on applique à la composante complexe estimée l'opérateur $\mathcal{F}$ puis on fixe son amplitude à la valeur objectif V.

Cette technique, illustrée sur la figure 2.3, permet de faire décroître la norme de Frobenius de $V - |\mathcal{F}(X)|$ au cours des itérations, comme cela a été prouvé dans GRIFFIN et LIM (1984). Par la suite, LE ROUX et al. (2008c) ont montré que cet algorithme faisait décroître la fonction d'inconsistance (2.4).

Il est enfin à noter que des approches similaires existent dans le cadre de la reconstruction de phase d'images de diffraction en optique : l'algorithme de Gerchberg-Saxton GERCHBERG et SAXTON (1972) est en effet très proche en essence de l'algorithme GL. En outre, cette idée qui consiste à utiliser la redondance de la transformée utilisée (et donc les dépendances qui existent entre points TF successifs) pour contraindre la phase est appliquée à d'autres types de transformations, comme la transformée en ondelettes MALLAT et WALDSPURGER (2015).

Algorithme de Griffin et Lim rapide L'algorithme GL est relativement lourd en temps de calcul puisqu'une itération requiert le calcul d'une TFCT et d'une TFCT inverse. Ainsi, ZHU et al. (2007) proposent une implémentation de l'algorithme qui permet un calcul en temps réel : c'est l'algorithme RTISI (pour *Real-Time Iterative Spectrogram Inversion*). Cette approche est basée sur le fait que pour estimer la trame t d'une TFCT, seules les trames précédentes sont nécessaires. Son approche conduit également à proposer une initialisation des phases qui permet une convergence beaucoup plus rapide qu'une initialisation aléatoire. D'autres améliorations ont depuis été proposées pour cet algorithme, comme dans BEAUREGARD et al. (2015) qui propose d'initialiser la phase d'une trame donnée par déroulé linéaire (nous reviendrons plus loin sur ce type de méthodes) ou encore de GNANN et SPIERTZ (2010) qui propose en plus de tenir compte de l'énergie des trames pour traiter prioritairement celles de plus grande énergie. Par ailleurs, il est proposé dans GNANN et SPIERTZ (2009) d'utiliser des tailles de fenêtre variables pour mieux estimer les phases des composantes transitoires.

PERRAUDIN et al. (2013) formule l'algorithme GL comme solution d'un problème non-convexe. Les règles de mise à jour sont modifiées pour que $\hat{X}^{(it)}$ ne dépende plus seulement de $\hat{X}^{(it-1)}$ mais également de sa valeur à l'itération précédente $\hat{X}^{(it-2)}$. Il observe expérimentalement une nette amélioration de la vitesse de convergence, mais n'a par contre plus de garantie théorique de convergence.

Cas de la séparation de sources Dans le cadre de la séparation de sources, l'algorithme peut être étendu à l'estimation des phases de plusieurs composantes en exploitant la phase du mélange Gunawan et Sen (2010). À chaque itération, un terme complémentaire est ajouté au calcul de $\hat{X}_k$ afin de tenir compte de l'erreur entre le mélange observé et le mélange estimé. Cet algorithme est dénommé MISI (pour *Multiple Input Spectrogram Inversion*).

Contrainte de consistance explicite

L'approche de Le Roux et al. (2008c) consiste à explicitement calculer la fonction d'inconsistance $\mathcal{I}$ donnée par (2.4) et à la minimiser directement. On obtient ainsi un algorithme itératif qui est en substance équivalent à celui de Griffin et Lim, mais a l'avantage d'être plus rapide, car certaines approximations permettent d'éviter de calculer l'intégralité de la TFCT et de la TFCT inverse à chaque itération.

En notant N_w la longueur de la fenêtre d'analyse utilisée, S le décalage (en échantillons) entre deux trames d'analyse et $Q = N_w/S$, la fonction d'inconsistance est explicitement donnée par $\mathcal{I} = \sum_{f,t} |I_X(f,t)|^2$ avec, $\forall (f,t)$:

$$I_X(f,t) = \sum_{p=-\frac{N_w}{2}}^{\frac{N_w}{2}-1} \sum_{q=-(Q-1)}^{Q-1} e^{2i\pi \frac{qf}{Q}} \alpha(p,q) X(f-p, t-q), \tag{2.5}$$

où α est un noyau qui dépend uniquement des fenêtres d'analyse w_a et de synthèse w_s ainsi que des paramètres de la TFCT :

$$\alpha(p,q) = \frac{1}{N_w} \sum_{k=0}^{N_w-1} w_a(k) w_s(k+qS) e^{2i\pi p \frac{k+qS}{N_w}} - \delta_p \delta_q, \tag{2.6}$$

avec $\delta_l = 1$ si $l = 0$ et 0 sinon. L'équation (2.5) montre que l'inconsistance est donnée par la convolution entre la TFCT X et le noyau α modulé par le terme $e^{2i\pi \frac{qf}{Q}}$. Une formule alternative pour le calcul de α est disponible dans Le Roux (2009). L'intérêt de la méthode de Le Roux est d'éviter le calcul de tout le produit de convolution en ne considérant que les valeurs $\alpha(p,q)$ ou p et q sont proches de 0, car ce noyau de consistance décroit rapidement (en module) lorsque p et q s'éloignent de 0. Il donne également une méthode pour construire les fenêtres d'analyse et de synthèse de façon à ce que l'énergie du noyau soit concentrée de façon maximale autour de $(0,0)$.

Sur la figure 2.4, nous représentons le module d'un noyau de consistance obtenu avec une fenêtre d'analyse et de synthèse égales à la racine carrée d'une fenêtre de Hann de longueur $N_w = 512$ échantillons (cette fenêtre est proposée dans Le Roux et al. (2008c)), avec 75 % de recouvrement ($S = 128$ et $Q = 4$). On constate que la majorité de l'énergie du noyau est contenue dans une amplitude d'environ 5 canaux fréquentiels. Pour être plus précis, nous pouvons examiner la proportion d'énergie du noyau contenue dans les $2P-1$ canaux fréquentiels situés de part et d'autre de 0 ($p \in [\![-P, P]\!]$) par rapport à son énergie totale (en fonction de P). Cette proportion d'énergie est :

$$\frac{\sum_{p=-(P-1)}^{P-1} \sum_{q=-(Q-1)}^{Q-1} |\alpha(q,p)|}{\sum_{p=-(\frac{N_w}{2}-1)}^{\frac{N_w}{2}-1} \sum_{q=-(Q-1)}^{Q-1} |\alpha(q,p)|}. \tag{2.7}$$

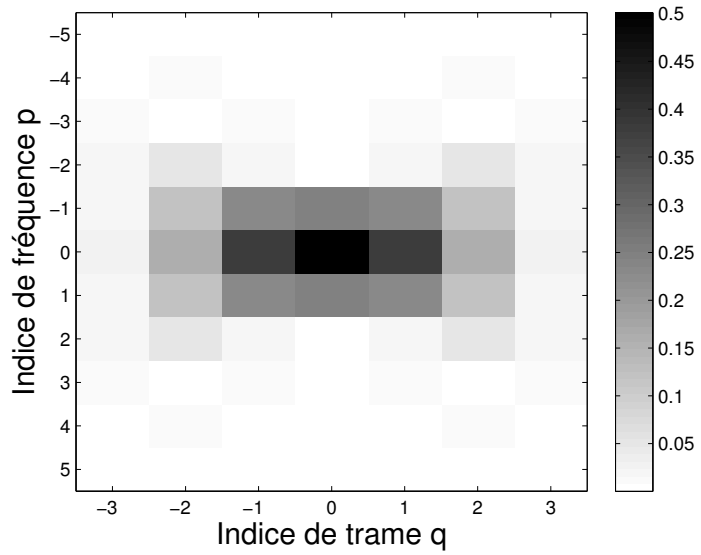


FIGURE 2.4 – Un exemple de module de noyau de consistance $|\alpha|$ obtenu à partir de racines carrées de fenêtres de Hann, de longueur 512 échantillons avec un recouvrement de 75 %.

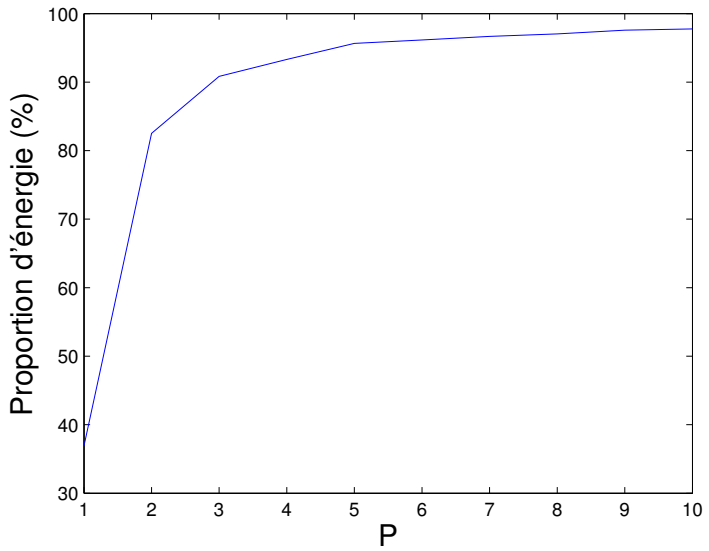


FIGURE 2.5 – Proportion d'énergie du noyau de consistance contenue dans les canaux fréquentiels d'indices $p \in [\![-P, P]\!]$ par rapport à son énergie totale.

La figure 2.5 montre que même pour de petites valeurs de P, la quasi-totalité de l'énergie du noyau est contenue dans ces quelques canaux fréquentiels : pour $P = 4$, cette proportion est de 93 %. L'idée au coeur de l'algorithme de Le Roux est donc de calculer la convolution (2.5) sur $2P - 1$ canaux fréquentiels plutôt que sur tous les canaux fréquentiels de la TFCT, en choisissant P tel que $P << N_w$:

$$I_X(f,t) \approx \sum_{p=-(P-1)}^{P-1} \sum_{q=-(Q-1)}^{Q-1} e^{j2\pi \frac{qf}{Q}} \alpha(p,q) X(f-p, t-q). \tag{2.8}$$

$\mathcal{I}$ est minimisée par un algorithme de descente de gradient, ce qui conduit à une règle de mise à jour de la phase :

$$\phi(f,t) \leftarrow \angle(I_X(f,t) + \alpha(0,0)X(f,t)), \tag{2.9}$$

où $\angle$ désigne l'argument complexe. D'autres simplifications (contrainte de parcimonie, utilisation des symétries de la fenêtre...) permettent d'améliorer encore les performances de l'algorithme. Une implémentation rapide de cette approche est détaillée dans Le Roux et al. (2010a).

Limites des approches consistantes

Lorsque la matrice non-négative V utilisée dans les approches consistantes est un spectrogramme consistant, les approches consistantes fournissent des résultats de bonne qualité. Néanmoins, lorsque ce n'est plus le cas (typiquement lorsque V est une approximation d'un spectrogramme de source, obtenue par exemple par NMF), le résultat obtenu n'est pas satisfaisant (*cf.* notre étude comparative menée au chapitre 3). Une idée intéressante serait alors de pouvoir modifier au fur et à mesure des itérations le spectrogramme afin que la consistance ne soit pas trop contraignante pour la reconstruction de phase.

En outre, les approches consistantes sont itératives, et donc souvent coûteuses en temps de calcul, même si des améliorations ont été faites sur ce point, comme nous l'avons rappelé. Enfin, ces approches visent toutes à résoudre un problème non-convexe, ce qui implique qu'il existe de nombreux minima locaux de la fonction de coût considérée (inconsistance). Certains auteurs (comme Sun et Smith (2012)) ont proposé de relâcher ce problème en un problème d'optimisation convexe. Néanmoins, cette opération a pour conséquence d'agrandir considérablement la dimension du problème, et pose à nouveau la question du temps de calcul.

2.1.4 Filtrage de Wiener consistant

Dans Le Roux et al. (2010b), les auteurs proposent de combiner la reconstruction de phase par approche de consistance et le filtrage de Wiener. On estime alors les sources via le schéma itératif suivant :

$$\hat{X}_k^{(it+1)} = \frac{\frac{1}{\sum_k V_k^{\odot 2}} \odot G_k \odot X + \gamma \mathcal{F}(\hat{X}_k^{(it)})}{\frac{1}{\sum_k V_k^{\odot 2}} + \gamma}, \tag{2.10}$$

où G_k est le gain de Wiener traditionnel (2.2) et γ est un paramètre de pondération qui ajuste l'importance de la contrainte de consistance. Cette approche fournit de meilleurs résultats que les deux approches (GL et filtrage de Wiener) prises séparément, mais oblige à actualiser à chaque itération la valeur du paramètre γ, ce qui peut s'avérer délicat en pratique.

Dans Sturmel et Daudet (2012), les auteurs proposent un raffinement de l'approche précédente. Il s'agit d'appliquer le masque de Wiener aux points TF où il n'y a pas de recouvrement, et d'appliquer une approche consistante dans les zones du domaine TF où des sources se recouvrent. Il faut donc procéder en une partition du domaine TF, d'où l'appellation de cette méthode PPR (*Partitioned Phase Retrieval*). Un domaine de confiance Ω_k est défini pour chaque source k :

$$\Omega_k = \{(f,t) \mid G_k(f,t) > \tau\}, \tag{2.11}$$

où G_k est gain de Wiener (2.2) et $\tau > 0$ est un seuil défini par l'utilisateur. En clair, le domaine de confiance Ω_k est constitué des points (f,t) où la k-ième source est dominante. On procède alors à l'initialisation $\hat{X}_k^{(0)} = G_k \odot X$, et à chaque itération :

$$\hat{X}_k^{(it+1)}(f,t) = \begin{cases} |\hat{X}_k^{(0)}(f,t)| e^{i\angle\mathcal{F}(\hat{X}_k^{(it)})(f,t)} & \text{si } (f,t) \notin \Omega_k, \\ \hat{X}_k^{(it)}(f,t) & \text{si } (f,t) \in \Omega_k. \end{cases} \tag{2.12}$$

Les résultats sont meilleurs que dans l'approche précédente, notamment au niveau du rejet d'interférences. Par ailleurs, le seuil τ est fixé à l'initialisation, et n'est plus actualisé par la suite.

Dans Sturmel et Daudet (2013), les auteurs proposent de combiner cette approche (PPR) avec l'idée d'exploiter la phase du mélange (MISI). Cela permet notamment de préserver l'énergie globale du mélange, puisque la somme des composantes estimées est alors égale aux observations.

Enfin, une amélioration de cette approche par partition de domaines est proposée dans Watanabe et Mowlaee (2013). Elle consiste à considérer les sources comme des mélanges de sinusoïdes, et donc à contraindre l'appartenance au domaine de confiance Ω_k non pas seulement par un seuil d'énergie, mais par l'appartenance au domaine sinusoïdal $\Omega_{k,sin}$ défini comme l'ensemble des bandes de fréquences du mélange de sinusoïdes de la source k.

Dans l'article Le Roux et Vincent (2013), Le Roux propose une nouvelle formulation du problème de filtrage de Wiener consistant. Des d'expériences sont menées et montrent que cette approche se compare favorablement à celles précédemment citées. Cependant, ces expériences cherchent à reconstruire les phases de deux sources seulement (parole et bruit). Par ailleurs, les spectrogrammes de ceux deux sources sont soit supposés connus (cas Oracle), soit le spectrogramme de bruit est connu et alors celui de parole non bruitée est estimé par soustraction spectrale. Nous proposons au chapitre 3 un cadre élargi, celui de la séparation de sources musicales lorsque les spectrogrammes sont estimés par NMF.

On pourra enfin se référer à Sturmel et Daudet (2011) pour une vue d'ensemble des techniques de reconstruction de phases basées sur ces approches consistantes.

2.1.5 Modèles de signaux

Alternativement, certaines méthodes de reconstruction de phase sont basées sur l'observation de signaux fondamentaux comme les mélanges de sinusoïdes. La modélisation de signaux musicaux par mélanges de sinusoïdes est fréquemment employée dans la littérature (modèle de McAulay et Quatieri McAuley et Quatieri (1986)). Dans l'algorithme du vocoder de phase Flanagan et Golden (1966), la phase de la TFCT d'une sinusoïde est explicitée. Cette approche exploite les relations naturelles qui existent entre phases de points TF successifs. Dans le vocoder de phase, cette idée est principalement appliquée à l'étirement temporel et à la modification de hauteur, et nécessite la phase de la TFCT originale. L'exploitation des phases de sinusoïdes a des applications dans divers domaines tels que la synthèse de parole Stylianou (2001) et plus généralement de signaux audio Girin et al. (2003).

Dans Gerkmann et al. (2012); Krawczyk et Gerkmann (2012, 2014), les auteurs utilisent une technique similaire pour reconstruire les phases de signaux de parole bruités. Ces approches modélisent des mélanges harmoniques et stationnaires, ce qui conduit à une propagation de l'erreur d'estimation de fréquence fondamentale à travers les partiels et les trames temporelles. Il est également intéressant de noter que le filtrage de Wiener et le modèle harmonique ont été combinés afin de calculer un masque TF qui tienne compte du modèle sinusoidal ainsi que de la phase du mélange Krawczyk et Gerkmann (2015).

Bronson et Depalle (2014) ont proposé une NMF complexe avec une contrainte de phase basée sur une modélisation sinusoïdale, que nous détaillons dans la section 2.3.1.

Les modèles sinusoïdaux sont au coeur de nombreux développements sur la reconstruction de phase, et permettent notamment l'obtention d'estimateurs MMSE des composantes. Connaissant l'amplitude et la phase du mélange, ainsi que les amplitudes des sources séparées, Mowlaee et al. (2012); Chacon et Mowlaee (2014) proposent une méthode d'estimation des phases des sources s'appuyant sur l'écriture explicite des composantes complexes dans le domaine TF (utilisant une décomposition polaire des composantes). Cette approche est par la suite étendue à l'estimation des amplitudes et est appliquée au rehaussement de la parole Mowlaee et Saeidi (2013) et à la séparation de sources Mowlaee et Martin (2012). Néanmoins, cette méthode ne s'applique qu'à des mélanges de deux sources (parole et bruit). Par ailleurs, ces dernières expériences sont conduites dans le cas où le signal de parole est connu, aussi seul le bruit est estimé via cette technique.

Enfin, ces contributions ont été reprises dans Mowlaee et al. (2013) et Mowlaee et al. (2014), où les auteurs insistent notamment sur la potentielle utilisation des structures au sein du champ de phases (délai de groupe, dérivées...) pour interpréter celui-ci et l'utiliser au mieux pour les applications audio. Dans Mowlaee et Saeidi (2014), il est proposé d'incorporer des contraintes sur les sauts de phases entre partiels de sinusoides et le délai de groupe pour améliorer l'estimation des phases dans ce contexte.

Ces approches ont pour intérêt principal l'utilisation de la phase pour une estimation optimale (au sens MMSE) des composantes complexes, ou bien l'utilisation de cet estimateur pour la reconstruction de la phase d'un signal de parole. Cependant, elles n'exploitent pas toute l'information sur la nature sinusoidale des signaux, et ne sont envisagées que dans des cas restreints (mélange de deux sources où l'une est connue) ou pour des applications particulières (débruitage de parole).

2.1.6 Modèles probabilistes

Des modèles de phase non-uniforme ont été introduits dans un cadre probabiliste. Les modèles Krawczyk et Gerkmann (2015); Sunnydayal et Kumar (2015) considèrent que la phase est une variable aléatoire suivant une loi circulaire non-uniforme dont le paramètre de localisation est égal à la phase obtenue par application d'un modèle sinusoïdal. Dans ces travaux, les signaux étudiés sont des signaux de parole et de bruit : le mélange ne comprend que deux sources, dont une qui est modélisée par un bruit blanc gaussien. Les lois employées sont notamment la distribution de Von Mises Mardia et Zemroch (1975) et la loi normale périodique dans le cadre de la modélisation de la parole Agiomyrgiannakis et Stylianou (2009).

Un verrou de ces approches est qu'il n'est pas encore évident de les généraliser à des modèles de sources multiples et musicales : il est en effet délicat d'estimer les paramètres de ces modèles ainsi que d'obtenir un estimateur des sources.

2.1.7 Méthodes temporelles

Dans Achan et al. (2003), il est proposé d'estimer la phase d'un signal de parole à partir de son spectrogramme via un modèle statistique : connaissant le spectrogramme dans le domaine TF, on suppose que le signal de parole dans le domaine temporel suit un modèle autorégressif (AR). Ainsi, on peut estimer le signal temporel en le choisissant de sorte à ce que son spectrogramme soit le plus proche possible de celui préalablement estimé. De fait, on n'estime pas directement la phase mais on reconstruit le signal temporel.

Certains travaux Le Roux et al. (2008a); Yoshii et al. (2013); Févotte et Kowalski (2014) proposent de séparer les sources dans le domaine temporel plutôt que dans le domaine TF afin de s'affranchir de la problématique de la reconstruction de phase. Ces méthodes reposent sur un modèle de mélange convolutif : les sources sont filtrées par une réponse de salle ce qui résulte en des signaux dits *sources images*. On introduit des modèles de type NMF pour structurer les paramètres des signaux sources. Néanmoins, ces méthodes sont très coûteuses en temps de calcul et peu robustes face aux variations de forme d'onde des atomes temporels d'une occurrence de la source à une autre. Enfin, ces méthodes sont appliquées dans un cadre supervisé, où le dictionnaire d'atomes temporels est appris au préalable.

Il existe de nombreux travaux sur les méthodes de séparation de sources temporelles, mais nous ne les détaillons pas davantage, car cette thèse s'intéresse à la séparation de sources dans le domaine TF.

Comme nous l'avons mentionné dans le préambule de ce chapitre, les méthodes de reconstruction de phase décrites dans cette section nécessitent l'estimation préalable d'un spectrogramme d'amplitude pour chaque source. Nous nous sommes intéressés aux approches NMF car celles-ci sont très populaires en audio Smaragdis et Brown (2003); Virtanen (2007). Nous effectuons donc ci-après une présentation générale des modèles NMF dans le cadre de la séparation de sources.

2.2 Factorisation en matrices non-négatives

La NMF est une technique qui a été utilisée dans de nombreux domaines, tels que le traitement d'images Lee et Seung (1999), la spectroscopie Liu et al. (2013) ou l'analyse de données textuelles Pauca et al. (2004). On pourra consulter Cichocki et al. (2009) pour une vue d'ensemble des applications de la NMF.

Dans le cadre du traitement du signal audio, des applications de la NMF sont par exemple la transcription automatique de musique en partitions Smaragdis et Brown (2003), la séparation de sources Wang et Plumbley (2005); Virtanen (2007) ou encore la restauration audio Le Roux et al. (2008b). La NMF a également été utilisée dans le domaine du rehaussement de la parole Wilson et al. (2008).

2.2.1 Principe général

Originellement, la NMF a été introduite comme une méthode de réduction de rang de matrices Lee et Seung (1999). Le problème de la NMF s'exprime de la façon suivante : si on considère une matrice V de dimensions $F \times T$ à coefficients non-négatifs, on cherche une approximation de V sous la forme factorisée suivante :

$$V \approx \hat{V} = WH, \tag{2.13}$$

où W et H sont deux matrices à coefficients non-négatifs de dimensions $F \times K$ et $K \times T$ respectivement. Pour réduire la dimension des données, K est choisi de sorte à ce que $K(F+T) \ll FT$.

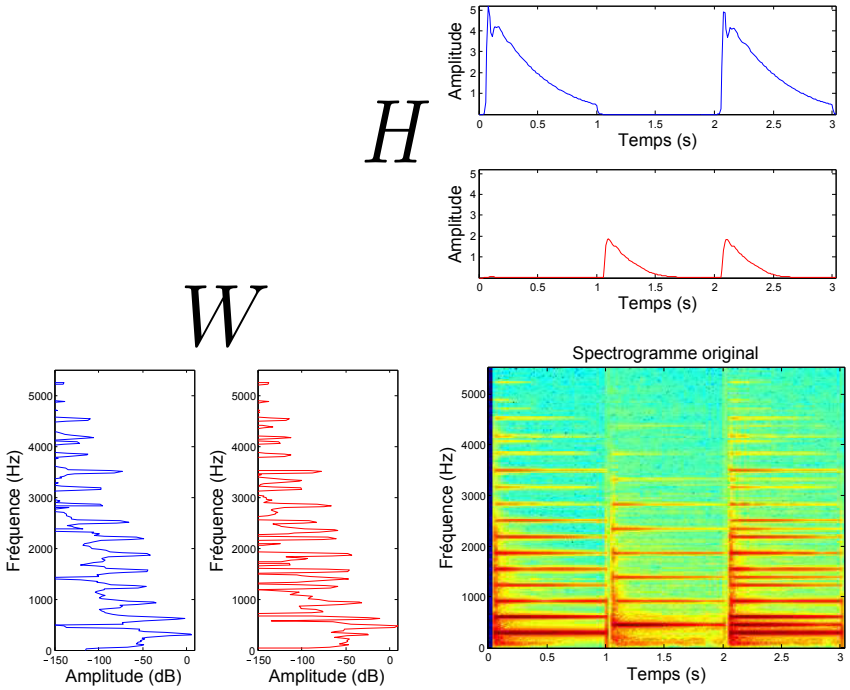


FIGURE 2.6 – Résultat de 100 itérations de règles multiplicatives d'une NMF euclidienne sur un spectrogramme constitué de deux notes de piano (E4 et B4).

En audio, V est généralement le spectrogramme $|X|^{\odot\alpha}$ d'une représentation TF X d'un signal. Si $\alpha = 1$, il s'agit du spectrogramme d'amplitude, et si $\alpha = 2$, il s'agit du spectrogramme de puissance (qui sont les deux représentations les plus couramment utilisées).

Un des principaux intérêts de la NMF est de fournir une factorisation qui soit interprétable intuitivement. On peut en effet voir W comme un dictionnaire d'atomes spectraux et H comme une matrice d'activations temporelles. Si W_k désigne la k-ième colonne de W et H_k la k-ième ligne de H, alors $\hat{V}_k = W_k H_k$ est le spectrogramme de la composante indexée par k. Par construction, on a :

$$\hat{V} = \sum_{k=1}^{K} \hat{V}_k, \tag{2.14}$$

ce qui traduit une propriété d'additivité des spectrogrammes. Si les V_k représentent des spectrogrammes empiriques, cette propriété n'est vérifiée que lorsque les sources ne se recouvrent pas dans le plan TF : le module de la somme des composantes n'est pas égal à la somme de leurs modules en général. Cependant, si les V_k représentent des spectrogrammes de puissance théoriques (c'est-à-dire des densités spectrales de puissance, ou plus généralement des paramètres de dispersion comme la variance dans des modèles probabilistes, présentés dans la section 2.2.3), alors cette propriété est vérifiée en moyenne.

La contrainte de non-négativité des données et des paramètres dans le modèle NMF est son principal atout par rapport aux autres techniques de réduction de rang (PCA, ICA...). Cette contrainte conduit à une décomposition qui fait sens, les atomes spectraux ainsi que les activations temporelles étant interprétables physiquement. On peut le voir sur la figure 2.6 qui montre un exemple de factorisation du spectrogramme d'un mélange de deux notes de piano.

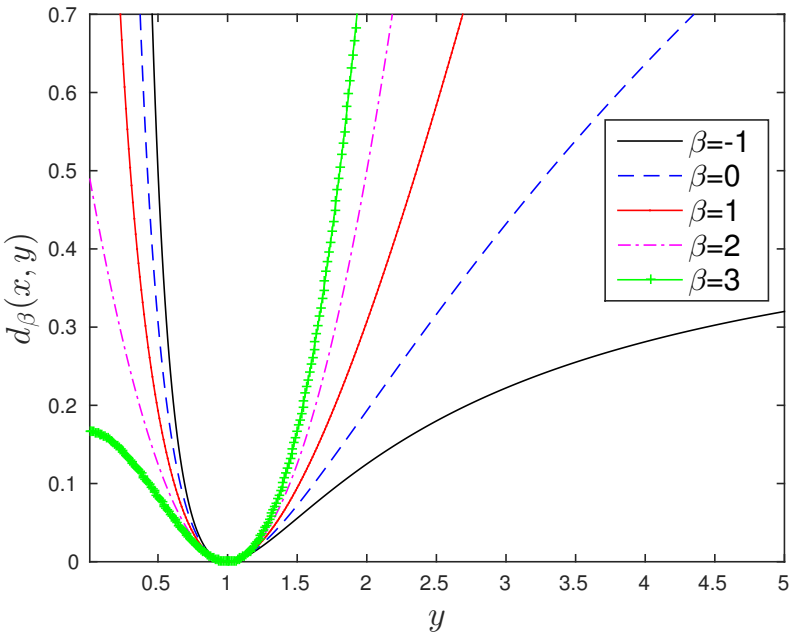


FIGURE 2.7 – Fonction de β-divergence $d_\beta(x,y)$ pour $x=1$ et plusieurs valeurs de β.

2.2.2 Fonctions de coût

La factorisation (2.13) s'obtient en minimisant une fonction de coût $D(V,\hat{V}) = D(V,WH)$. On utilise des fonctions qui présentent deux propriétés :

— Ce sont des *divergences*, c'est-à-dire des fonctions à valeurs positives telles que $D(X,Y) = 0 \Leftrightarrow X = Y$. À la différence des distances, elles ne vérifient pas forcément les propriétés de symétrie et d'inégalité triangulaire.

— Elles sont *séparables* : $D(X,Y) = \sum_{f,t} d(X(f,t),Y(f,t))$.

De nombreux choix de fonctions de coût sont possibles. Une classe de fonctions est particulièrement populaire en traitement du signal audio, les β-divergences CICHOCKI et AMARI (2010), définies comme suit :

$$d_\beta(x,y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\ln(\frac{x}{y}) + y - x & \beta = 1 \\ \frac{x}{y} - \ln(\frac{x}{y}) - 1 & \beta = 0. \end{cases} \tag{2.15}$$

Cette famille de divergences généralise plusieurs fonctions fréquemment utilisées en analyse de signaux :

— pour $\beta = 2$, c'est la distance euclidienne (Euc),

— pour $\beta = 1$, c'est la divergence de Kullback-Leibler (KL) KULLBACK et LEIBLER (1951),

— pour $\beta = 0$, c'est la divergence d'Itakura-Saito (IS) ITAKURA et SAITO (1968).

La figure 2.7 illustre quelques β-divergences.

Les divergences Euc et KL sont utilisées comme fonctions de coût pour estimer le modèle NMF dans LEE et SEUNG (1999). La divergence KL présente notamment l'intérêt d'être

plus adaptée à la perception humaine, qui est construite sur une échelle logarithmique, que la distance Euclidienne. La NMF avec divergence IS a été introduite dans Févotte et al. (2009). Cette dernière présente l'avantage de l'invariance d'échelle :

$$\forall a \in \mathbb{R}_+,\ d_0(ax, ay) = d_0(x, y). \tag{2.16}$$

Cela signifie que les zones du plan TF où il y a peu d'énergie comptent autant dans le calcul de la divergence que les zones de forte énergie. Cette propriété est pratique car des bandes de fréquences de faible énergie en audio peuvent contribuer perceptivement autant que des bandes de plus forte énergie (au niveau des harmoniques aigus notamment). Des études ont par ailleurs été menées pour déterminer le paramètre β optimal dans un contexte de séparation de sources musicales Fitzgerald et al. (2008).

2.2.3 Modélisation probabiliste

Les approches probabilistes pour la modélisation et la séparation de sources ont été très populaires pendant cette dernière décennie. En effet, le cadre probabiliste permet non seulement de modéliser les sources ainsi que les erreurs, mais fournit en outre un cadre rigoureux pour introduire un certain nombre de contraintes dans le modèle via des à priori sur ses paramètres. Enfin, il ouvre également la voie à de nouvelles techniques d'estimation des modèles. On pourra se référer à Vincent et al. (2010) pour une vue d'ensemble de ces modèles.

Modèles génératifs

Le principe des modèles *génératifs* est de modéliser le mécanisme qui gouverne la production des données. Les observations sont alors vues comme la réalisation d'un processus aléatoire qui dépend de variables dites latentes, car non observées (les sources) ainsi que de certains paramètres qui les caractérisent. Le principe des modèles NMF probabilistes est alors de structurer non pas les réalisations des variables latentes (comme effectué précédemment) mais leurs paramètres de dispersion (comme les variances). Ces paramètres sont estimés par diverses méthodes, comme la technique du maximum de vraisemblance (ML pour *Maximum Likelihood*) ou du maximum à postériori (MAP). Dans certains cas, on peut montrer qu'une estimation ML du modèle est équivalente à un problème de minimisation tel que présenté précédemment. Dans Févotte et Cemgil (2009), les auteurs présentent en effet trois modèles probabilistes dont l'estimation est équivalente au problème de NMF avec la distance Euclidienne et les divergences KL et IS.

Prenons l'exemple de Févotte et al. (2005). Dans ce travail, les sources sont modélisées par des distributions gaussiennes, sous l'hypothèse que tous les points TF sont indépendants :

$$X(f,t) = \sum_{k=1}^{K} X_k(f,t) \text{ avec } X_k(f,t) \sim \mathcal{N}(0, \sigma_k(f,t)^2), \tag{2.17}$$

où $\mathcal{N}$ désigne la loi normale circulaire complexe. Les variances sont structurées par un modèle NMF : $\sigma_k(f,t)^2 = W(f,k)H(k,t)$. On peut alors montrer que la maximisation de la log-vraisemblance des données est équivalente à la minimisation de la divergence IS entre les observations $|X|^{\odot 2}$ et le modèle NMF WH. Ce modèle est appelé ISNMF.

De façon similaire, les auteurs dans Virtanen et al. (2008) utilisent un modèle de Poisson pour les sources. Ils montrent ainsi que l'estimation ML du modèle revient à effectuer une NMF avec divergence de KL (on parlera alors de KLNMF) entre les données $|X|$ et le modèle WH. La loi de Poisson modélise cependant des variables aléatoires discrètes, aussi certains

développements ont eu lieu pour fournir un cadre plus rigoureux à l'utilisation de cette loi lorsque l'on traite des variables aléatoires continues Hoffman (2012).

De très nombreux modèles de sources ont été proposés dans la littérature, aussi il ne nous semble pas justifié d'en faire ici l'inventaire exhaustif. Néanmoins, sur la base de ce qui vient d'être dit, une question intéressante émerge : dans le modèle gaussien, la factorisation est faite sur $|X|^{\odot 2}$ alors que dans le modèle de Poisson, elle est faite sur $|X|$. On peut donc se demander s'il existe un exposant optimal du spectrogramme d'amplitude sur lequel appliquer un modèle NMF, c'est-à-dire une valeur de l'exposant qui vérifie autant que possible une propriété d'additivité. Cette question a fait l'objet de plusieurs travaux Hennequin (2011); Liutkus et Badeau (2015); Voran (2015), et une piste possible pourrait être de s'intéresser à une famille de distributions qui est celle des loi α-stables Nolan (2015). Celles-ci ont en effet de bonnes propriétés (stabilité et robustesse notamment) et généralisent la loi normale ainsi que la loi de Cauchy Liutkus et al. (2015). Elles fournissent un cadre théorique pour obtenir un estimateur des sources par filtrage de Wiener généralisé (que nous avons évoqué dans la section 2.1.2). Nous avons par ailleurs proposé certains développements sur les distributions α-stables, qui sont présentés dans le chapitre 9.

Outre les sources, il est possible de modéliser le bruit, qui peut traduire une erreur entre les données et l'approximation, ou bien un modèle physique de bruit, comme par exemple les perturbations liées à l'acquisition des données. Ce bruit peut notamment être additif (par exemple un bruit gaussien, *cf.* Schmidt et Laurberg (2008)) ou multiplicatif (par exemple de loi Gamma *cf.* Févotte et al. (2009)).

Ces modèles supposent l'indépendance des points TF. Cette propriété est utilisée pour son côté pratique et simplifie grandement les calculs effectués. Néanmoins, elle est peu réaliste : même pour des signaux très simples (une sinusoïde), les points TF sont dépendants les uns des autres. Ainsi, les relations entre points adjacents ne sont pas prises en compte. Des propositions ont été faites pour prendre en compte ces relations : utilisation de chaîne de Markov pour modéliser la dépendance des amplitudes entre trames adjacentes Mysore et al. (2010), ou modèle autorégressif par bande de fréquences sur les sources complexes (c'est le modèle de NMF à haute résolution Badeau (2011) que nous détaillons dans la section 2.3.2). L'introduction de dépendances améliore la qualité des résultats obtenus et fournit une représentation plus réaliste physiquement. Néanmoins, elle a tendance à compliquer les modèles, et donc leur estimation. Il est donc nécessaire de trouver un compromis entre le pouvoir expressif d'un modèle génératif et notre capacité à en estimer les paramètres.

Modèle de comptage

Alternativement aux modèles génératifs, l'analyse en composantes latentes (ou PLCA de l'anglais *Probabilistic Latent Component Analysis*) Smaragdis et al. (2006, 2007) consiste en un modèle de comptage. Les observations non-négatives V sont vues comme l'histogramme issu du tirage des variables aléatoires f et t. Ces variables ont une loi jointe $P(f,t)$ qui dépend d'une variable cachée (composante latente) k. On peut alors écrire, en utilisant la règle de Bayes ainsi que l'indépendance des variables :

$$P(f,t) = \sum_{k=1}^{K} P(f|k)P(k,t). \tag{2.18}$$

On constate qu'estimer la log-vraisemblance des observations V est équivalent à une NMF avec divergence KL Shashanka et al. (2008), en posant $W(f,k) = P(f|k)$ et $H(k,t) = P(k,t)$. W est alors normalisée (les coefficients somment à 1).

2.2.4 Estimation du modèle NMF

Que l'on adopte une approche déterministe ou probabiliste pour structurer des observations par un modèle NMF (2.13), son estimation se ramène à la minimisation d'une fonction de coût $\mathcal{C}(\theta)$. Bien souvent, $\mathcal{C}$ est une β-divergence, à laquelle peut être ajoutée une ou plusieurs pénalités, et θ est l'ensemble des paramètres (constitué uniquement de W et H dans une NMF classique).

De nombreux algorithmes existent pour effectuer cette minimisation : algorithme à gradient projeté, méthode de Newton, moindres carrés alternés... On pourra se référer à Berry et al. (2007) pour une présentation de ces algorithmes, mais nous nous limiterons ici aux principales techniques rencontrées dans la littérature, et qui serviront dans la suite de ce manuscrit.

Approche heuristique

Nous présentons ici l'approche qui a été initialement utilisée pour l'estimation du modèle NMF Lee et Seung (1999). Cette approche est encore aujourd'hui très largement utilisée par la communauté scientifique pour sa simplicité, et en raison du fait qu'elle conduit à des règles de mise à jour multiplicatives (MUR pour *Multiplicative Update Rules*), efficaces sur le plan du temps de calcul. L'idée est d'écrire le gradient de la fonction de coût $\mathcal{C}$ comme la différence de deux composantes positives :

$$\nabla_\theta \mathcal{C}(\theta) = \nabla_\theta^+ - \nabla_\theta^-. \tag{2.19}$$

On considère alors la mise à jour suivante :

$$\theta \leftarrow \theta \times \frac{\nabla_\theta^-}{\nabla_\theta^+}. \tag{2.20}$$

Une telle mise à jour est construite de sorte que le paramètre θ varie dans le sens de la décroissance locale de $\mathcal{C}$. Néanmoins, ce n'est aucunement une garantie de la décroissance de la fonction de coût. Celle-ci est démontrée dans certains cas (où il s'avère que ces règles de mises à jour sont les mêmes que celles obtenues par des méthodes plus rigoureuses) mais ce n'est pas systématique. Dans Badeau et al. (2010) les auteurs montrent qu'il est parfois préférable d'utiliser des règles de mise à jour qui ne font pas décroître la fonction de coût de façon monotone, car elles accélèrent la vitesse de convergence de l'algorithme.

Dans le cadre de la NMF, Févotte et al. (2009) fournit les règles de mise à jour pour les β-divergences :

$$H \leftarrow H \odot \frac{W^T((WH)^{\odot[\beta-2]} \odot V)}{W^T(WH)^{\odot[\beta-1]}}, \tag{2.21}$$

$$W \leftarrow W \odot \frac{((WH)^{\odot[\beta-2]} \odot V)H^T}{(WH)^{\odot[\beta-1]}H^T}, \tag{2.22}$$

où $\odot$ (respectivement la barre de fraction) désigne la multiplication (respectivement la division) de matrices terme à terme. La décroissance de la fonction de coût sous ces règles de mise à jour a été démontrée pour $\beta = 1$ et 2 dans Lee et Seung (2001). Elle a ensuite été étendue au cas $\beta \in [1, 2]$ dans Kompass (2007), et la démonstration a été généralisée dans Févotte et Idier (2011) à $\beta \in [0, 2]$ (ce qui correspond au cas d'application pratique).

Ces règles multiplicatives garantissent la propriété de positivité des matrices W et H, à condition que l'initialisation vérifie aussi cette propriété. En effet, les termes ∇_θ^+ et ∇_θ^- étant positifs, le signe de θ ne change pas au cours des itérations dans (2.20).

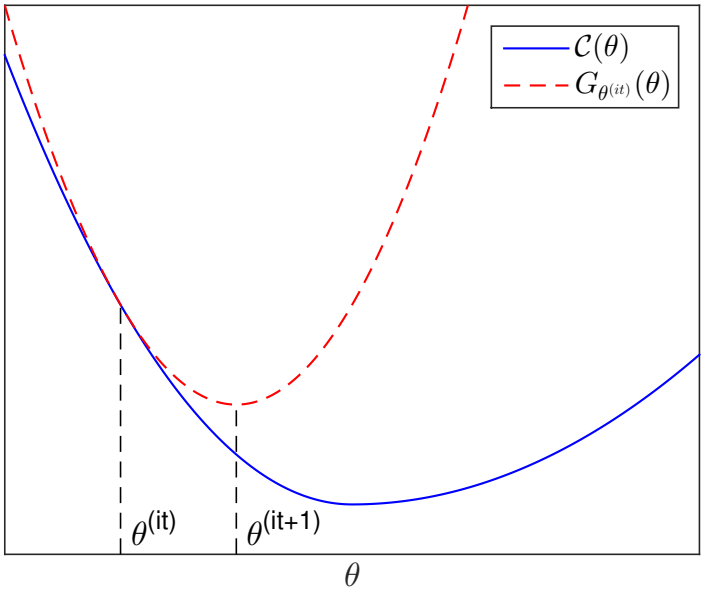


FIGURE 2.8 – Illustration de l'approche Majoration-Minimisation : la fonction de coût $\mathcal{C}$ est majorée à l'itération (it) par la fonction auxiliaire $G_{\theta^{(it)}}$, dont la minimisation conduit à un nouveau paramètre $\theta^{(it+1)}$.

Pour $\beta \in [1, 2]$, les β-divergences sont convexes par rapport aux variables W (à H fixé) et à H (à W fixé). Ce n'est plus le cas lorsque $\beta \notin [1, 2]$, comme cela a été montré notamment dans BERTIN et al. (2009b). En outre, elles ne sont pas conjointement convexes par rapport à W et H. Cela explique pourquoi une minimisation alternée par rapport à chacune de ces deux variables est utilisée. Par ailleurs, cela pose le problème de la non-unicité des solutions obtenues, en raison du grand nombre de minima locaux de la fonction. Le problème de la non-unicité des solutions vient également de la structure de la factorisation. Par exemple, si (W, H) est une solution, alors $(WB, B^{-1}H)$ en est une également (à condition que WB et $B^{-1}H$ soit non-négatives). On lève en général cette indétermination en normalisant une des deux matrices (il est commun de normaliser les colonnes de W) pour avoir l'unicité par rapport au facteur d'échelle, i.e. pour des matrices B diagonales. Par ailleurs, on peut imposer un ordre (par exemple, les colonnes de W doivent être rangées par ordre d'énergie croissante) pour éviter l'indétermination sur la permutation.

Algorithme Majoration-Minimisation (MM)

L'algorithme Majoration-Minimisation (MM) HUNTER et LANGE (2004) fournit un cadre théorique rigoureux pour obtenir les règles précédentes. Le principe de l'algorithme MM est de majorer la fonction de coût $\mathcal{C}(\theta)$ (dans le cas de la NMF, $\theta = \{W, H\}$) en un point $\theta^{(it)}$ par une fonction auxiliaire $G_{\theta^{(it)}}$, dont la minimisation est possible analytiquement, et conduit à la décroissance de la fonction de coût [2]. Cette méthode est illustrée sur la figure 2.8.

Une telle fonction doit vérifier les propriétés suivantes :

— elle est égale à la fonction de coût au point $\theta^{(it)}$: $G_{\theta^{(it)}}(\theta^{(it)}) = \mathcal{C}(\theta^{(it)})$,

2. De façon complètement équivalente, lorsque l'on souhaite faire croître un certain critère, on pourra construite une fonction auxiliaire minorante et maximiser celle-ci, comme on le fait dans l'algorithme EM présenté un peu plus loin.

— elle doit majorer la fonction de coût : $\forall\theta$, $G_{\theta^{(it)}}(\theta) \geq \mathcal{C}(\theta)$.

L'idée est alors de minimiser $G_{\theta^{(it)}}$, ce qui conduit à la mise à jour :

$$\theta^{(it+1)} = \arg\min_{\theta} G_{\theta^{(it)}}(\theta). \tag{2.23}$$

Une telle mise à jour permet de faire décroître le critère $\mathcal{C}(\theta^{(it)})$. En effet, par définition de $\theta^{(it+1)}$, on a $\forall\theta$:

$$G_{\theta^{(it)}}(\theta^{(it+1)}) \leq G_{\theta^{(it)}}(\theta). \tag{2.24}$$

En particulier pour $\theta = \theta^{(it)}$, cela donne :

$$G_{\theta^{(it)}}(\theta^{(it+1)}) \leq G_{\theta^{(it)}}(\theta^{(it)}) = \mathcal{C}(\theta^{(it)}). \tag{2.25}$$

Par ailleurs, par définition de la fonction auxiliaire, on sait que :

$$G_{\theta^{(it)}}(\theta^{(it+1)}) \geq \mathcal{C}(\theta^{(it+1)}). \tag{2.26}$$

Ainsi, en combinant (2.25) et (2.26), on aboutit à $\mathcal{C}(\theta^{(it+1)}) \leq \mathcal{C}(\theta^{(it)})$ ce qui prouve la décroissance du critère.

Dans le cadre de la NMF, cet algorithme a été utilisé pour justifier la décroissance de la fonction de coût avec les règles (2.21) et (2.22) pour la distance Euclidienne et la divergence KL Lee et Seung (2001). Dans Févotte et Idier (2011), les auteurs ont montré que pour toute β-divergence telle que $\beta \in [1, 2]$, la méthode MM conduit exactement à ces règles de mise à jour.

La difficulté qui se pose en pratique est la construction d'une telle fonction auxiliaire. Dans Févotte et Idier (2011), les auteurs suggèrent de décomposer la fonction de coût $\mathcal{C}$ en une partie convexe et une partie concave : la partie convexe est majorée en utilisant l'inégalité de Jensen, et la partie concave est majorée par sa tangente. Nous utiliserons cette technique dans ce manuscrit, notamment au chapitre 9.

Méthode de la fonction auxiliaire

Il est parfois compliqué de trouver une fonction auxiliaire qui permette d'appliquer la technique MM présentée précédemment. On peut alors concevoir une approche similaire, mais qui consiste à augmenter la taille de l'espace des paramètres en introduisant des paramètres auxiliaires $\tilde{\theta}$. La *méthode de la fonction auxiliaire* considère une fonction $g(\theta, \tilde{\theta})$ telle que :

$$\mathcal{C}(\theta) = \min_{\tilde{\theta}} g(\theta, \tilde{\theta}). \tag{2.27}$$

On peut alors montrer que $\mathcal{C}$ est décroissante sous les règles de mise à jour suivantes :

$$\tilde{\theta} \leftarrow \arg\min_{\tilde{\theta}} g(\theta, \tilde{\theta}) \text{ et } \theta \leftarrow \arg\min_{\theta} g(\theta, \tilde{\theta}), \tag{2.28}$$

ce qui revient à minimiser non plus $\mathcal{C}$ directement, mais g en alternant les mises à jour sur θ et $\tilde{\theta}$. En effet, considérons une valeur des paramètres $\theta^{(it)}$. Les mises à jour s'écrivent :

$$\tilde{\theta}^{(it+1)} \leftarrow \arg\min_{\tilde{\theta}} g(\theta^{(it)}, \tilde{\theta}) \text{ et } \theta^{(it+1)} \leftarrow \arg\min_{\theta} g(\theta, \tilde{\theta}^{(it+1)}). \tag{2.29}$$

En combinant la définition de g d'après (2.27) et la mise à jour sur $\tilde{\theta}$, il est clair que :

$$\mathcal{C}(\theta^{(it)}) = g(\theta^{(it)}, \tilde{\theta}^{(it+1)}). \tag{2.30}$$

En utilisant à présent la mise à jour sur θ, on trouve que $\forall\theta,\ g(\theta^{(it+1)},\tilde{\theta}^{(it+1)}) \leq g(\theta,\tilde{\theta}^{(it+1)})$. En particulier, pour $\theta = \theta^{(it)}$, on a :

$$g(\theta^{(it+1)},\tilde{\theta}^{(it+1)}) \leq g(\theta^{(it)},\tilde{\theta}^{(it+1)}). \tag{2.31}$$

Enfin, par définition de la fonction auxiliaire on a $\mathcal{C}(\theta^{(it+1)}) = \min_{\tilde{\theta}} g(\theta^{(it+1)},\tilde{\theta})$, soit :

$$\forall\tilde{\theta},\ \mathcal{C}(\theta^{(it+1)}) \leq g(\theta^{(it+1)},\tilde{\theta}), \tag{2.32}$$

soit en particulier pour $\tilde{\theta} = \tilde{\theta}^{(it+1)}$:

$$\mathcal{C}(\theta^{(it+1)}) \leq g(\theta^{(it+1)},\tilde{\theta}^{(it+1)}). \tag{2.33}$$

Finalement, en combinant les équations (2.30), (2.31) et (2.33), on obtient :

$$\mathcal{C}(\theta^{(it+1)}) \leq \mathcal{C}(\theta^{(it)}), \tag{2.34}$$

ce qui prouve la décroissance de $\mathcal{C}$.

Cette technique est employée pour estimer les modèles de NMF complexe Kameoka et al. (2009) et de NMF complexe consistante Le Roux et al. (2009) qui seront détaillés dans la section 2.3.1. Un intérêt fort de la méthode de la fonction auxiliaire est de découpler les paramères θ à estimer. En effet, ceux-ci sont liés dans la fonction objectif de départ $\mathcal{C}$ alors que leur estimation peut se faire de façon indépendante lorsqu'on agit sur g. Plusieurs algorithmes présentés dans ce manuscrit sont obtenus en utilisant la méthode de la fonction auxiliaire.

Approches probabilistes

Maximum de vraisemblance Dans un cadre probabiliste (*cf.* section 2.2.3), les observations X sont vues comme la réalisation d'un processus aléatoire dépendant de certains paramètres θ, processus défini par une loi $p(X|\theta)$. La méthode naturelle d'estimation des paramètres, dans ce contexte, consiste à maximiser la vraisemblance des observations (méthode ML), ou de façon équivalente, leur log-vraisemblance, donnée par :

$$L(\theta) = \log p(X|\theta). \tag{2.35}$$

En général (et notamment pour les modèles NMF qui nous intéressent ici), celle-ci peut être réécrite sous la forme :

$$L(\theta) \propto -\mathcal{C}(\theta), \tag{2.36}$$

où $\propto$ désigne l'égalité à une constante multiplicative positive et une constante additive près (constantes qui ne dépendent pas des paramètres à estimer). Ainsi, l'estimation ML des paramètres se ramène à un problème de minimisation d'un critère $\mathcal{C}$.

Maximum à postériori Alternativement, lorsque l'on souhaite introduire un à priori sur un ou plusieurs paramètres, on maximise plutôt la distribution à postériori des paramètres sachant les observations $p(\theta|X)$ (on parle alors d'estimateur du maximum à postériori MAP). Si on note $p(\theta)$ cet à priori, on a, en vertu de la règle de Bayes :

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}, \tag{2.37}$$

et donc :

$$\log p(\theta|X) \stackrel{c}{=} L(\theta) + \log p(\theta), \tag{2.38}$$

où $\overset{c}{=}$ désigne l'égalité à une constante additive près (en effet, $p(X)$ ne dépend pas des paramètres θ). L'estimation MAP des paramètres donc revient à minimiser un critère qui inclut un terme d'attache aux données (provenant de la vraisemblance $L(\theta)$) et un terme d'à priori sur les paramètres.

Dans ces deux approches, il est nécessaire de résoudre un problème d'optimisation pour lesquels s'appliquent l'algorithme MM ou la méthode de la fonction auxiliaire.

Algorithme Espérance-Maximisation Dans certains modèles à variables latentes, lorsque les distributions considérées sont sophistiqués, il devient parfois délicat de minimiser directement la fonction de coût. On peut alors voir l'algorithme Espérance-Maximisation (EM) Bishop (2006) comme un cas particulier de l'algorithme MM adapté aux modèles probabilistes à variables latentes, puisque cet algorithme est fondé sur la construction d'une fonction minorante particulière. L'idée de base de cet algorithme est qu'il est plus simple d'estimer la vraisemblance des données complètes (observations et variables latentes) que la vraisemblance des observations seules Dempster et al. (1977).

En notant Z ces variables latentes et q une densité de probabilité sur ces variables, on montre que l'on peut écrire la log-vraisemblance sous la forme :

$$L(\theta) = \mathcal{L}(q, \theta) + D_{KL}(q, p). \tag{2.39}$$

Le terme $\mathcal{L}(q, \theta)$ est appelé *énergie variationnelle libre*, et le terme $D_{KL}(q, p)$ est la divergence KL entre les fonctions $q(z)$ et $p(z|X; \theta)$. Cette divergence étant positive, l'énergie variationnelle libre est une minorante de la vraisemblance (avec cas d'égalité pour $q(z) = p(z|X; \theta)$). Le principe de l'algorithme est de maximiser l'énergie variationnelle libre par rapport à θ, que l'on peut réécrire :

$$\mathcal{L}(q, \theta) = Q(q, \theta) + \mathcal{H}(q), \tag{2.40}$$

où Q est l'espérance des données complètes :

$$Q(q, \theta) = \mathbb{E}_q(\log p(X, z|\theta)), \tag{2.41}$$

et le terme $\mathcal{H}(q)$ est l'entropie de la distribution q.

Originellement, l'algorithme EM consistait à considérer la "meilleure" minorante possible pour L, c'est-à-dire $\mathcal{L}(q, \theta)$, lorsque $D_{KL}(q, p) = 0$ dans (2.39). À une itération (it) donnée, on a donc $q^{(it)}(z) = p(z|X; \theta^{(it)})$. Maximiser L par rapport à θ revient donc à maximiser

$$Q(q^{(it)}, \theta) = \mathbb{E}_{q^{(it)}}(\log p(X, z; \theta)) = \mathbb{E}_{Z|X;\theta^{(it)}}(\log p(X, z; \theta)) \tag{2.42}$$

par rapport à θ, car l'entropie ne dépend pas de θ. Synthétiquement, l'algorithme EM consiste donc en l'alternance des étapes suivantes :

E Calcul de l'espérance : $\mathbb{E}_{Z|X;\theta^{(it)}}(\log p(X, z; \theta))$;

M Maximisation : $\theta^{(it+1)} = \arg\max_\theta \mathbb{E}_{Z|X;\theta^{(it)}}(\log p(X, z; \theta))$.

Cet algorithme permet d'estimer un grand nombre de modèles probabilistes, en se basant sur l'approche MM, garantissant ainsi la décroissance du critère de coût associé, c'est-à-dire la croissance de la vraisemblance.

EM Variationnel Lorsqu'il est difficile d'estimer la distribution à postériori $p(z|X; \theta)$, on peut chercher à réduire la divergence KL dans (2.39) plutôt que de l'annuler : on fait alors croître une minorante de L plutôt que L directement. Cette idée est le fondement des approches variationnelles (VEM) Beal et Ghahramani (2003), dans lesquelles q est recherchée dans

un sous-ensemble de densités ayant des propriétés intéressantes. L'approximation *mean field* consiste à écrire $q(z)$ comme un produit de densités : $q(z) = \prod_k q_k(z_k)$. Une telle approximation permet de découpler les variables latentes, ce qui rend les calculs plus aisés, et fournit des algorithmes rapides. Par exemple, le modèle de NMF à haute résolution Badeau (2011) a été initialement estimé par un algorithme EM, mais celui-ci conduisait à un temps de calcul trop élevé. Une approche par algorithme VEM a été proposée Badeau et Dremeau (2013) et mène à des performances similaires pour un coût en temps de calcul nettement moindre.

Algorithme SAGE Il est possible de simplifier les étapes E et M de l'algorithme EM en raisonnant sur les paramètres θ de la façon suivante. Notons l'ensemble des paramètres sous la forme $\theta = \{\theta_k\}$ et supposons que chaque variable latente Z_k ne dépende que du paramètre θ_k. On peut montrer que la maximisation de Q revient à maximiser le critère suivant en fonction de θ_k, pour tout k :

$$Q_k(\theta_k, \theta) = \mathbb{E}_{Z_k|X;\theta}(\log p(z_k; \theta_k)). \tag{2.43}$$

Lorsqu'il est plus simple d'estimer la distribution $p(z_k; \theta_k)$ pour chaque k plutôt que directement la loi jointe $p(z; \theta)$, cette approche rend les calculs plus aisés. En outre, à la différence de l'approche VEM, cette méthode (algorithme SAGE de l'anglais *Space Alternating Generalized EM* Fessler et Hero (1994)) garantit la croissance de la vraisemblance. Néanmoins, SAGE requiert une mise à jour séquentielle des différents paramètres (chaque couple d'étapes E-M est effectué en utilisant les valeurs des autres paramètres les plus actuelles). Ainsi, elle peut être plus lourde en temps de calcul que les approches variationnelles. Cette approche est notamment employée pour l'estimation de modèles NMF dans Févotte et al. (2009) et Bertin et al. (2010).

2.2.5 Extensions

Le modèle NMF (2.13) ne fournit généralement une décomposition satisfaisante que sur des signaux simples. Lorsqu'on traite des données réalistes et complexes, il est nécessaire d'enrichir ce modèle afin que la décomposition obtenue respecte certaines propriétés souhaitables. Nous présentons ci-après ces extensions dans les grandes lignes.

NMF informée

Une approche efficace pour améliorer la qualité de la décomposition est d'incorporer dans le modèle certaines informations sur les paramètres. Par exemple, il existe des cas où la matrice d'atomes spectraux W est connue : on parle alors de NMF semi-supervisée. En général, on a à disposition une base d'apprentissage constituée de sources séparées, à partir de laquelle il est aisé d'apprendre un dictionnaire W. Lors de l'estimation du modèle NMF sur un morceau test, on suppose alors connue la matrice W et on estime seulement H. Cette approche est par exemple utilisée dans Laroche et al. (2016) qui utilise des dictionnaires de percussions W adaptés au genre musical, ou dans Dessein et al. (2010) où les spectres de notes de piano sont préalablement appris sur des notes isolées.

Il est également possible d'incorporer une information externe sur la partition d'un morceau de musique pour aider la décomposition Hennequin et al. (2011b); Ewert et al. (2014a). Ces approches sont toutefois limitées à des morceaux pour lesquels cette information est disponible.

Enfin, la séparation de sources *informée* Liutkus et al. (2013) repose sur l'incorporation d'information sur les sources dans le signal de mélange, lors du procédé de mixage. Par exemple, les spectrogrammes des sources isolées peuvent être encodés dans le mélange et récupérés en sortie : la séparation est alors effectuée par filtrage de Wiener avec un masque

construit avec ces valeurs des spectrogrammes Liutkus et al. (2012). Afin de réduire la quantité d'information à transmettre, on approche les spectrogrammes des sources isolées par NMF à l'encodage.

Contraintes

Une façon naturelle de guider la décomposition consiste en l'incorporation de contraintes dans le modèle, sous la forme d'ajout de pénalités dans la fonction objectif. Celle-ci s'écrit sous la forme :

$$\mathcal{C}(W,H) = D(V, WH) + \sum_{c=1}^{C} \sigma_c \tilde{D}_c(W,H), \tag{2.44}$$

où D est un terme d'attache aux données (en général une β-divergence entre le modèle et les données) et les D_c sont des termes qui représentent un écart à un à priori, auquel est affecté un poids σ_c qui ajuste l'importance relative de chaque contrainte. D'un point de vue probabiliste, on peut voir $-D$ comme étant la log-vraisemblance et $-D_c$ comme les distributions à priori dans le cadre d'une estimation MAP du modèle (*cf.* section 2.2.4).

De telles contraintes assurent que la décomposition NMF possède certaines propriétés désirables. Parmi les plus courantes, citons :

— La parcimonie Hoyer (2004). Dans Smaragdis et Brown (2003), cette contrainte est introduite sous la forme d'un terme de pénalité dans la fonction de coût de la forme $||H||_p$ (norme p de H) avec $p \in]0, 2[$;

— La décorrélation des activations temporelles Zhang et Fang (2007) ;

— La régularité temporelle Virtanen (2007) qui conduit à des activations H lisses ;

— L'harmonicité Bertin et al. (2009a, 2010) ou l'inharmonicité (dans le cas du piano) Rigaud et al. (2013) des atomes spectraux W.

Les difficultés liées à la mise en oeuvre de ces contraintes sont par exemple la non-convexité du critère, certains problèmes numériques (convergence) ou encore le choix des poids σ_c. La contrainte de parcimonie est notamment très employée, mais son implémentation intuitive Smaragdis et Brown (2003) conduit, après normalisation de W et remise à échelle de H, à rendre la fonction de coût non-monotone. Le Roux et al. (2015) propose une nouvelle formulation de la NMF parcimonieuse qui permet d'éviter cet écueil.

Des contraintes mathématiques sur W et H peuvent également être formulées dans le but de réduire l'espace des solutions du problème NMF, comme par exemple la NMF orthogonale. Dans Laroche et al. (2015), une NMF structurée projective est proposée pour la séparation de sources harmoniques / percussives : les composantes harmoniques sont stockées dans une sous-matrice de W à composantes orthogonales, et les composantes percussives sont stockées dans les autres colonnes de W.

Variations d'enveloppes spectrales et temporelles

Le modèle NMF suppose la stationnarité des spectres des évènements sonores. Cette hypothèse n'est pas vérifiée par exemple pour des signaux comme ceux de parole ou d'instruments à cordes frottées, qui contiennent des vibratos (variations de fréquences fondamentales). En outre, supposer que H ne dépend pas de la fréquence revient à dire que tous les harmoniques constituant un atome ont la même enveloppe temporelle. Pour les signaux de piano par exemple, on sait que le coefficient d'amortissement d'amplitude dépend de l'harmonique considéré.

Pour dépasser ces limitations, certains travaux introduisent les variations de fréquences et d'enveloppes temporelles dans les modèles NMF. Le modèle de Durrieu (2011) représente les variations d'enveloppe spectrale de la voix. Hennequin et al. (2011a) propose un modèle de mélange de filtres (un par source) autorégressif à moyenne ajustée (ARMA), ce qui permet de représenter des signaux à fréquences fondamentales variables. Les modèles source-filtre Virtanen et Klapuri (2006); Bouvier et al. (2016) permettent également de représenter des signaux à fréquence variable, comme ceux de parole.

Utilisation de la phase

Le modèle NMF est basé sur la propriété d'additivité des données non-négatives (2.14), mais cette propriété n'est pas vérifiée lorsque plusieurs sources interfèrent dans un point du plan TF. Comment, dans ce cadre, intégrer des informations sur la phase ?

Ewert et al. (2014b) proposent d'introduire un masque de pondération pour pénaliser la fonction de coût aux points TF où il y a recouvrement de plusieurs sources. Après application d'un premier algorithme d'estimation de NMF sur un mélange, un masque est calculé à partir des énergies des sources estimées, afin d'identifier les zones du plan TF où les sources se recouvrent (ce qui revient à considérer les points où les sources ne sont pas en phase). Une nouvelle NMF, dite pondérée, est calculée en tenant compte de ce masque. Les règles de mises à jour multiplicatives pour la minimisation d'une telle fonction de coût sont fournies dans Blondel et al. (2007) pour la divergence KL et la distance euclidienne, et étendues aux β-divergences dans Limem et al. (2013). Des développements sur la NMF pondérée en ont notamment amélioré les techniques d'estimation Kim et Choi (2009).

La phase peut être exploitée de façon explicite pour dépasser ce problème de non-additivité des spectrogrammes. Dans Parry et Essa (2007), les auteurs calculent le spectrogramme d'un mélange de deux sources complexes dans le cas général où celles-ci ne sont pas en phase. L'expression du spectrogramme du mélange fait alors apparaître un terme de différence de phase entre les composantes, qui est supposée suivre une loi uniforme. Une telle démarche permet de raffiner l'estimation des spectrogrammes des sources séparées, mais est limitée à deux sources uniquement.

L'utilisation de la phase pour affiner l'estimation du spectrogramme dépasse par ailleurs le cadre de la séparation de sources par NMF. En effet, on pourra se référer par exemple à Gerkmann et Krawczyk (2013), où un estimateur MMSE du spectrogramme d'un signal de parole est obtenu à partir de la donnée de la phase du signal non bruité, dans un contexte de débruitage de la parole.

Ces approches exploitent la phase pour améliorer l'estimation des spectrogrammes, mais ne s'intéressent toutefois pas à sa reconstruction.

2.2.6 Clustering

En traitement du signal musical, ce que l'on entend par *source* dans l'expression *séparation de sources* possède différents sens. En effet, il peut s'agir d'un instrument, d'une note, où même d'un harmonique composant une note. Selon le rang de la factorisation dans le modèle NMF, on obtiendra donc une décomposition qui s'interprétera différemment. En général, une source est considérée comme étant une piste correspondant à un instrument de musique donné. Le j-ième instrument est donc constitué de la somme de K_j composantes dans la factorisation NMF WH. Si on note K le rang total de la factorisation et qu'il y a J instruments qui

composent le mélange, alors $\sum_j K_j = K$, et on cherche une partition de $[\![1, K]\!]$ sous la forme :

$$\{\mathcal{K}_j \subset [\![1, K]\!],\, j \in [\![1, J]\!] \;/\; \bigcup_{j=1}^{J} \mathcal{K}_j = [\![1, K]\!] \text{ et, } \forall i \neq j,\, \mathcal{K}_i \cap \mathcal{K}_j = \emptyset\}. \tag{2.45}$$

Le spectrogramme de la j-ième source est alors donné par $\sum_{k \in \mathcal{K}_j} W_k H_k$. Plusieurs approches existent pour obtenir une telle partition (ou *clustering*). Le clustering oracle Barker et Virtanen (2013) suppose la connaissance des sources isolées : chaque atome spectral W_k est comparé aux sources de référence et associé à celle avec laquelle il a la plus forte similarité (la comparaison se faisant en général par le calcul du SDR, sur lequel nous reviendrons dans la section 2.4). Cette méthode est notamment présentée dans Spiertz et Gnann (2009).

Dans un contexte où l'on ne connaît plus la vérité terrain sur les sources séparées, le clustering se fait par similarité spectrale : l'idée principale est de regrouper entre eux les atomes spectraux qui se "ressemblent". Dans Casey et Westner (2000), les auteurs proposent une mesure de similarité entre atomes spectraux basée sur la divergence KL. Spiertz et Gnann (2009) proposent d'utiliser les MFCC (*Mel Frequency Cepstral Coefficients*) qui permettent de caractériser le timbre des instruments.

Une autre direction consiste à effectuer une décomposition de type NMF "translatée" (*Shifted NMF* en anglais) Fitzgerald et al. (2005). Dans un tel modèle, les atomes composant la matrice spectrale W sont les translatés en fréquence d'atomes de référence. L'hypothèse sous-jacente est que les notes issues d'un même instrument ont la même enveloppe spectrale, et que seule la fréquence fondamentale change. Un tel modèle regroupe naturellement les atomes par instruments. L'hypothèse d'enveloppe spectrale constante n'est cependant pas toujours vérifiée en pratique et des développements ont été effectués en ce sens Jaiswal et al. (2011).

Ceci soulève quelques questions importantes, qui sont liées à des problématiques de sélection de modèles : comment choisir le rang de la factorisation K, le nombre de clusters (c'est-à-dire le nombre d'instruments) J, et la taille de ceux-ci ? Certaines réponses peuvent être trouvées par des approches non-paramétriques pour la sélection de modèle Gershman et Blei (2011), ainsi que par la fusion de modèles Jaureguiberry et al. (2013).

2.3 Estimation conjointe des spectrogrammes et des phases

Plutôt que d'effectuer une factorisation de spectrogramme par NMF d'une part, et de reconstruire la phase de chaque source dans un second temps, des modèles ont été proposés pour effectuer conjointement ces deux opérations.

2.3.1 NMF Complexe

Principe

La NMF complexe (CNMF) Kameoka et al. (2009) consiste à factoriser un spectrogramme d'amplitude tout en reconstruisant un champ de phases pour chaque source. Le mélange observé X est donc approché par le modèle $\hat{X}$ suivant :

$$\forall (f,t),\, \hat{X}(f,t) = \sum_{k=1}^{K} \hat{X}_k(f,t) = \sum_{k=1}^{K} W(f,k) H(k,t) e^{i\phi_k(f,t)}. \tag{2.46}$$

Il est à noter que le terme de "NMF complexe" peut prêter à confusion. En effet, le modèle de CNMF n'est ni un modèle de données non-négatives (on traite des coefficients complexes),

ni une factorisation au sens strict. Le modèle est estimé via la minimisation d'une fonction de coût qui est la norme de Frobenius de la différence entre les observations X et le modèle $\hat{X}$:

$$D(X, \hat{X}) = ||X - \hat{X}||_2^2 = \sum_{f,t} |X(f,t) - \hat{X}(f,t)|^2. \tag{2.47}$$

À ce terme est généralement ajoutée une pénalité afin de promouvoir la parcimonie des activations :

$$\mathcal{C}_s(H) = 2||H||_p^p = 2\sum_{k,t} |H(k,t)|^p, \tag{2.48}$$

où p est un paramètre de parcimonie (choisi entre 0 et 2). De plus amples détails sur la procédure d'optimisation et les algorithmes d'estimation de ce modèle peuvent être trouvés dans Sawada et al. (2011), et on pourra se référer à King et Atlas (2012) pour une implémentation de cette méthode.

Ce modèle combine les deux problématiques précédentes (séparation de spectrogrammes et reconstruction de phase) mais, comme nous le verrons dans le chapitre suivant, ne permet pas d'obtenir de résultats satisfaisants sans contraindre la phase. Un réglage fin des paramètres est nécessaire, mais conduit soit à obtenir une phase aléatoire (aucune forme de cohérence), soit à ce que chaque source possède la phase du mélange.

Néanmoins, dans le cas où les bases spectrales W sont apprises préalablement, un tel modèle peut conduire à des résultats intéressants, comme cela a été étudié dans King et Atlas (2010, 2011). On pourra se référer à la thèse de Brian King King (2012) qui a conduit un certain nombre de développements sur les factorisations de matrices complexes.

Contrainte de consistance

Le Roux et al. (2009) a proposé de contraindre le modèle précédent avec une contrainte de consistance. Le principal avantage de cette méthode est d'estimer conjointement amplitudes et phases, plutôt que d'estimer la phase depuis une amplitude imposée, comme dans la combinaison de la NMF et de l'algorithme de Le Roux et al. (2008c). Le spectrogramme est itérativement rendu de plus en plus consistant. La fonction de coût devient alors :

$$\mathcal{C}(W, H, \phi) = D(X, \hat{X}) + \lambda \mathcal{C}_s(H) + \gamma \sum_{k=1}^{K} \mathcal{I}(\hat{X}_k), \tag{2.49}$$

où $\mathcal{I}$ est la fonction d'inconsistance (2.5).

La méthode de la fonction auxiliaire présentée dans la section 2.2.4 est utilisée pour obtenir une procédure de mise à jour des paramètres. Nous utiliserons cette approche comme une référence dans notre étude comparative au chapitre 3.

Autres formes de contraintes

Enfin, d'autres contraintes ont été proposées pour la CNMF, basées sur la modélisation de signaux. Dans Bronson et Depalle (2014), un modèle de phase basé sur les mélanges de sinusoïdes est introduit afin de contraindre les composantes complexes dans le cadre d'une CNMF. La fonction de coût est donnée par :

$$\mathcal{C}(W, H, \phi) = D(X, \hat{X}) + \lambda \mathcal{C}_s(H) + \sigma \mathcal{C}_\phi(\phi), \tag{2.50}$$

où le terme $\mathcal{C}_\phi(\phi)$ traduit la fonction d'évolution de la phase selon un modèle sinusoïdal :

$$\mathcal{C}_\phi(\phi) = \sum_{t,k,r} \sum_{f \in \mathcal{N}_{k,r}} |e^{i\phi_k(f,t)} - e^{i\phi_k(f,t-1)} e^{i2\pi\nu_{0_k} rS}|^2, \tag{2.51}$$

où $\mathcal{N}_{k,r}$ est l'ensemble des canaux fréquentiels qui composent le lobe principal de la transformée de Fourier de la fenêtre d'analyse centrée autour de la fréquence réduite $\nu_{r_k} = r\nu_{0_k}$, ν_{0_k} étant la fondamentale, r l'indice d'harmonique et S est le décalage temporel (en échantillons) entre deux trames consécutives.

Cette approche, qui exploite la structure de signaux pour contraindre la phase, est développée dans le cadre de signaux strictement harmoniques, et requiert la connaissance de la fréquence fondamentale et du nombre d'harmoniques de chaque source. Cela la rend mal adaptée à la séparation de sources aveugle, ou lorsque les signaux diffèrent du modèle considéré (transitoires, vibratos, signaux percussifs, sinusoïdes amorties, mélanges non harmoniques, signaux à fréquences variables tels que la parole...).

Alternativement, l'approche de Kirchhoff et al. (2014) repose sur une hypothèse d'invariance de certains paramètres de phase des sources. Elle suppose en effet que les écarts de phase entre partiels sont constants au cours du temps, ce qui permet de structurer les phases des sources par cet invariant. Cette propriété est observée expérimentalement sur des sons de saxophone, et un modèle de mélange de sources complexes est obtenu. Néanmoins, celui-ci n'est estimé que lorsqu'il n'y a qu'un instrument (l'algorithme fourni n'est pas applicable à davantage de sources), mais cela montre l'intérêt d'exploiter les propriétés physiques des signaux pour contraindre les phases.

2.3.2 NMF Haute-Résolution

Le modèle de NMF à Haute Résolution (HRNMF) a été introduit par Badeau (2011). L'idée est de modéliser chaque bande de fréquences de la représentation TF X d'un signal par filtrage AR. Une telle technique, agissant sur les données complexes directement, capture naturellement les relations de phase et les dépendances temporelles des composantes. Le mélange est modélisé comme suit :

$$X(f,t) = n(f,t) + \sum_{k=1}^{K} \hat{X}_k(f,t), \tag{2.52}$$

où $n(f,t)$ est un bruit blanc gaussien. Chaque source $\hat{X}_k(f,t)$ est obtenue par filtrage temporel AR d'un signal $b_k(f,t)$:

$$\hat{X}_k(f,t) = b_k(f,t) + \sum_{p=1}^{P(k,f)} a_p(k,f)\hat{X}_k(f,t-p), \tag{2.53}$$

où $P(k,f)$ est l'ordre du filtre pour la source k dans le canal f, de coefficients $a_p(k,f)$. Enfin, $b_k(f,t)$ suit une loi normale centrée de variance $\sigma_k(f,t)^2$ telle que $\sigma_k(f,t)^2 = W(f,k)H(k,t)$, tous les $b_k(f,t)$ étant indépendants. Ce modèle permet de généraliser certaines approches :

- Si $P(k,f) = 0$ pour tout f et k, alors le modèle est équivalent au mélange de gaussiennes ISNMF décrit dans Févotte et al. (2009).
- Si $H(k,t) = 1$ pour tout t alors $\hat{X}_k$ est simplement un processus AR d'ordre $P(k,f)$ dans le canal fréquentiel f.
- Lorsque $H(k,t)$ est une impulsion, chaque source $\hat{X}_k$ peut être écrite comme un polynôme complexe qui correspond au modèle de sinusoïdes exponentielles (ESM) Badeau (2012); Badeau et al. (2006), fréquemment utilisé dans les méthodes à haute résolution, ce qui explique le nom du modèle HRNMF.

Les paramètres du modèle peuvent être estimés par un algorithme EM, qui est assez lourd en temps de calcul : la complexité est alors de $O(K^3FT(1+P)^3)$ où $P = \max_{k,f} P(k,f)$. Une approche variationnelle bayésienne (VBEM) Badeau et Dremeau (2013) permet un calcul plus rapide sans véritable perte de qualité : la complexité peut alors être diminuée jusqu'à $O(KFT(1+P))$. Alternativement, l'estimation du gradient de la fonction de coût a permis de remplacer le calcul de l'étape M par des mises à jours multiplicatives Badeau et Ozerov (2013), accélérant la vitesse de convergence de l'algorithme. Le choix de l'initialisation de l'algorithme est critique (*cf.* chapitre 3).

Le modèle HRNMF a été étendu au cas multicanal et aux mélanges convolutifs Badeau et Plumbley (2013a,b). Il a également été repris et généralisé dans Badeau et Plumbley (2013c) sous la forme d'un modèle probabiliste apte à représenter une plus grande variété de signaux (processus ARMA, bruits et transitoires d'attaque avec une haute résolution temporelle). Il est enfin à présent capable de modéliser les corrélations entre bandes de fréquences Badeau et Plumbley (2014).

2.4 Qualité de la séparation de sources

La mesure de la qualité de la séparation est encore aujourd'hui un problème ouvert. Cela s'explique par le fait que celle-ci est avant tout un critère perceptif et subjectif. La complexité des phénomènes perceptifs présents dans les signaux musicaux est donc difficilement synthétisable en un jeu d'indicateurs à vocation universelle.

Il existe principalement deux boîtes à outils qui fournissent de tels indicateurs. Le premier est un jeu d'indicateurs objectifs, alors que le deuxième est lié à des expériences subjectives. Globalement, ces méthodes visent à quantifier l'écart entre les sources réelles x_k (vérité terrain qui est disponible lorsqu'on travaille sur des bases de données, mais pas en pratique) et les sources estimées $\hat{x}_k$.

2.4.1 BSS Eval

BSS Eval (pour *Blind Source Separation Evaluation*) est une boîte à outils qui permet de calculer des critères objectifs de qualité de séparation de sources, à partir des sources originales et des sources estimées. Introduite dans Vincent et al. (2006), elle a été étendue au cas multicanal dans Vincent et al. (2007). Nous décrivons brièvement ici le principe de construction de ces indicateurs.

La différence entre les sources x_k et leurs estimées $\hat{x}_k$ est décomposée en trois composantes :

$$x_k - \hat{x}_k = e_k^{target} + e_k^{interf} + e_k^{artif}, \tag{2.54}$$

où les trois composantes sont respectivement l'erreur par rapport à la cible, la composante d'interférence et la composante d'artéfact. Ces composantes sont calculées par projection de $\hat{x}_k$ sur divers sous-espaces. Par exemple, e_k^{target} est obtenu par projection de $\hat{x}_k$ sur l'espace engendré par les x_l, $l \in [\![1, K]\!]$. Une fois ces composantes obtenues, on calcule divers rapports d'énergies qui quantifient la qualité de séparation, à partir de définitions similaires au rapport signal sur bruit :

— le SDR (*Signal to Distortion Ratio*) qui évalue la qualité globale de l'estimation,

— le SIR (*Signal to Interference Ratio*) qui mesure le rejet d'interférences,

— le SAR (*Signal to Artifact Ratio*) qui évalue le rejet d'artéfacts.

Ces indicateurs, largement utilisés par la communauté scientifique, permettent de comparer de nombreuses méthodes. Ils apparaissent comme des indicateurs significatifs des rejets d'artéfacts et d'interférences, même si certains phénomènes (comme le masquage perceptif) ne sont pas toujours bien pris en compte par ces quantités.

2.4.2 PEASS

Dans le but de proposer un ensemble de critères qui corresponde le plus possible à des appréciations subjectives, Emiya et al. (2011) a proposé la boîte à outils PEASS (pour *Perceptual Evaluation of Audio Source Separation*).

Similairement à BSS Eval, la première étape consiste à décomposer l'erreur entre sources originales et sources estimées en trois composantes de distorsion (*cf.* équation (2.54)). La technique pour obtenir ces composantes diffère néanmoins de celle employée dans BSS Eval, dans le but d'améliorer la cohérence avec la perception.

Ils proposent d'utiliser des scores perceptifs plutôt que des rapports d'énergie, qui d'après eux ne sont pas bien corrélés à la perception (les différences en basses fréquences notamment affectent grandement les rapports d'énergie mais peu la perception). Les critères subjectifs proposés sont au nombre de quatre :

— OPS : *Overall Perceptual Score*,

— TPS : *Target-related Perceptual Score*,

— IPS : *Interference-related Perceptual Score*,

— APS : *Artifacts-related Perceptual Score*.

Il faut donc relier les composantes de distorsion calculées précédemment avec ces scores. Pour cela, les auteurs proposent d'appliquer une fonction non-linéaire aux composantes pour obtenir les scores. La fonction est de forme sigmoïde, et ses paramètres sont appris en minimisant l'erreur quadratique entre les résultats calculés grâce à cette fonction et ceux obtenus par un test subjectif d'écoute. Des améliorations ont par ailleurs été apportées dans Vincent (2012) en jouant notamment sur les paramètres de cette fonction.

Des expériences ont montré que ces critères représentaient mieux que BSS Eval la perception humaine. Cette boîte à outils est moins bien adaptée à des morceaux de musique réalistes car très coûteuse en temps de calcul. Ainsi, nous avons choisi d'utiliser BSS Eval pour nos différents tests dans cette thèse, pour une raison de temps de calcul. Les résultats obtenus n'ont pas été significativement différents selon que l'on utilisait une boîte à outils plutôt qu'une autre.

2.5 Motivation

En conclusion de cette présentation des méthodes de reconstruction de phase dans les approches NMF pour la séparation de sources, nous résumons les principaux verrous scientifiques des méthodes de l'état de l'art, qui motivent la suite de cette thèse.

Tout d'abord, la méthode du filtrage de Wiener, qui consiste à attribuer la phase du mélange à chaque source, ne donne pas de bons résultats lorsque les sources se recouvrent dans le domaine TF, ce qui est pourtant fréquent en musique. Cette méthode conduit notamment à des interférences entre sources estimées, et à des artéfacts dans les basses fréquences, qui sont particulièrement marqués dans les pistes de basse et de batterie. On peut donc se demander si ces interférences et artéfacts proviennent de la seule estimation des spectrogrammes de puissance pour le calcul du masque de Wiener, ou bien si l'estimation de la phase joue également un rôle dans cette propriété du filtrage de Wiener.

Les approches par consistance, comme l'algorithme de Griffin et Lim ou de Le Roux, utilisent une propriété de la TFCT pour contraindre les phases des signaux estimés. Elles ont connu certains développements et ont été combinées au filtrage de Wiener. Mais peut-on assurer que la consistance est synonyme de qualité audio ?

Le cadre probabiliste est prometteur car il permet de modéliser l'incertitude sur des candidats potentiels de phase estimés par modèles de signaux. Ces approches s'avèrent complexes à mettre en oeuvre étant donné que les distributions de probabilités sur des variables circulaires (comme la phase) mènent à des modèles de mélanges dont on ne sait que rarement exprimer analytiquement les lois des variables latentes. Les modèles présentés dans la section 2.1.6 sont par ailleurs principalement appliqués au rehaussement de la parole, et pas à la séparation de sources musicales.

Enfin, le potentiel des approches qui estiment conjointement amplitudes et phases (*cf.* section 2.3) est encore incertain. La NMF complexe consistante repose sur une propriété de la TFCT, alors que le modèle de NMF à haute résolution utilise une structuration de la TFCT issue de modèles de signaux. Quels sont les potentiels de ces méthodes ? Quelle approche choisir ?

Chapitre 3

Étude comparative

Sommaire

Nous avons introduit, dans le chapitre précédent, les principales techniques de séparation de sources basées sur la NMF qui utilisent en complément de la séparation des spectrogrammes une technique de reconstruction de phase. Étant donné que ce dernier aspect a été nettement moins étudié ces dernières années dans le contexte de la séparation de sources, il nous a paru intéressant de comparer les principales approches existantes et d'identifier les avantages et inconvénients de chacune, afin de pouvoir orienter la suite de nos travaux vers des méthodes de reconstruction de phase performantes.

Pour cela, nous avons réalisé une étude comparative de diverses méthodes, sur plusieurs jeux de données, et avec deux approches, afin d'en mesurer non seulement les performances, mais également le potentiel d'amélioration. Il parait irréaliste de comparer exhaustivement tous les modèles. Nous avons donc retenu les approches basées sur la NMF "classique", c'est-à-dire sans injection de connaissance à priori et sans contrainte, afin de centrer spécifiquement l'étude sur la performance des méthodes en matière de reconstruction de phase. Nous avons également souhaité examiner le potentiel de certaines extensions de la NMF qui permettent la reconstruction de phase (NMF complexe et NMF à Haute Résolution). Nous avons donc étudié les méthodes suivantes :

— **NMF-Wiener** NMF avec filtrage de Wiener Févotte et al. (2009),
— **NMF-GL** NMF avec algorithme de Griffin et Lim Griffin et Lim (1984),
— **NMF-LR** NMF avec algorithme de Le Roux Le Roux et al. (2008c),
— **CNMF** NMF complexe non contrainte Kameoka et al. (2009),
— **CNMF-LR** NMF complexe avec contrainte de consistance Le Roux et al. (2009),
— **HRNMF** NMF à Haute Résolution Badeau et Plumbley (2014).

Il est à noter que les méthodes de CNMF avec contraintes de phase par modèles de signaux Bronson et Depalle (2014); Kirchhoff et al. (2014) n'ont pas été retenues pour cette étude : la première suppose la connaissance de certains paramètres (fréquences fondamentales et nombres d'harmoniques), et le deuxième modèle n'est estimé que dans le cas d'une seule source, et n'offre pas un cadre général de séparation. En outre, ces deux modèles reposent sur une hypothèse de mélanges harmoniques, plus restrictif que les autres méthodes.

Les principales conclusions de cette étude comparative ont fait l'objet d'une publication à la conférence ICASSP 2015 Magron et al. (2015d).

La section 3.1 présente la méthodologie employée dans cette étude : les différentes approches, les jeux de données ainsi que le protocole y sont décrits. Dans la section 3.2, nous nous intéressons au problème de l'initialisation et du choix de l'algorithme pour le modèle HRNMF. La section 3.3 détaille les résultats en termes de séparation de sources. La section 3.4 propose d'étudier le filtrage de Wiener consistant, et la section 3.5 s'intéresse à l'influence de la représentation TF utilisée. Enfin, nous effectuons un bilan de cette étude dans la section 3.6, et justifions ainsi l'orientation de la suite de nos travaux de thèse.

3.1 Méthodologie

3.1.1 Approches aveugle et oracle

Pour évaluer le potentiel (et donc les possibilités d'amélioration) de chaque méthode, nous avons comparé les résultats obtenus avec une approche aveugle et avec une approche oracle. L'approche aveugle consiste à estimer le modèle directement depuis le mélange de sources, sans utiliser d'à priori sur les sources isolées. L'approche oracle, quant à elle, consiste à évaluer la meilleure performance possible de chaque technique. Les paramètres du modèle sont appris sur les sources séparées. Ainsi, pour les méthodes **CNMF**, **CNMF-LR** et **HRNMF**, il n'y

a pas a proprement parler d'étape de séparation puisque les estimateurs des sources selon ces modèles sont calculés uniquement en utilisant les sources séparées et non le mélange. Pour la méthode **NMF-Wiener** (et donc en conséquence pour les approches consistantes qui utilisent **NMF-Wiener** comme initialisation), les modèles NMF sont appris sur les sources séparées, puis les sources sont estimées en appliquant un filtrage de Wiener au mélange : c'est ce qui correspond au bloc "séparation" sur le schéma de la figure 3.1 qui illustre ces approches. La comparaison entre les approches aveugle et oracle nous informe sur le potentiel et les possibilités d'amélioration de chaque méthode.

Il est à noter qu'il existe une approche intermédiaire, dite semi-supervisée. Par exemple, le dictionnaire d'atomes spectraux W peut être appris au préalable, et seules les activations H sont estimées. Cette approche, utile en pratique lorsqu'on connaît par exemple l'instrument qui a servi à produire les sons, n'est pas étudiée ici car on s'intéresse au potentiel de chaque méthode : l'approche oracle nous fournit cette information.

3.1.2 Données

Plusieurs jeux de données ont été utilisés :

— Des mélanges synthétiques de sinusoïdes harmoniques amorties, dont les amplitudes, les phases à l'origine, les fréquences et les coefficients d'amortissement sont aléatoires. Dans la moitié des cas, on force un recouvrement temps-fréquence.

— La base de données MAPS (*MIDI Aligned Piano Sounds*) Emiya et al. (2010) fournit de nombreuses données qui permettent de fabriquer des mélanges de sons de piano. Afin de tester les modèles sur des données réelles, nous avons considéré 30 mélanges de deux notes de piano tirées aléatoirement dans la base de données MAPS.

— Enfin, nous avons testé les modèles sur un court extrait MIDI d'un peu moins de 2 secondes. Il est composé de plusieurs occurrences de trois notes de basse et d'un accord de guitare, chacun de ces évènements étant représenté par un atome NMF (ainsi $K = 4$).

Pour les données synthétiques et de piano, chaque source est activée seule successivement, puis les deux sources sont ensuite activées simultanément. Un exemple de spectrogrammes de mélanges synthétiques (avec et sans recouvrement) est donné sur la figure 3.2.

Ces signaux sont simples. Ce choix de notre part est volontaire, car nous avons voulu utiliser des données qui permettent un contrôle précis des résultats. Notons enfin que dans ce chapitre, chaque atome NMF correspond à une source : nous ne sommes donc pas confrontés au problème du clustering de ces atomes.

3.1.3 Protocole

Il est important de préciser que pour le modèle HRNMF, nous avons choisi un ordre de filtrage autoregressif de 1 pour toutes les sources et les bandes de fréquences. Ainsi, ce modèle utilise deux fois plus de paramètres (dictionnaire d'atomes W et coefficients de filtrage a) que la NMF standard (W seulement). Pour que la comparaison soit plus équitable, nous avons donc calculé la TFCT avec deux fois plus de précision en travaillant sur la NMF standard. Notons que la CNMF utilise beaucoup plus de paramètres que les autres modèles (puisque les phases sont libres), mais il n'est pas nécessaire de régler le nombre de paramètres finement puisque comme nous le verrons, ce modèle fournit de moins bons résultats que les autres, alors qu'il utilise plus de paramètres.

Les modèles NMF (avec divergence KL) et CNMF sont estimés par 30 itérations de règles de mise à jour multiplicatives, et la reconstruction de phase est effectuée par 50 itérations (dans le cas des procédures itératives de GL et de LR). HRNMF est initialisé avec 30 itérations

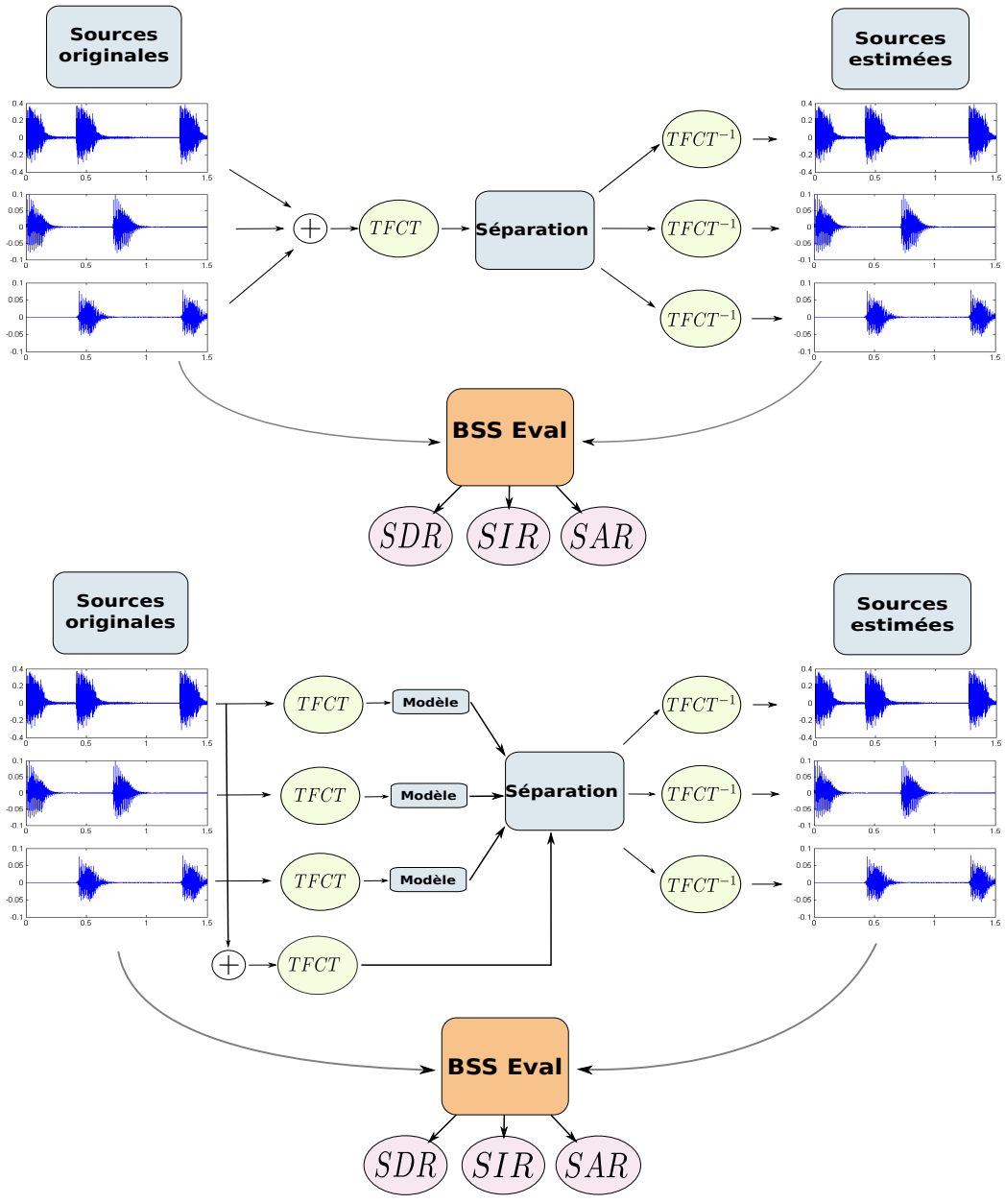


FIGURE 3.1 – Schéma de fonctionnement de notre étude. Deux approches complémentaires sont utilisées : une approche aveugle (en haut) et une approche Oracle (en bas).

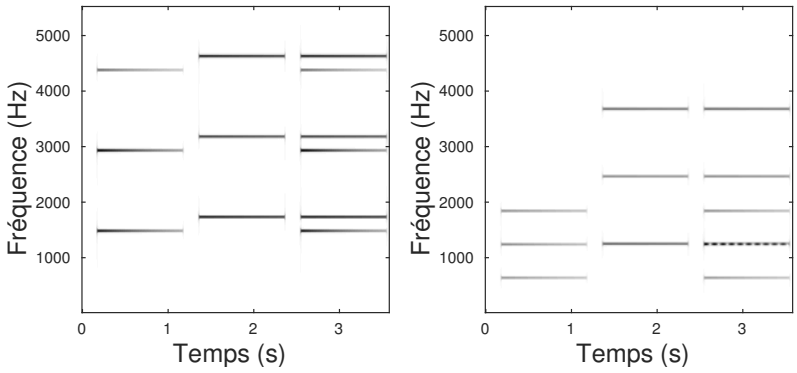


FIGURE 3.2 – Spectrogrammes de mélanges synthétiques : sans recouvrement TF (gauche) et avec recouvrement TF (droite).

de NMF et estimé par 30 itérations de l'algorithme VBEM (pour l'approche aveugle) et 10 itérations de cet algorithme pour chaque source (pour l'approche oracle). Ces nombres d'itérations sont choisis de sorte que la performance n'est pas améliorée au-delà. Enfin, les scores sont calculés sur 30 initialisations aléatoires afin de garantir la robustesse des résultats.

Afin de mesurer la qualité de la séparation de sources, nous utilisons la boîte à outils BSS EVAL VINCENT et al. (2006), un ensemble de critères objectifs qui sont adaptés à cette problématique. Notons que la boîte à outils PEASS EMIYA et al. (2011) a fourni des résultats similaires à BSS EVAL pour nos tests, nous avons donc ici retenu la première pour un critère de rapidité de calcul (*cf.* chapitre 2 section 2.4).

3.2 Initialisation et algorithme pour HRNMF

Le modèle HRNMF requiert une initialisation bien choisie pour produire des résultats satisfaisants, probablement à cause du grand nombre de minima locaux de la fonction de coût. Nous testons donc différentes initialisations : aléatoire, par KLNMF LEE et SEUNG (2001) ou par ISNMF FÉVOTTE et al. (2009), calculés à l'aide de règles multiplicatives (MUR). Nous comparons également les algorithmes Espérance-Maximisation (EM) BADEAU (2011) et EM variationnel Bayésien (VBEM) BADEAU et PLUMBLEY (2014). Les tests sont effectués sur des mélanges de notes de piano.

Précisons que pour cette expérience, ainsi que pour toutes celles conduites dans ce manuscrit, les simulations sont effectuées sur un ordinateur muni d'un CPU cadencé à 3.6 GHz et de 16 Go de RAM.

Algorithme	Initialisation	SDR	SIR	SAR	Temps (s)
EM	Aléatoire	5.3	6.4	14.3	379
	ISNMF	15.0	21.2	17.0	376
	KLNMF	17.0	22.2	18.7	377
VBEM	Aléatoire	1.4	2.8	11.1	1.03
	ISNMF	**16.9**	**25.3**	**17.7**	**0.95**
	KLNMF	**16.9**	**24.5**	**17.8**	**0.89**

TABLEAU 3.1 – Influence de l'initialisation et du choix de l'algorithme pour HRNMF sur la performance de séparation

Les résultats sont présentés dans le tableau 3.1, la meilleure performance étant mise en

valeur en gras. Nous remarquons qu'initialiser HRNMF avec une NMF améliore significativement les résultats par rapport à une initialisation aléatoire. Le choix d'une NMF avec divergence KL ou IS ne semble pas influencer grandement les résultats. Nous remarquons également que l'algorithme VBEM fournit des résultats similaires à EM, avec un gain très important en matière de temps de calcul. Nous utiliserons donc pour le reste de notre étude l'algorithme VBEM avec une initialisation KLNMF afin d'estimer le modèle HRNMF.

3.3 Résultats de séparation de sources

3.3.1 Mélanges synthétiques

Les résultats des tests sur les données synthétiques sont présentés sur la figure 3.3. Les boîtes à moustaches représentent les résultats de l'approche aveugle : chaque boîte à moustaches est constituée d'une ligne centrale indiquant la médiane des indicateurs, de bords inférieurs et supérieurs indiquant les 1^{er} et 3^{eme} quartiles, et les moustaches indiquent les valeurs extrémales. Les étoiles indiquent la performance de l'approche oracle.

Ces résultats montrent que les algorithmes de reconstruction de phase par approches consistantes (GL et LR) ne mènent pas à des résultats satisfaisants en ce qui concerne la qualité audio [1]. Ces algorithmes minimisent par construction l'inconsistance des composantes estimées, mais diminuent les SDR et SAR par rapport au filtrage de Wiener initial, diminution légère dans le cas aveugle mais nettement plus marquée dans le cas Oracle. Il est à noter que cette conclusion a déjà été suggérée dans une précédente étude Yoshii et al. (2013). Forcer l'amplitude à être constante (égale à une valeur cible) au cours des itérations semble être trop contraignant pour améliorer la qualité audio.

La NMF complexe avec contrainte de consistance **CNMF-LR** est supposée être une réponse à ce problème, puisque les spectrogrammes des sources sont ajustés au cours des itérations afin de compenser la contrainte de consistance, mais on constate en réalité que ce modèle ne conduit pas à une amélioration par rapport à **NMF-LR**. Nous observons que la NMF complexe non contrainte **CNMF** donne de meilleurs résultats que **CNMF-LR**, ce qui confirme que la consistance n'est pas forcément un critère adapté à la qualité audio.

Les résultats chutent globalement lorsque les sources se recouvrent dans le domaine TF, à l'exception du SAR : le rejet d'artefacts semble amélioré lorsqu'il y a recouvrement.

Enfin, la séparation aveugle avec le modèle HRNMF fournit des résultats légèrement meilleurs qu'avec les autres approches (excepté dans le cas de recouvrement, où les performances de **CNMF** et **HRNMF** sont similaires). Ce modèle fournit la meilleure performance dans la comparaison oracle. **NMF-Wiener** reste par contre la méthode la plus rapide (40 ms), les autres étant exécutées en environ 1.5 s. Les temps de calcul sont comparables sur les données de piano.

Remarque : Des tests complémentaires sur des mélanges synthétiques avec vibratos conduisent à des résultats similaires : le modèle HRNMF surpasse significativement les autres modèles dans la comparaison oracle, ce qui montre sa capacité à représenter une grande variété de signaux. À ce sujet, mentionnons qu'il peut être intéressant de travailler dans le domaine de modulation de spectrogramme afin de prendre en compte les variations d'amplitude et de fréquence des sources. Nous avons par ailleurs contribué à l'étude Stöter et al. (2016) qui proposait de comparer HRNMF et des méthodes de NMF dans le domaine de modulation de spectrogramme, montrant des résultats assez similaires.

1. Nous supposons ici que les indicateurs de SDR, SIR et SAR traduisent la qualité audio. Cette hypothèse est cependant sujette à controverse, et il est fréquent dans la littérature de voir la pertinence de ces indicateurs remise en question. Il faut donc garder à l'esprit que lorsqu'on se réfère ici à la "qualité audio", il est question de ces indicateurs.

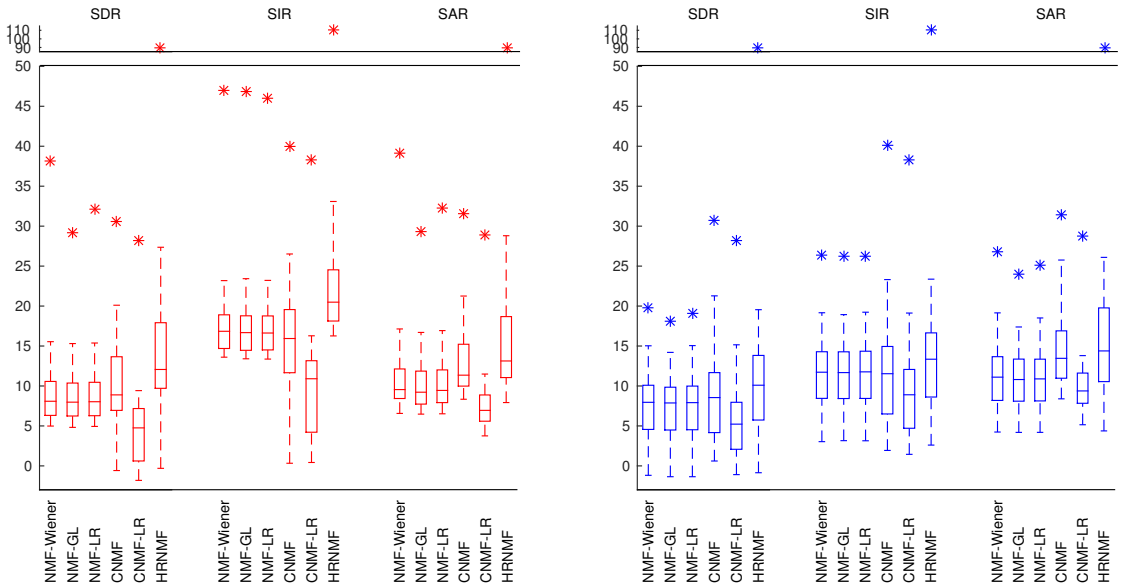


FIGURE 3.3 – Performance de la séparation de mélanges synthétiques sans recouvrement TF (gauche) et avec recouvrement TF (droite). Approches aveugle (boîtes à moustaches) et oracle (étoiles).

3.3.2 Notes de piano

Les résultats des tests sur les notes de piano sont présentés sur la figure 3.4. Les algorithmes ne conduisent pas à des performances particulièrement plus mauvaises que sur les données synthétiques, à l'exception de **CNMF**, dont la performance devient moins bonne que **NMF-Wiener**, inversement au cas des signaux synthétiques. Comme précédemment, le modèle HRNMF montre un potentiel très élevé par rapport aux autres méthodes (résultats oracle).

3.3.3 Extrait MIDI

La figure 3.5 présente les résultats expérimentaux sur un extrait MIDI. Ces résultats montrent une baisse significative des performances des algorithmes en comparaison avec les tests précédents. La complexité de ces signaux semble induire une baisse de qualité en termes de séparation de sources. L'estimation **HRNMF** n'améliore pas le résultat sur l'initialisation avec KLNMF en ce qui concerne les SDR et SIR dans le cas aveugle. Cependant, l'approche oracle montre toujours le potentiel de cette méthode. **NMF-Wiener** est estimé en 60 ms et les autres modèles entre 3 et 4 secondes.

3.3.4 En résumé

Les principaux résultats de cette étude comparative sont donc :

— Le modèle HRNMF possède le plus fort potentiel pour la séparation de sources, au vu des résultats de l'approche oracle. La modélisation des dépendances temporelles des composantes semble être une approche efficace pour améliorer la qualité de séparation.

— Ce modèle souffre néanmoins d'une estimation coûteuse en temps de calcul, malgré les efforts faits sur le sujet, notamment grâce à l'algorithme VBEM.

— Il y a une grande différence entre l'approche aveugle et oracle pour ce modèle. HRNMF semble bien fonctionner lorsque des informations sur les sources sont disponibles et

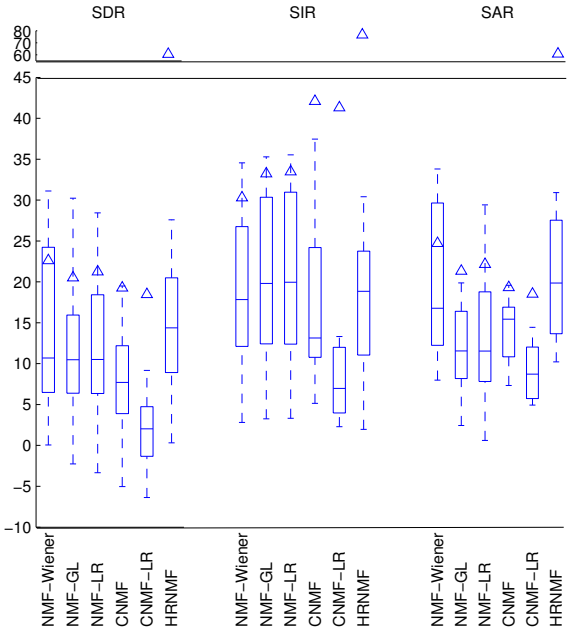


FIGURE 3.4 – Performance de la séparation de notes de pianos. Approches aveugle (boîtes à moustaches) et oracle (triangles).

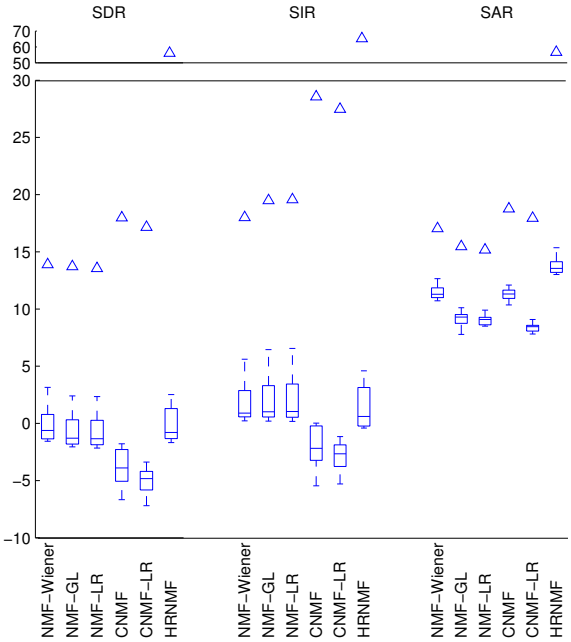


FIGURE 3.5 – Performance de la séparation des sources sur l'extrait MIDI. Approches aveugle (boîtes à moustaches) et oracle (triangles).

fonctionne moins bien en cas de séparation aveugle. Des applications en séparation supervisée peuvent donc être envisagées.

— Le filtrage de Wiener fournit un estimateur des sources (et donc implicitement de la phase) efficace et rapide. Néanmoins, lorsque les sources se recouvrent en temps et en fréquence, ses performances baissent significativement. Des phénomènes comme les battements créent alors des interférences entre sources.

— Les approches par consistance ne semblent pas adaptées à la séparation de sources car la consistance de la représentation ne s'avère en réalité pas être un critère corrélé à la qualité audio. Les contraintes de phase devraient donc reposer sur la consistance des données (comme le fait HRNMF) plutôt que sur la consistance de la représentation (ici la TFCT, ce que font GL et LR).

— La comparaison entre les résultats de la CNMF et de la CNMF consistante confirment ce diagnostic : contraindre les sources obtenues à être la TFCT d'un signal ne semble pas améliorer les SDR, SIR et SAR. Les NMF complexes ne fournissent par ailleurs pas de meilleurs résultats que les NMF traditionnelles, probablement en raison de la nature des contraintes (ou de l'absence de contrainte) sur les phases. La non-réduction de la dimensionnalité des données de phase est par ailleurs handicapante pour ces méthodes. Ces résultats ont déjà été partiellement observés précédemment (*cf.* King (2012)).

3.4 Filtrage de Wiener consistant

Les différentes méthodes qui consistent à combiner filtrage de Wiener et approche consistante ont été présentées dans le chapitre 2, section 2.1.4. Nous n'avons pas retenu ces approches dans notre comparatif puisque nous voulions évaluer indépendamment le potentiel des approches consistantes et du filtrage de Wiener dans le cas où il y a recouvrement TF des sources.

On peut néanmoins se demander si une approche qui combine phase du mélange et contrainte de consistante peut dépasser les performances de ces deux approches prises séparément, limites que nous venons d'identifier. D'après Le Roux et Vincent (2013), la méthode la plus aboutie, et celle qui fournit les meilleurs résultats parmi ces approches est le filtrage de Wiener consistant. Nous proposons donc de tester cette approche dans le cadre de la séparation de sources et de la comparer au filtrage de Wiener traditionnel et à l'algorithme GL.

On considère un jeu de données constitué de 30 mélanges de notes de piano qui se recouvrent dans le domaine TF. Nous appliquons donc les méthodes sus-citées à partir d'estimations des spectrogrammes obtenues par KLNMF sur les sources séparées (amplitudes oracle). Le filtrage de Wiener consistant dépend d'un paramètre γ ajustant l'importance relative de la contrainte de consistance, aussi nous faisons varier ce paramètre de 10^{-2} à 10^{7}. Les résultats moyennés sur la base de données sont présentés sur la figure 3.6.

On constate que pour une valeur du paramètre γ bien choisie (autour de 10^2), on obtient un compromis entre les différents indicateurs. Ceux-ci sont alors supérieurs aux valeurs obtenues par les deux méthodes (filtrage de Wiener et algorithme GL). Ce résultat montre l'intérêt d'une telle approche. Néanmoins, celui-ci est à relativiser : tout d'abord, l'amélioration des résultats reste modérée (le gain est de l'ordre de 0.1 dB en SDR et SAR, et de 0.2 dB en SIR). Par ailleurs, le paramètre γ optimal est fortement dépendant du jeu de données utilisé : en effet, dans les expériences conduites dans Le Roux et Vincent (2013), le paramètre optimal obtenu se situe autour de 10^6 alors qu'il est de 10^2 ici. Il est donc crucial, pour mettre efficacement en oeuvre ce type d'approche, de disposer d'une base d'apprentissage

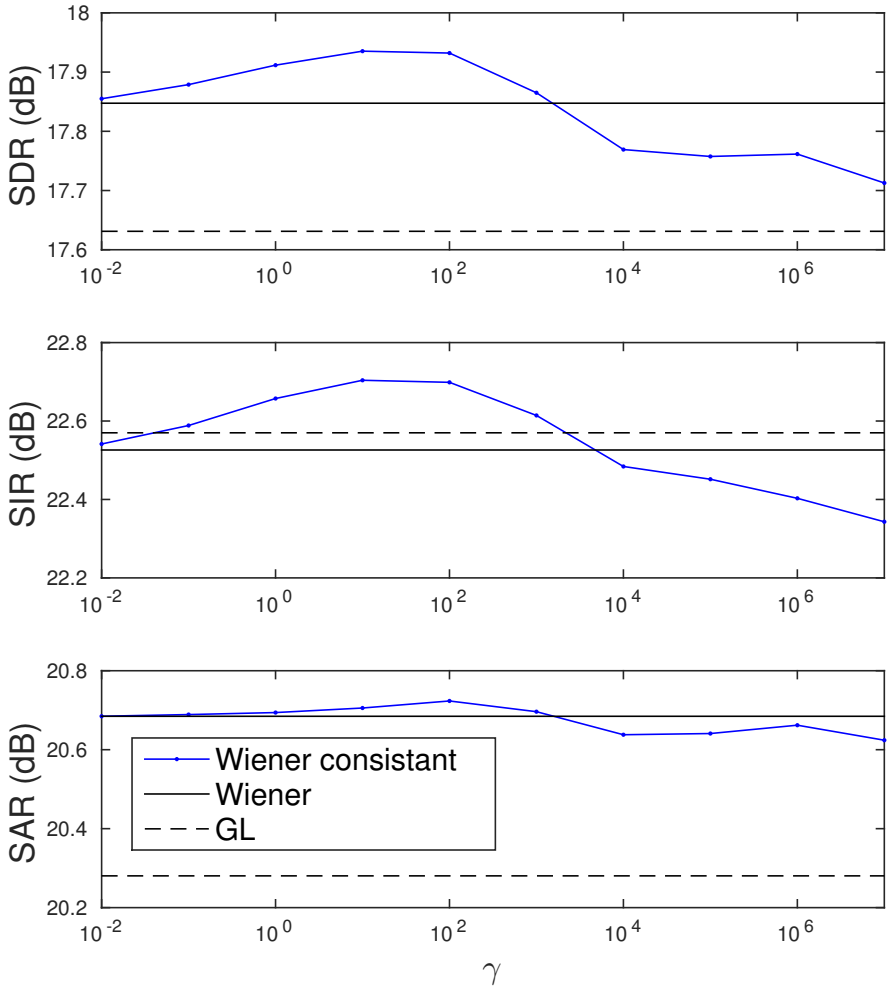


FIGURE 3.6 – Performance du filtrage de Wiener consistant en séparation de sources (mélanges de notes de piano).

relativement similaire à la base de tests. On pourrait par exemple choisir γ de sorte à le "lier" aux données, par exemple en prenant $\gamma = \tilde{\gamma}||X||_2$. Ce choix de définition du paramètre est notamment employé dans BRONSON et DEPALLE (2014) pour ajuster la contrainte de phase dans un modèle de NMF complexe, ou encore dans KAMEOKA et al. (2009) pour la contrainte de parcimonie.

Enfin, l'algorithme de filtrage de Wiener consistant est relativement lourd en matière de temps de calcul : sur un morceau plus réaliste (avec 4 sources instrumentales et d'une durée d'environ 10 secondes), le filtrage de Wiener consistant est effectué en environ 30 secondes contre moins d'une seconde pour le filtrage de Wiener traditionnel.

3.5 Influence de la transformation temps-fréquence

Les expériences précédentes montrent les limites des approches consistantes qui n'améliorent pas les performances en matière de qualité audio par rapport au filtrage de Wiener. Ainsi, ces approches étant basées sur la cohérence des informations redondantes de la trans-

formation TF utilisée (ici la TFCT), il est légitime de se demander si cette propriété de redondance bénéficie effectivement à la séparation de sources.

Nous avons donc complété cette étude par l'expérience suivante qui vise à comparer différentes représentations TF :

— La transformée en cosinus discrète modifiée (MDCT de l'anglais *Modified Discrete Cosine Transform*) Princen et Bradley (1986). Celle-ci a montré de bons résultats en séparation de sources musicales Plumbley et al. (2010). Elle permet notamment d'augmenter la parcimonie des sources Tan et Févotte (2005).

— La transformée de Fourier discrète modifiée (MDFT de l'anglais *Modified Discrete Fourier Transform*) Karp et Fliege (1999), qui vise notamment à s'affranchir des problèmes de recouvrement spectral inhérents à la transformée de Fourrier.

— La transformée à Q constant (CQT de l'anglais *Constant-Q Transform*) Fillon et Prado (2012) qui a été rendue inversible récemment Holighaus et al. (2013). Cette transformation est particulièrement adaptée au traitement du signal audio puisque sa résolution variable est adaptée à la perception humaine. Nous avons utilisé la boîte à outils MATLAB telle que décrite dans Schorkhuber et al. (2014).

Nous avons également étudié la TFCT avec plusieurs taux de recouvrement (0, 50 et 75 %). Les données utilisées et le protocole sont les mêmes que précédemment.

La figure 3.7 présente les résultats moyennés sur 30 signaux de mélanges de notes de piano (les résultats sont similaires sur les autres types de données). On constate qu'une séparation basée sur la TFCT semble donner les meilleurs résultats. Par ailleurs, le recouvrement de celle-ci influe sur les résultats : plus celui-ci est important, plus la séparation est de meilleure qualité, ce qui fait écho à la conclusion de Raki et al. (2005).

Ainsi, même si l'optimisation directe de la fonction d'inconsistance ne semblent pas améliorer les SDR, SIR et SAR par rapport au filtrage de Wiener initial, il semble que le taux de recouvrement, qui est à la base de ces approches, soit tout de même important. On peut donc suggérer que la consistance de la représentation puisse être utilisée dans un but de reconstruction de phase, mais peut-être pas via une optimisation directe de ce critère.

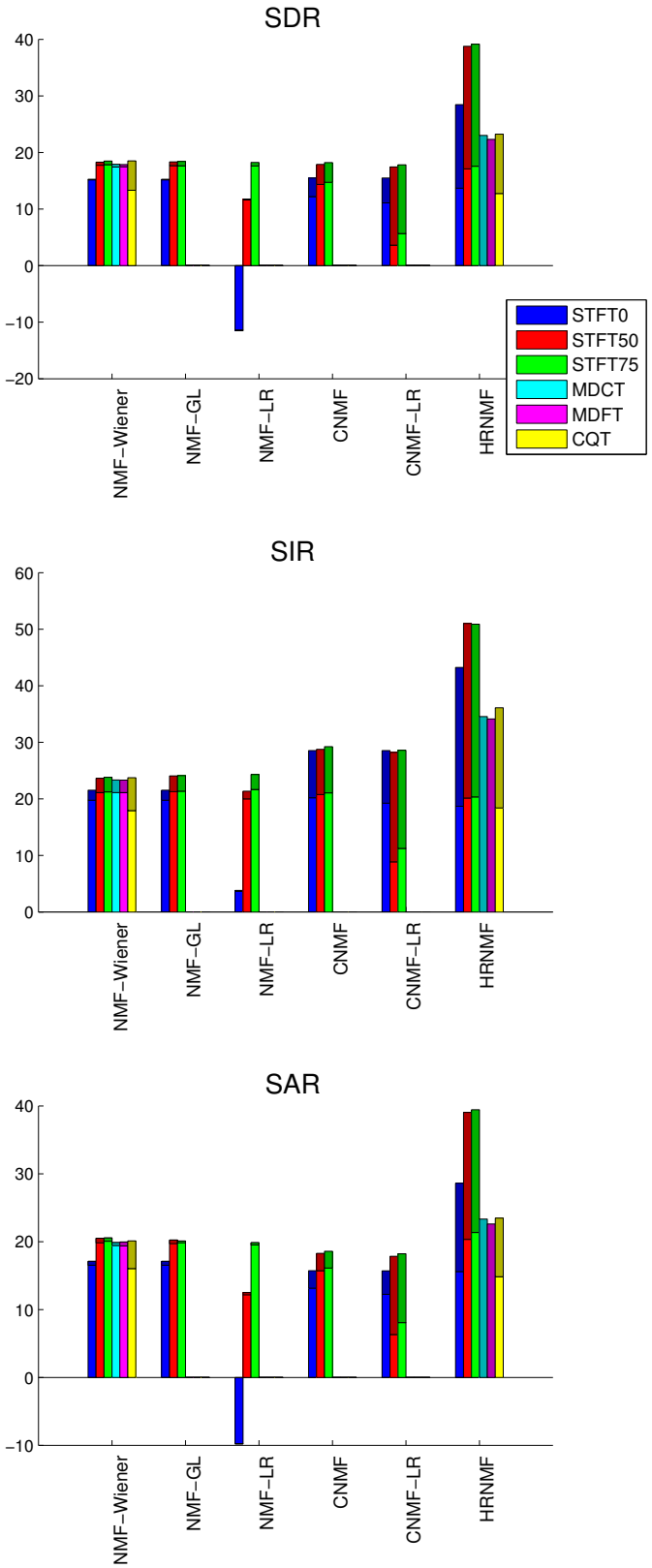


FIGURE 3.7 – Influence de la représentation temps-fréquence sur la performance de la séparation de sources (mélanges de notes de piano en relations harmoniques) : SDR (haut), SIR (milieu) et SAR (bas) en dB. Les barres claires représentent la performance aveugle, les barres foncées représentent la performance oracle.

3.6 Bilan de l'étude et approche

Les résultats de cette étude comparative soulignent la nécessité de mettre au point de nouvelles techniques de reconstruction de phase. En effet, les résultats Oracle de cette étude montrent que même lorsqu'un estimateur oracle du spectrogramme d'amplitude est disponible, la qualité de la séparation est toujours limitée par la méthode de restauration de phase utilisée.

L'utilisation d'une transformation redondante joue un rôle dans la cohérence du signal à travers les phases de sa TFCT, mais la propriété de consistance ne devrait pas être utilisée comme critère à maximiser pour reconstruire les phases. Le filtrage de Wiener est une technique rapide et efficace, mais lorsque les sources se recouvrent en temps et en fréquence, ses performances baissent significativement. Des phénomènes comme les battements créent alors des interférences entre sources, qui persistent dans le cas Oracle. Le filtrage de Wiener consistant combine ces deux aspects, mais n'améliore pas significativement les performances, même dans un cas Oracle, et est gourmand en temps de calcul. Il est à noter qu'une piste intéressante a été envisagée : elle consiste en une initialisation des algorithmes consistants qui exploite un modèle sinusoïdal Gnann et Spiertz (2010). Cela combine une propriété du signal et la consistance de la représentation.

Le modèle HRNMF tire son potentiel de la modélisation des signaux. La méthode d'estimation de ce modèle limite son emploi en pratique, même si des pistes d'amélioration peuvent être envisagées (méthodes à haute résolution Hua et al. (2004) ou méthodes MCMC Andrieu et al. (2003)).

La NMF complexe propose un cadre général utile car on peut facilement y inclure des contraintes. En ce sens, l'approche de Bronson et Depalle (2014) est prometteuse.

La modélisation des signaux confère à HRNMF son potentiel, à la NMF complexe une contrainte efficace Bronson et Depalle (2014), et aux approches consistantes une initialisation de qualité Gnann et Spiertz (2010). La suite de nos recherches portera donc sur la modélisation des signaux afin d'obtenir des contraintes de phase qui permettent de renforcer certaines propriétés désirables des signaux de musique (comme la précision des attaques ou la continuité temporelle). Ces contraintes pourront être intégrées dans le cadre de la séparation de sources, notamment dans un modèle de NMF complexe contrainte. Nous pourrons enfin mettre au point un modèle probabiliste de sources basé sur une phase non-uniforme, utilisant ce type de modèles.

Deuxième partie

Reconstruction de phase par modèles de signaux

Chapitre 4

Déroulé linéaire de phase par modèle sinusoïdal

Sommaire

Les conclusions du chapitre précédent ont orienté nos recherches sur la reconstruction de phase par modèles de signaux. Nous nous intéressons dans ce chapitre au modèle de mélange de sinusoïdes McAuley et Quatieri (1986), qui est fréquemment utilisé dans la littérature, comme par exemple dans l'algorithme du vocodeur de phase Flanagan et Golden (1966), la séparation de sources par NMF contrainte Bronson et Depalle (2014) ou encore le rehaussement de la parole Mowlaee et Kulmer (2015).

Nous proposons une généralisation de cette approche qui consiste à contraindre les phases de signaux musicaux dans le domaine TF par un modèle de mélanges de sinusoïdes. Nous obtenons un algorithme de déroulé de phases *horizontal*, à travers les trames temporelles. Notre technique s'applique à plusieurs types de signaux musicaux, tels que des sons de guitare ou de piano. L'estimation locale (dans chaque trame) des fréquences instantanées [1] étend le domaine de validité de cette méthode aux signaux non stationnaires comme des sons de violon ou de parole.

Les principaux résultats liés à ce travail ont fait l'objet d'une publication à la conférence EUSIPCO Magron et al. (2015b), et un rapport technique plus détaillé a été déposé dans la base de données de Télécom ParisTech Magron et al. (2015c).

La section 4.1 présente le modèle sinusoïdal à partir duquel est obtenu le déroulé horizontal. La section 4.2 présente une évaluation expérimentale de cette technique, et la section 4.3 applique cette méthode à la suppression de clics dans les enregistrements audio. Nous introduisons dans la section 4.4 un modèle d'impulsion pour la reconstruction de phase dans les trames d'attaque, avant de conclure dans la section 4.5.

4.1 Modèle sinusoïdal

4.1.1 Signal stationnaire

Considérons une sinusoïde complexe de fréquence instantanée $\nu_0 \in]-\frac{1}{2},\frac{1}{2}]$, de phase à l'origine $\phi_0 \in]-\pi,\pi]$ et d'amplitude $A_0 > 0$:

$$\forall n \in \mathbb{Z},\ x(n) = A_0 e^{2i\pi\nu_0 n + i\phi_0}. \tag{4.1}$$

On rappelle l'expression de la TFCT, pour chaque bande de fréquences $f \in [\![0, F-1]\!]$ et trame temporelle $t \in \mathbb{Z}$:

$$X(f,t) = \sum_{n=0}^{N_w-1} x(n+tS)w(n)e^{-2i\pi\frac{f}{F}n}, \tag{4.2}$$

où w est une fenêtre d'analyse de longueur N_w échantillons et S est le décalage temporel entre deux trames. Soit $W(\nu) = \sum_{n=0}^{N_w-1} w(n)e^{-2i\pi\nu n}$ la Transformée de Fourier à Temps Discret (TFTD) de la fenêtre d'analyse w à la fréquence réduite $\nu \in]-\frac{1}{2},\frac{1}{2}]$. La TFCT de la sinusoïde (4.1) est :

$$X(f,t) = A_0 e^{2i\pi\nu_0 St + i\phi_0} W\left(\frac{f}{F} - \nu_0\right). \tag{4.3}$$

On note $\phi = \angle X$ la phase de X ($\angle$ désigne l'argument complexe). Elle s'écrit alors sous la forme :

$$\phi(f,t) = \phi_0 + 2\pi S\nu_0 t + \angle W\left(\frac{f}{F} - \nu_0\right). \tag{4.4}$$

1. Dans ce manuscrit, l'expression "fréquence instantanée" désigne une estimation de la fréquence dans une trame t : il s'agit en toute rigueur de la fréquence instantanée (définie pour tout échantillon n dans le domaine temporel) moyenne dans la trame d'analyse. Néanmoins, nous utiliserons l'expression "fréquence instantanée" par commodité de langage.

Cela conduit à une relation entre points TF successifs :

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu_0. \tag{4.5}$$

On voit qu'une telle équation permet, dans un canal fréquentiel donné, d'estimer la phase dans une trame t en fonction de la phase dans la trame précédente et de la fréquence instantanée de la sinusoïde ν_0.

L'approche que nous proposons est donc similaire à l'étape de synthèse du vocoder de phase : nous estimons les fréquences instantanées pour en déduire l'incrément de phase $2\pi S\nu_0$. La différence est que le vocodeur de phase utilise les différences entre phases (supposées connues à l'analyse) pour calculer la fréquence instantanée (et dérouler ensuite une phase de synthèse), alors que nous proposons d'estimer cette fréquence par une méthode alternative, qui n'utilise que les amplitudes.

4.1.2 Sinusoïdes multiples

Considérons à présent un mélange de P sinusoïdes de paramètres notés A_p, ν_p et $\phi_{p,0}$:

$$x(n) = \sum_{p=1}^{P} A_p e^{2i\pi\nu_p n + i\phi_{p,0}}. \tag{4.6}$$

Sa TFCT s'écrit :

$$X(f,t) = \sum_{p=1}^{P} A_p e^{2i\pi\nu_p St + i\phi_{p,0}} W\left(\frac{f}{F} - \nu_p\right). \tag{4.7}$$

Nous supposons qu'il y a au plus une sinusoïde active par bande de fréquences, ce qui signifie que dans une bande de fréquences donnée, la TFCT X peut simplement s'écrire comme étant égale à la contribution d'un seul partiel (c'est-à-dire une composante sinusoïdale). Cette hypothèse est peu réaliste pour des signaux de musique réalistes où plusieurs sources se recouvrent dans le plan TF, mais nous opérerons dans ce cas de figure (*cf.* chapitre 5) sur les sources séparées.

Nous proposons de découper l'espace des canaux fréquentiels en régions, dites *régions d'influence* Laroche et Dolson (1999), pour s'assurer que la phase dans un canal fréquentiel donné soit déroulée selon la fréquence instantanée appropriée.

Dans la trame t, on observe donc une amplitude $|X(f,t)|$ que nous notons $V(f)$ (on s'affranchit de l'indice de trame par souci de lisibilité). Les canaux qui correspondent aux pics d'amplitude sont notés f_p. Nous définissons les limites des régions d'influence comme suit :

$$\forall p \in [\![2, P]\!],\ l_p = \left\lfloor \frac{V(f_p)f_{p-1} + V(f_{p-1})f_p}{V(f_p) + V(f_{p-1})} \right\rfloor, \tag{4.8}$$

où $\lfloor . \rfloor$ désigne la partie entière, et $l_1 = 0$, $l_{P+1} = F$. On définit alors la p-ième région d'influence :

$$I_p = [\![l_p, l_{p+1} - 1]\!]. \tag{4.9}$$

Une telle définition présente deux avantages. Tout d'abord, plus le pic $V(f_p)$ est important devant ses voisins, plus étendue sera la région d'influence correspondante. En outre, cette définition assure que l'ensemble des régions d'influence forme une partition de l'ensemble des canaux fréquentiels :

$$\forall p \neq q,\ I_p \cap I_q = \emptyset \text{ et } \bigcup_{p=1}^{P} I_p = [\![0, F-1]\!], \tag{4.10}$$

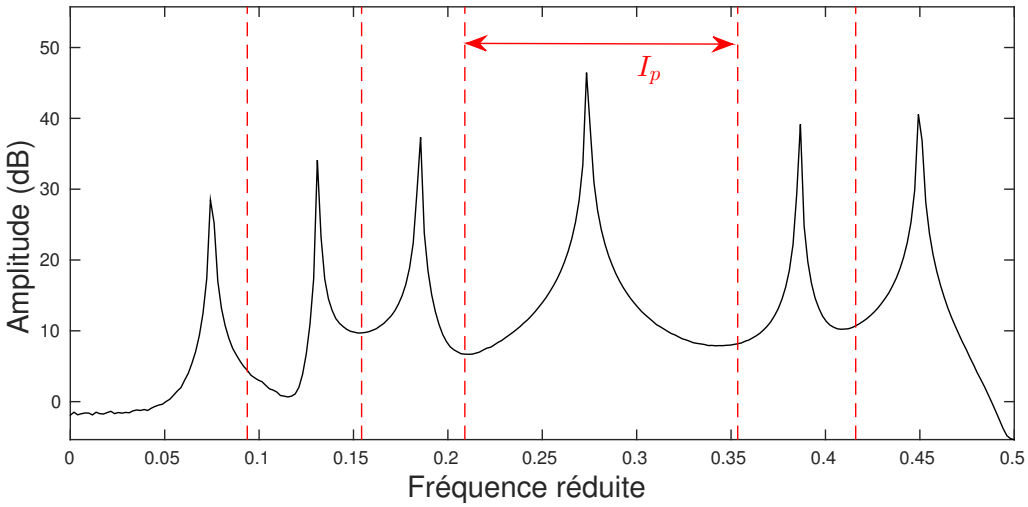


FIGURE 4.1 – Découpage en régions d'influence : un spectre (courbe en traits pleins) est segmenté en régions d'influence, qui sont d'autant plus larges qu'un pic est plus important que ses voisins. Les traits en pointillés représentent les frontières entre ces régions.

ce qui signifie que traiter toutes les régions I_p permet de traiter la totalité des canaux fréquentiels. Ce découpage en régions d'influence est illustré sur la figure 4.1.

D'autres choix de régions d'influence sont possibles LAROCHE et DOLSON (1999). Par exemple, la limite entre deux pics d'amplitude peut être le canal de plus petite énergie. Toujours d'après LAROCHE et DOLSON (1999), on peut choisir comme limite entre deux régions d'influence le milieu entre deux canaux fréquentiels correspondant aux pics d'amplitude consécutifs. Nous avons choisi d'utiliser la définition (4.9) pour sa simplicité et sa facilité d'implémentation.

Ainsi, si nous considérons à présent un canal dans la p-ième région d'influence, la TFCT X (4.7) devient :

$$\forall f \in I_p,\ X(f,t) \approx A_p e^{2i\pi\nu_p St + i\phi_{p,0}} W\left(\frac{f}{F} - \nu_p\right), \tag{4.11}$$

ce qui conduit à :

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu_p. \tag{4.12}$$

Nous pouvons donc proposer l'équation de *déroulé linéaire* suivante, qui généralise (4.5) :

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f), \tag{4.13}$$

telle que $\forall p \in [\![1,P]\!]$, $\forall f \in I_p$, $\nu(f) = \nu_p$.

4.1.3 Signal à fréquence variable

On peut calculer la phase de la TFCT d'un signal dont la fréquence instantanée varie au cours du temps (pour un vibrato par exemple). Le calcul est conduit dans ABE et SMITH (2005) pour des signaux continus, et peut être étendu aux signaux à temps discret : si la variation de fréquence entre deux trames consécutives $t-1$ et t est petite devant la largeur d'un canal fréquentiel, c'est-à-dire si un pic d'amplitude reste dans le même canal fréquentiel, alors nous pouvons généraliser (4.13) :

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f,t). \tag{4.14}$$

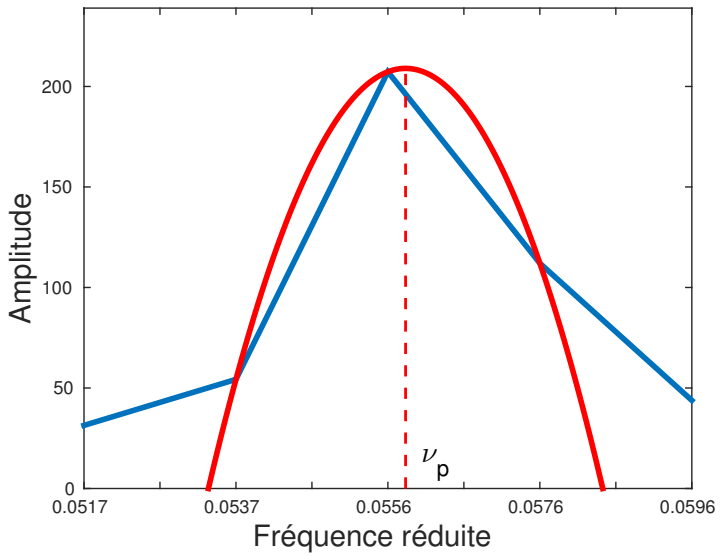


FIGURE 4.2 – Illustration de la QIFFT : un pic d'amplitude est approché par une parabole, et le calcul du maximum de cette parabole conduit à une estimation de la fréquence instantanée.

La fréquence instantanée est alors estimée dans chaque trame temporelle, afin de représenter des signaux à fréquence variable tels que les vibratos, qui sont souvent présents en musique (signaux de voie chantée ou de violon par exemple).

Notons enfin que l'on peut toujours appliquer ce résultat lorsque la variation de fréquence instantanée devient plus importante LAROCHE et DOLSON (1999), et que le canal fréquentiel correspondant au p-ième pic devient variable au cours du temps : on note ce canal $f_p(t)$. On estime alors la phase dans le point TF $(f_p(t), t)$ à partir de la phase dans le point TF $(f_p(t-1), t-1)$:

$$\phi(f_p(t), t) = \phi(f_p(t-1), t-1) + 2\pi S\nu(f_p(t), t). \tag{4.15}$$

4.1.4 Estimation de fréquences instantanées

Pour estimer les fréquences instantanées, nous utilisons la technique d'interpolation quadratique de FFT (QIFFT pour *Quadratic Interpolated FFT*) ABE et SMITH (2004a). Cette méthode consiste à approcher la forme d'un spectre au voisinage d'un pic d'amplitude par une parabole. Cette approximation parabolique est justifiée théoriquement pour des fenêtres d'analyse gaussiennes MARQUES et ALMEIDA (1986), et utilisée en pratique pour n'importe quel type de fenêtre ABE et SMITH (2004a). Le calcul du maximum de cette parabole fournit l'estimation de la fréquence instantanée. Il est à noter que cette méthode n'est valable que si une seule sinusoïde est active par bande de fréquences. La figure 4.2 illustre cette technique.

Le biais de cette méthode dépend d'une part du type de fenêtre, et d'autre part du nombre de points utilisés pour le calcul de la transformée de Fourier. Dans ABE et SMITH (2004b), les auteurs donnent des méthodes pour réduire arbitrairement ce biais, notamment en utilisant du bourrage de zéro. Des méthodes plus poussées existent pour l'estimation de fréquences instantanées dans le domaine TF. Néanmoins, celles-ci sont généralement basées sur l'hypothèse de mélanges harmoniques (somme et produit harmonique spectral, ou de façon plus sophistiquée, l'algorithme PEFAC GONZALEZ et BROOKES (2014)), ou bien agissent sur des données complexes. Ainsi, la QIFFT semble être un choix approprié dans notre cadre d'étude.

Algorithme 2 Reconstruction de phase par déroulé linéaire

Entrées :
Spectrogramme d'amplitude $V \in \mathbb{R}_+^{F \times T}$,
Trames d'attaque t_m, $\forall m \in [\![0, M]\!]$,
Phases d'attaque $\phi(f, t_m)$, $\forall m \in [\![0, M-1]\!]$.
pour $m = 0$ à $M-1$ **faire**
  pour $t = t_m + 1$ à $t_{m+1} - 1$ **faire**
    Calculer $v(f) = V(f, t)$.
    Localisation de pics f_p à partir de $v(f)$.
    Fréquences instantanées ν_p par QIFFT autour des pics f_p.
    Régions d'influence I_p à partir des pics f_p et des amplitudes $v(f_p)$.
    Attribution des fréquences $\forall f \in I_p$, $\nu(f) = \nu_p$.
    Déroulé de phase $\forall f$, $\phi(f, t) = \phi(f, t-1) + 2\pi S \nu(f)$.
  fin pour
fin pour
Sortie :
$\phi \in \mathbb{R}^{F \times T}$

4.1.5 Algorithme de déroulé horizontal

Nous présentons dans l'algorithme 2 la procédure de reconstruction des phases d'une TFCT à partir de son spectrogramme d'amplitude. On suppose connues les trames d'attaque (qui peuvent être calculées, par exemple, à partir du spectrogramme via la boîte à outils MATLAB Tempogram Toolbox Grosche et Müller (2011)). La détection des transitoires d'attaque est en effet une problématique qui dépasse le cadre de cette thèse, et nous ne nous y sommes pas intéressés directement : il existe en effet de multiples méthodes pour les estimer (on pourra trouver dans Daudet (2005) une présentation de diverses méthodes d'extraction de transitoires d'attaque). On note ces trames t_m avec $m \in [\![0, M-1]\!]$, où M est le nombre d'attaques. Pour éviter tout problème d'indices au bord, on note également $t_M = T$, ainsi $t_M - 1 = T - 1$ désigne la dernière trame de la TFCT. Les phases dans les trames d'attaque doivent être fournies à l'algorithme, puisque celui-ci repose sur une relation récursive. Dans les expériences conduites dans la prochaine section, elles seront supposées connues, mais on s'intéressera par la suite à des méthodes alternatives d'estimation de ces phases d'attaque.

4.2 Évaluation expérimentale

4.2.1 Protocole et données

La boîte à outils MATLAB Tempogram Grosche et Müller (2011) fournit une estimation rapide et robuste des trames d'attaque[2] à partir d'un spectrogramme. Nous utilisons plusieurs jeux de données :

A : 30 morceaux de piano tirés de la base MAPS Emiya et al. (2010) ;

B : 6 morceaux de guitare extraits de la base IDMT-SMT-GUITAR Kehling et al. (2014) ;

2. En réalité, cette boîte à outils est concue pour évaluer le tempo. Néanmoins, elle calcule l'ensemble des trames d'attaque de façon intermédiaire, c'est donc l'information que nous avons exploitée.

Données	A	B	C	D
Erreur $\tilde{\epsilon}$ (%)	0.48	0.62	0.58	0.35

TABLEAU 4.1 – Erreur entre estimées de fréquence par QIFFT et vocodeur de phase sur plusieurs jeux de données.

C : 12 quatuors à cordes tirés de la base de données SCISSDB (*SCore Informed Source Separation DataBase*) HENNEQUIN et al. (2011b) ;

D : 40 extraits de parole de la base ChiME (*Computational Hearing in Multisource Environments*) BARKER et al. (2013) ;

E : 50 morceaux de musique de divers styles (pop, rock, électronique...) issus de la base DSD100 (*Demixing Secret Database*) : cette base de données est une version remasterisée de la base mise à disposition pour la campagne SiSEC (*Signal Separation Evaluation Campaign*) ONO et al. (2015).

Les signaux sont échantillonnés à $F_s = 44100$ Hz. La boîte à outils BSS EVAL VINCENT et al. (2006) est utilisée pour évaluer la performance de la reconstruction : on quantifie celle-ci en calculant le SDR entre le signal original et son estimé. L'algorithme itératif de Griffin et Lim (GL) est utilisé comme référence, 200 itérations de cet algorithme étant effectuées (la performance n'étant pas améliorée au-delà). Il est initialisé avec des phases aléatoires, sauf dans les trames d'attaque où la phase est supposée connue.

4.2.2 Estimation des fréquences instantanées

Dans cette expérience, nous évaluons la qualité de la technique de QIFFT pour estimer les fréquences instantanées. La TFCT est calculée avec une fenêtre de Hann de longueur 4096 échantillons (soit 92 ms), 75 % de recouvrement et pas de bourrage de zéros.

On considère dans un premier temps des signaux synthétiques constitués de mélanges de sinusoïdes. De tels signaux nous permettent de connaître la vérité terrain (les fréquences instantanées). On les compare alors aux valeurs estimées par QIFFT. Les signaux contiennent en moyenne 40 harmoniques, et on effectue cette tâche sur 50 signaux. L'erreur moyenne d'estimation des fréquences est de 0.002 %, ce qui montre l'efficacité de la QIFFT pour l'estimation de fréquences instantanées de signaux sinusoïdaux.

On effectue une expérience similaire sur des signaux réalistes (données A à D). On note $\nu^*(f,t)$ l'estimée de la fréquence instantanée par QIFFT au point temps-fréquence (f,t) et $\nu(f,t)$ sa valeur calculée grâce à la technique du vocoder de phase LAROCHE et DOLSON (1999), c'est-à-dire en supposant la phase connue. Notons que cette estimation est une référence (et non pas la vérité terrain), le but dans cette expérience étant d'évaluer la différence entre un estimateur basé sur la phase (vocoder) et un estimateur basé sur l'amplitude (QIFFT).

La figure 4.3 illustre un spectrogramme de signal qui comporte des vibratos marqués, ainsi que les fréquences instantanées estimées par ces deux méthodes. Les deux méthodes conduisent à un résultat similaire.

L'erreur relative moyenne d'estimation en fréquence est :

$$\tilde{\epsilon} = \frac{1}{|\Upsilon|} \sum_{(f,t)\in\Upsilon} \frac{|\nu^*(f,t) - \nu(f,t)|}{\nu(f,t)}, \tag{4.16}$$

où Υ désigne l'ensemble des points du plan TF qui correspondent aux pics détectés et $|\Upsilon|$ désigne le cardinal de l'ensemble Υ.

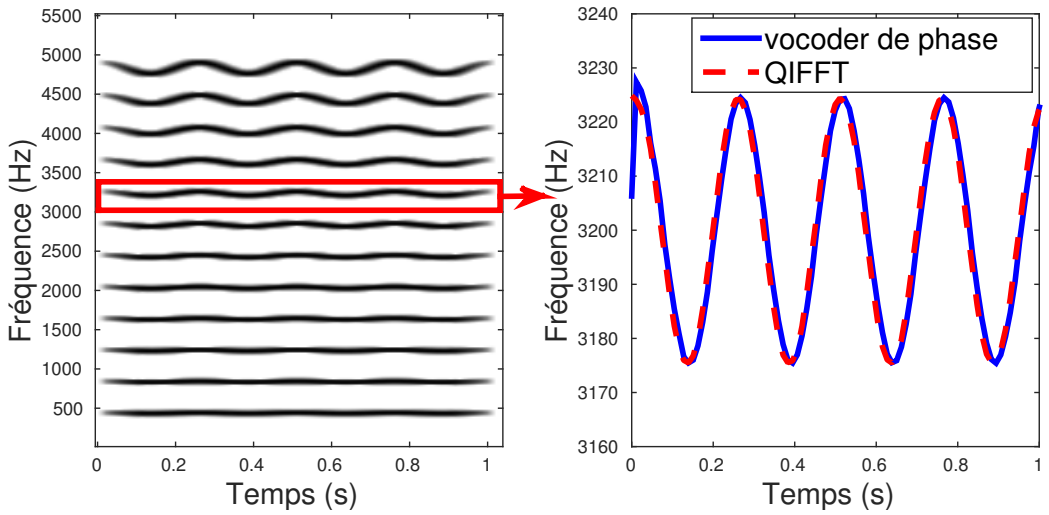


FIGURE 4.3 – Spectrogramme d'un mélange synthétique avec vibrato (gauche) et fréquences instantanées correspondant au partiel oscillant autour de 3200 Hz (droite).

Dans le tableau 4.1, on peut lire l'erreur d'estimation moyenne (4.16) pour différents jeux de données. Ces résultats confirment que la QIFFT conduit à une estimation de fréquence très proche de celle obtenue en utilisant l'information de phase. Ce résultat confirme les travaux plus extensifs de BETSER et al. (2008).

Le choix de l'estimation de fréquence par la méthode du vocodeur de phase comme valeur de référence est tout à fait arbitraire : cette valeur n'est en effet pas égale à la vérité terrain, indisponible ici. Cette expérience amène donc à la conclusion que l'estimation de fréquences instantanées n'utilisant qu'une information d'amplitude (QIFFT) conduit à des résultats proches d'une méthode utilisant également une information de phase (vocodeur de phase). Ainsi, si on suppose que l'estimation par vocodeur de phase est de qualité relativement bonne BETSER et al. (2008), on peut considérer que la QIFFT est un outil adapté à cette tâche sur des signaux réalistes.

4.2.3 Comparaison à l'algorithme de Griffin et Lim

Dans cette expérience, nous testons l'algorithme 2 sur les jeux de données introduits préalablement. La TFCT est calculée comme précédemment. Le tableau 4.2 fournit les résultats de la reconstruction avec l'algorithme GL et avec notre approche. Nous considérons deux cas : les amplitudes peuvent être connues (cas Oracle) ou bien approchées par une KLNMF, qui utilise 30 itérations de règles de mise à jour multiplicatives, et un rang de factorisation égal à 30. Ce scénario non-Oracle nous renseigne sur la dégradation de performance des algorithmes, qui dépendent tous les deux des spectrogrammes d'amplitudes, lorsque ceux-ci ne sont plus égaux à la vérité terrain.

Notre approche donne des résultats significativement meilleurs que l'approche de Griffin et Lim. Les composantes stationnaires et à fréquence variable sont reconstruites avec une meilleure précision dans les deux scénarios. Bien que les deux approches conduisent à une performance moindre lorsque les amplitudes ne sont plus connues exactement, notre approche semble plus prometteuse au niveau du SDR que l'algorithme GL.

Enfin, nous calculons la valeur de l'inconsistance (définie par l'équation (2.4) au chapitre 2) pour les TFCT estimées par ces deux méthodes, dans le cas Oracle (amplitudes connues) sur le jeux de données B. Cette valeur, moyennée sur les données, est de 2×10^2 pour l'algorithme

	Scénario Oracle		Scénario non-Oracle	
	Griffin et Lim	Déroulé de phase	Griffin et Lim	Déroulé de phase
A	0.4	**5.8**	−0.2	**4.7**
B	−0.5	**2.2**	−11.2	**−9.7**
C	−6.5	**0.4**	−8.9	**−4.7**
D	**1.1**	−1.8	−11.8	**−11.6**

TABLEAU 4.2 – Performance de reconstruction (SDR en dB) pour divers jeux de données.

GL contre 1×10^5 pour notre approche. Cela signifie donc qu'un meilleur SDR (ce que l'on interprète comme un signal mieux reconstruit) peut être obtenu au détriment d'une inconsistance plus élevée. Ce résultat confirme les conclusions du chapitre 3 : bien que l'inconsistance puisse être un critère important pour attester de la qualité d'une TFCT complexe estimée, on ne peut pas établir un lien direct entre consistance et critère objectif de reconstruction (comme le SDR). Ainsi, l'optimisation directe de ce critère d'inconsistance n'est pas nécessairement la meilleure approche pour reconstruire la phase d'une TFCT. Il pourrait être intéressant d'envisager des méthodes alternatives pour prendre en compte cette propriété.

Notons enfin que les valeurs de SDR obtenues sont relativement faibles : ainsi, même si la méthode de déroulé conduit à de meilleurs résultats que l'algorithme GL, les signaux restaurés sont corrompus par des artéfacts (que nous identifions plus précisément dans l'expérience suivante) dus à une propagation de l'erreur de phase qui est amplifiée à travers les trames. Il n'est donc pas souhaitable d'utiliser cette méthode dans ce contexte (où il y a un grand nombre de trames à restaurer et pas ou peu d'information disponible). Ainsi, nous considérerons deux applications plus réalistes qui utilisent cette technique : la suppression de craquements (présentée dans la section 4.3) où le nombre de trames successives à restaurer est faible, et la séparation de sources (qui fait l'objet du chapitre 5) où l'on peut exploiter la phase du mélange pour réduire les artéfacts.

4.2.4 Influence des paramètres de TFCT

Longueur de la fenêtre d'analyse

On s'intéresse dans cette expérience à l'influence de la longueur de la fenêtre d'analyse sur la qualité de reconstruction de signal par déroulé linéaire. En effet, cette méthode dépend des fréquences instantanées, et donc de la qualité d'estimation celles-ci. En augmentant la longueur de la fenêtre d'analyse, on augmente également la résolution fréquentielle, et on peut supposer que cela conduira à une meilleure estimation des fréquences instantanées.

On considère des signaux des jeux de données A (morceaux de piano) et C (quatuors à cordes). La TFCT est calculée avec une longueur de fenêtre N_w variable et utilise toujours un taux de recouvrement de 75 % (et pas de bourrage de zéros). Nous présentons sur la figure 4.4 les résultats obtenus.

On constate qu'il existe une grande disparité de SDR selon la longueur de fenêtre utilisée. En particulier, on note la présence d'un pic de SDR pour chaque jeu de données : il semble qu'une valeur optimale de fenêtre existe. En écoutant les résultats obtenus, on identifie deux phénomènes qui caractérisent la dégradation du signal audio :

— Le *bruit musical*. Celui-ci est d'autant plus important que la fenêtre d'analyse est courte. Une fenêtre d'analyse courte implique une résolution fréquentielle faible : l'estimation des fréquences graves est alors peu précise, et le déroulé dans les basses fréquences utilisant une valeur de fréquence instantanée erronnée peut conduire à de tels artéfacts.

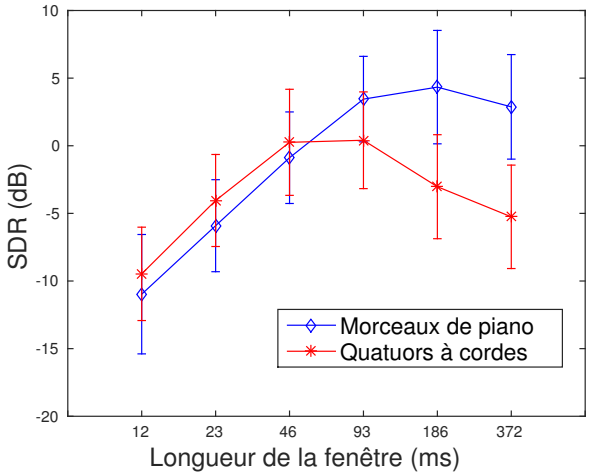


FIGURE 4.4 – Influence de la longueur de fenêtre d'analyse sur la qualité de reconstruction du signal (SDR en dB). Les marqueurs centraux représentent la moyenne et les barres horizontales l'écart-type calculés sur le jeu de données correspondant.

— La perte de précision des transitoires d'attaque. Ce phénomène est aussi connu sous le nom de *phasiness* ou de *reverberation* et a déjà fait l'objet d'études approfondies dans le cadre du vocodeur de phase LAROCHE et DOLSON (1997). Il se manifeste d'autant plus que la longueur de la fenêtre d'analyse est importante : la perte de résolution temporelle conduit à une mauvaise estimation des phases au niveau des attaques.

Il semble donc qu'il faille trouver un compromis entre des longueurs de fenêtre importantes (qui créent de la réverbération) et des fenêtres plus courtes (qui créent du bruit musical) [3]. Le pic de SDR observé expérimentalement pourrait correspondre à ce compromis. Néanmoins, il n'est pas évident que le SDR capture à la fois l'information de bruit musical et de *phasiness*. Perceptivement, certaines valeurs de la fenêtre d'analyse différentes de celle qui conduit au pic de SDR conduisent à des résultats plus satisfaisants et équilibrés au niveau de l'écoute, même s'il s'agit là d'une appréciation subjective de notre part. On constate enfin que pour des quatuors à corde (signaux non-stationnaires en fréquence), le pic de SDR est obtenu pour une fenêtre plus courte que pour les morceaux de piano. Cela peut s'expliquer par le fait qu'en augmentant la longueur de la fenêtre, on perd l'hypothèse de stationarité locale des fréquences, ce qui conduit à dégrader la performance pour des signaux à fréquence variable.

Taux de recouvrement

Dans cette expérience, on évalue l'impact du taux de recouvrement de la TFCT sur la qualité du signal reconstruit. En effet, on suppose qu'en augmentant ce recouvrement, on peut aboutir à une meilleure restitution des phases, notamment au niveau des transitoires d'attaque, tout en conservant une résolution fréquentielle importante. On choisit une fenêtre d'analyse de 4096 échantillons et on fait varier le taux de recouvrement.

Comme on l'a rappelé dans l'annexe A, la condition de reconstruction parfaite est vérifiée, pour des fenêtres de Hann, Hamming et Blackman, pour certaines valeurs du taux de recouvrement. Nous considérons donc les taux de 50 %, 75 % et 87.5 %. La figure 4.5 présente les résultats obtenus sur des morceaux de pianos (données A).

3. Cette notion de compromis entre résolutions temporelle et fréquentielle est centrale en analyse temps-fréquence. Comme on le constate ici, celle-ci n'impacte pas seulement les amplitudes ou les densités spectrales de puissance, mais est également déterminante dans le domaine de la reconstruction de phase.

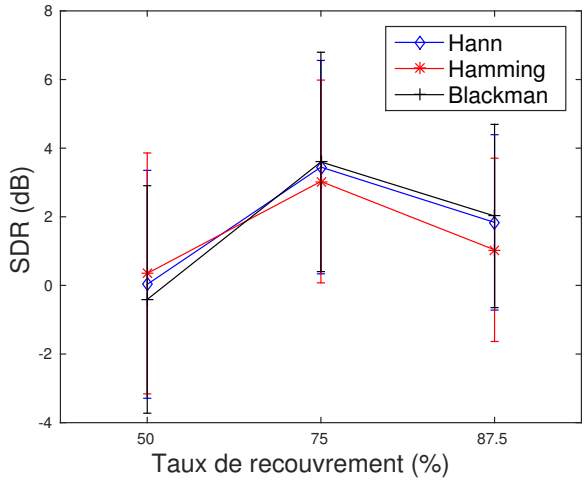


FIGURE 4.5 – Influence du taux de recouvrement de la TFCT sur la qualité de reconstruction du signal (SDR en dB).

Les meilleurs résultats sont obtenus pour un taux de 75 %. Ce taux semble être un bon candidat, puisqu'il conduit à de meilleurs résultats (pour ce jeux de données) qu'un taux de 50 %, pour un temps de calcul moindre qu'un taux de 87.5 % (qui ne conduit en outre pas à améliorer les résultats).

Bourrage de zéros

Nous étudions à présent l'impact du facteur de bourrage de zéros sur la qualité de la restauration de phase. Le facteur de bourrage de zéros est :

$$\tau = \frac{N_{fft}}{N_w}, \tag{4.17}$$

où N_{fft} est le nombre de points utilisés pour calculer la transformée de Fourier, soit ici $2(F-1)$. On s'attend effectivement à ce qu'augmenter la précision fréquentielle permette une meilleure estimation des basses fréquences, et atténue ainsi le bruit musical lorsque la fenêtre d'analyse est courte. La figure 4.6 présente les résultats obtenus sur des morceaux de piano et sur des quatuors à cordes, la TFCT utilisant une fenêtre de Hann et un recouvrement de 75 %.

Pour chaque jeu de données, l'augmentation du facteur de bourrage de zéros améliore la qualité de reconstruction. Cette augmentation n'est toutefois significative que lorsqu'on passe de $\tau = 1$ à $\tau = 2$. Augmenter davantage τ ne produit pas d'amélioration notable de la qualité (aussi bien en matière de SDR que d'un point de vue perceptif). Par ailleurs, cette augmentation est plus marquée lorsque la fenêtre d'analyse est courte (cas $N_w = 512$) que lorsqu'elle est longue (cas $N_w = 4096$). En d'autres termes, augmenter ce paramètre est pratique lorsque la fenêtre d'analyse est courte, ce qui correspond au cas où la résolution fréquentielle est faible, et donc où les fréquences instantanées sont mal estimées (*cf.* expériences précédentes). Néanmoins, son importance s'amoindrit lorsqu'on considère des fenêtres plus longues, puisque la résolution fréquentielle est augmentée : un gain artificiel de précision ne raffine pas l'estimation des fréquences.

On remarque cependant que même avec un facteur τ important, la restitution avec une fenêtre courte n'atteint pas les résultats de celle avec une fenêtre longue sans bourrage de zéros. En fin de compte, cette piste n'est pas satisfaisante pour réduire à la fois le bruit

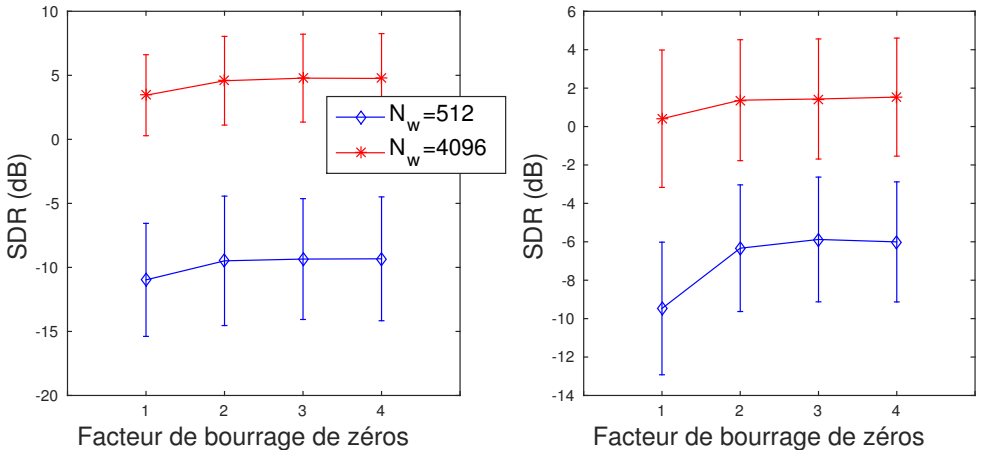


FIGURE 4.6 – Influence du facteur de bourrage de zéros sur la qualité de reconstruction du signal (SDR en dB) pour des morceaux de piano (gauche) et des quatuors à cordes (droite).

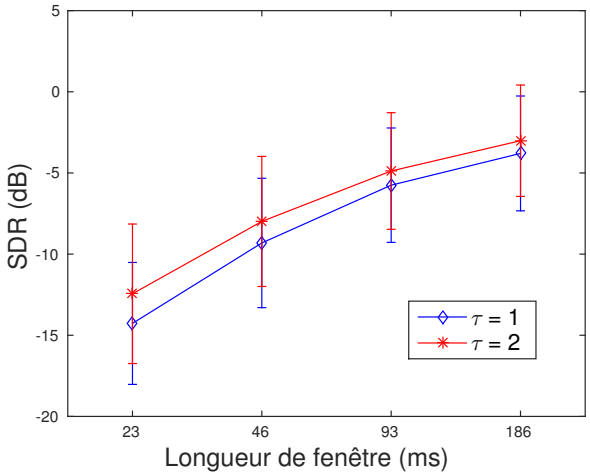


FIGURE 4.7 – Influence de la longueur de fenêtre et du facteur de bourrage de zéros sur la qualité de reconstruction du signal (SDR en dB). Les marqueurs centraux représentent la moyenne et les barres horizontales l'écart-type calculés sur le jeu de données correspondant (jeu de données E).

musical et le phénomène de réverbération. Pour de futures recherches, on pourra considérer une approche multi-résolution pour traiter spécifiquement les transitoires d'attaques, comme c'est notamment proposé dans une version améliorée du vocodeur de phase RÖBEL (2003b,a).

Expérience sur base DSD100

Considérons les morceaux de musique issus de la base DSD100 (jeu de données E). La TFCT est calculée avec une fenêtre de Hann et un recouvrement de 75 %. Nous faisons varier la longueur de la fenêtre d'analyse ainsi que le facteur de bourrage de zéros. Nous étudions donc l'impact de ces paramètres sur la reconstruction, dans le cas de données polyphoniques.

Les résultats présentés sur la figure 4.7 confirment les diagnostics précédemment établis sur des morceaux plus simples (avec piano ou cordes frottées uniquement). Augmenter le facteur de bourrage de zéros améliore les résultats (entre 1 et 2 dB selon la longueur de la fenêtre), et une fenêtre plus longue donne de meilleurs résultats qu'une fenêtre courte, avec en contrepartie l'apparition du phénomène de réverbation qui n'est peut-être pas capturé par

le SDR.

Pour traiter des signaux de musique réalistes, une bonne approche semble être d'effectuer une TFCT avec une fenêtre de longueur comprise entre 46 et 92 ms, et d'éviter le bourrage de zéro, puisque celui-ci n'améliore pas significativement les résultats, mais induit un coût de calcul nettement plus important.

Rappelons cependant que, en théorie, la QIFFT (qui est un des éléments de base de notre méthode) n'est valable que lorsqu'il n'y a qu'une seule sinusoïde par canal fréquentiel. Pour les signaux du jeu de données E, les instruments se recouvrent et cette hypothèse n'est plus vérifiée, ce qui peut expliquer les valeurs assez faibles de SDR obtenues. En séparation de sources (*cf.* chapitre 5) cette méthode de reconstruction sera appliquée à des spectrogrammes de sources isolées, ce qui permet de s'affranchir du problème de recouvrement : même si les sources se recouvrent dans le domaine TF, on supposera que pour chaque source isolée, il y a au plus une sinusoïde active par canal fréquentiel et par source.

4.3 Application à la suppression de clics

Nous proposons de tester la méthode de déroulé linéaire de phase dans le cadre de la restauration de signaux audio corrompus par des clics Charbit et Cappé (1997). Les clics sont des bruits de courte durée (de l'ordre de quelques échantillons), souvent observés dans de vieux enregistrements (bandes magnétiques ou disques vinyles déteriorés) et se traduisent perceptivement par des craquements. La restauration d'enregistrements est un thème de recherche vivant en traitement du signal audio, aussi nous avons souhaité examiner le potentiel d'un algorithme de reconstruction de phase pour une telle application. Nous supposons ici que l'information de phase dans certaines trames (correspondant aux clics) est perdue, ce qui signifie que l'on ne peut pas exploiter d'information supplémentaire (contrairement à la séparation de sources, où la phase du mélange est disponible).

4.3.1 Méthodes de restauration audio

Synthèse des craquements

Nous considérons des signaux audio non corrompus et les déteriorons synthétiquement par des clics. Ce protocole permet de comparer les sons originaux et restaurés. Pour fabriquer les clics, nous avons dérivé des fenêtres de Hann d'une durée d'environ 1 ms, que nous avons ajoutées au signal original comme montré sur la figure 4.8. Pour une application réaliste, ceux-ci représentent au total moins de 1 % de la durée du signal original.

Détection

Dans les approches telles que Janssen et al. (1986), les clics sont détectés par le biais d'une modélisation autorégressive (AR) du signal dans le domaine temporel. Ce modèle est également utilisé pour la restauration. Les écarts entre les données et le modèle AR permettent d'identifier la présence de craquements Esquef et al. (2002).

Dans le domaine TF, comme c'est suggéré dans Kahrs et Brandenburg (1998), nous détectons les clics en étudiant l'énergie des signaux dans les hautes fréquences. En effet, les clics étant des signaux quasi-impulsifs, ils sont localisés en temps et étalés en fréquence (ce que montre la figure 4.9). Ainsi, le calcul des énergies spectrales dans les hautes fréquences (dans lesquelles le signal original n'a que peu d'énergie) permet de localiser les clics.

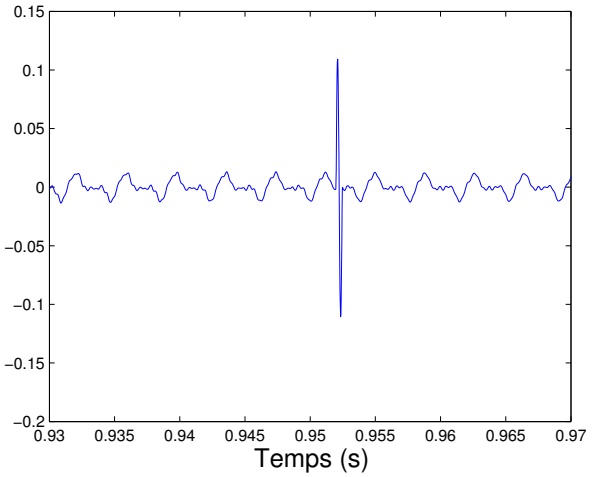


FIGURE 4.8 – Exemple d'un signal de piano en présence d'un craquement.

Méthode temporelle

La méthode de débruitage JANSSEN et al. (1986); GODSILL et RAYNER (1998) est une méthode temporelle qui consiste en une modélisation AR du signal non bruité. Les craquements sont modélisés comme des impulsions apparaissant à des instants aléatoires. L'idée de la méthode est d'abord d'estimer les positions des craquements en mesurant l'écart entre le modèle AR et le signal observé, puis de restaurer à ces endroits le signal par application du filtrage AR dont les coefficients ont été préalablement estimés (par exemple via les équations de Yule-Walker).

Cette méthode est simple à mettre en oeuvre et rapide, mais elle ne conduit à de bons résultats que lorsque l'ordre du filtre AR est relativement faible (des artéfacts peuvent apparaître lorsqu'il y a de nombreux instruments dans le signal). En outre, elle requiert un réglage fin du seuil de détection ainsi que de l'ordre du filtre AR.

Méthodes Temps-Fréquence

On peut restaurer le signal dans le domaine TF ADLER et al. (2012). Il est envisageable de restaurer directement la TFCT corrompue (à valeurs complexes) grâce par exemple au modèle HRNMF : les paramètres du modèle sont appris sur la partie de la TFCT non corrompue, puis, par application du modèle appris, on peut restaurer toute la TFCT BADEAU (2011). L'estimation du modèle HRNMF est cependant coûteuse en temps de calcul (*cf.* chapitre 3).

L'approche que nous proposons consiste à procéder en deux temps. Tout d'abord, on restaure le spectrogramme. Nous proposons d'utiliser une interpolation linéaire sur le logarithme de l'amplitude pour restaurer les amplitudes des points TF manquants (hypothèse d'amplitudes exponentiellement décroissantes BADEAU (2012)). Ce procédé est illustré par la figure 4.9.

Dans un deuxième temps, nous restaurons la phase de ces points corrompus par différentes méthodes :

— L'algorithme GL, en supposant connues (et donc fixes) les phases des points TF où le signal n'est pas corrompu ;

— L'algorithme de déroulé linéaire. Plusieurs stratégies sont alors possibles :

  — Le déroulé peut être fait dans le sens des temps croissants ("déroulé avant") ;

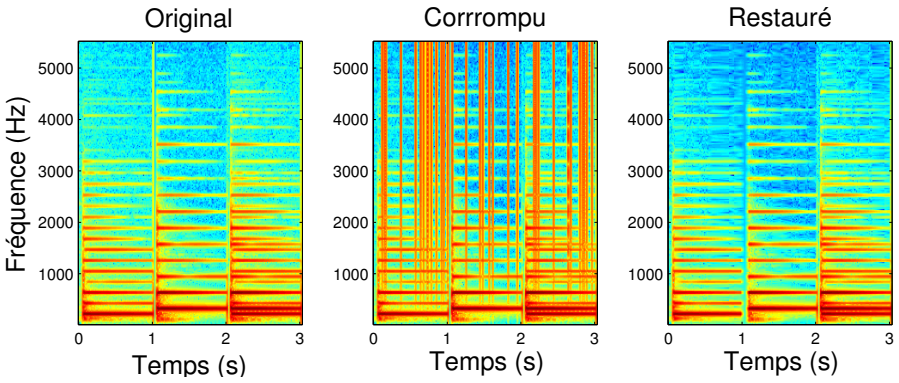


FIGURE 4.9 – Restauration de spectrogrammes par interpolation linéaire de log-amplitude sur un mélange de notes de piano : original (gauche), corrompu par des clics (centre) et restauré (droite).

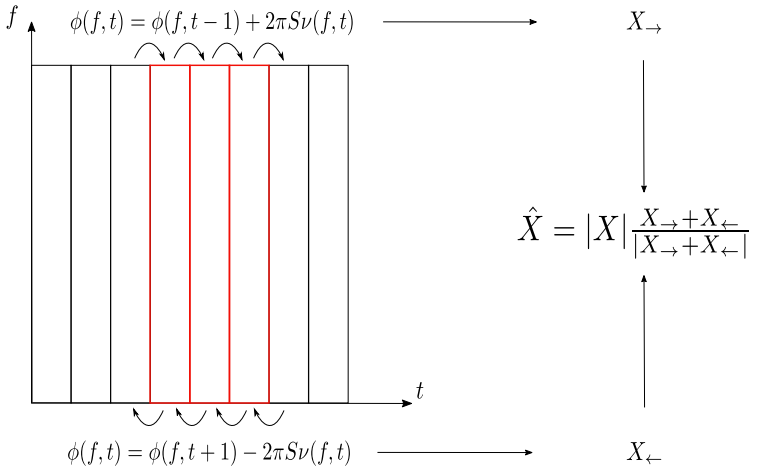


FIGURE 4.10 – Méthode de déroulé de phase pour reconstruire des trames temporelles corrompues : on combine un déroulé avant et un déroulé arrière que l'on moyenne ensuite.

— Similairement, on peut effectuer un "déroulé arrière" dans le sens des temps décroissants ;

— Ces deux approches conduisent à une discontinuité de phase au niveau de la limite entre la zone corrompue et celle qui ne l'est pas. Pour la réduire, on peut moyenner les deux résultats. On parlera alors de "déroulé moyen". Cette méthode est illustrée sur la figure 4.10.

Remarque : On aurait pu directement moyenner les phases avant et arrière dans cette dernière approche. Néanmoins, en raison de la 2π-périodicité de celle-ci, si on avait une estimée légèrement supérieure à 0 et une autre légèrement inférieure à 2π, la moyenne donnerait une phase autour de π, alors que la phase souhaitée est proche de 0. Pour éviter ce problème, on choisit donc plutôt de moyenner les composantes complexes $X_{\rightarrow}$ et $X_{\leftarrow}$.

4.3.2 Résultats expérimentaux

La première expérience utilise les données A à D qui sont majoritairement monophoniques, alors que la deuxième expérience sera consacrée au jeu de données E constitué de morceaux de musique polyphonique (ce sont les jeux de données présentés dans la section 4.2.1). Les signaux sont échantillonnés à 44100 Hz. La TFCT est calculée avec une fenêtre de Hann de longueur 512 échantillons et un taux de recouvrement de 50 %. En effet, même si les craquements

Données	Modèle AR	HRNMF	Griffin et Lim	Déroulé linéaire		
				Avant	Arrière	Moyen
A	**26.4**	19.1	16.0	17.5	17.2	18.4
B	24.1	18.4	15.5	22.1	22.1	**25.1**
C	**25.0**	18.3	15.3	17.9	17.9	19.1
D	**22.6**	20.6	19.2	19.3	19.4	20.2

Tableau 4.3 – Suppression de clics : performance (SDR en dB) sur plusieurs jeux de données.

sont de durée relativement courte, ils corrompent toute une trame d'analyse dans le plan TF. Des fenêtres courtes et un recouvrement réduit limitent la proportion de trames corrompues. L'algorithme GL utilise 50 itérations et la qualité de la restauration est mesurée par le SDR (en dB).

Signaux monophoniques

Les résultats sur des signaux monophoniques (données A à D) sont présentés dans le tableau 4.3.

La méthode temporelle AR fournit globalement les meilleurs résultats, excepté pour le jeu de données B. La méthode AR est à priori très bien adaptée à ce type de données ce qui explique cette bonne performance.

La méthode HRNMF donne également de bons résultats, légèrement inférieurs à ceux de la méthode AR, et comparables à ceux de la technique proposée. Alors que le modèle HRNMF est appris sur les trames de la TFCT non corrompues par les clics, notre méthode est aveugle. Il serait donc intéressant d'incorporer la connaissance sur les phases de toute la partie non corrompue (et pas seulement des trames directement adjacentes à la zone corrompue) à notre méthode pour raffiner la restauration.

L'algorithme GL donne des résultats en deça de notre technique. Cela signifie qu'à restauration d'amplitude égale (la même technique est utilisée), la reconstruction de phase par déroulé linéaire fournit de meilleurs résultats que l'algorithme GL. Enfin, on constate que le fait de combiner un déroulé avant et arrière pour les moyenner améliore les performances par rapport à un simple déroulé avant. Cela se remarque au niveau du SDR, mais également perceptivement : les artéfacts dûs aux discontinuités s'en voient réduits.

Morceaux de musiques polyphoniques

Nous présentons les résultats obtenus pour le jeu de données E sur la figure 4.11, en omettant volontairement les méthodes de déroulé avant et arrière, pour ne laisser que le résultat du déroulé moyen.

Notre méthode fournit des résultats nettement supérieurs à la méthode traditionnelle (AR), et légèrement inférieurs à la technique basée sur le modèle HRNMF (environ -0.9 dB). Cet exemple montre la limite de la méthode AR dans le cas où il y a de nombreux instruments qui se recouvrent dans le domaine temporel : le filtre AR à estimer a alors un ordre très élevé, et la technique traditionnelle n'est plus capable de fournir d'aussi bons résultats que précédemment.

Notre méthode donne des résultats similaires à HRNMF, tout en étant plus rapide et aveugle. Par ailleurs, la technique de déroulé ne nécessite pas le réglage des paramètres des méthodes AR et HRNMF (nombre de sources, nombre d'itérations, ordre du filtre AR, paramètre de seuil etc.). Enfin, notre méthode est basée sur la reconstruction de phase, et il est possible que sa performance soit limitée par la reconstruction d'amplitude.

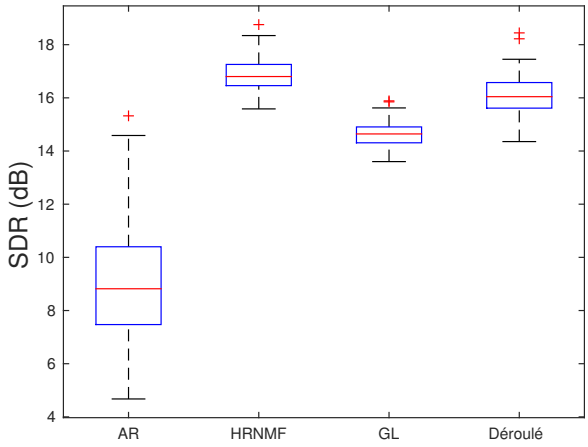


FIGURE 4.11 – Performance de la suppression de clics (SDR en dB) sur le jeu de données E pour plusieurs méthodes.

Pour une comparaison équitable en matière de reconstruction de phase (qui agit donc uniquement sur la phase de la TFCT), il est logique de comparer notre méthode à l'algorithme GL puisque ces deux approches utilisent la même technique préalable de reconstruction d'amplitude. Notre méthode fournit de meilleurs résultats que l'algorithme GL (+1.5 dB) et elle est rapide en temps de calcul d'un facteur 2. Cependant, ce dernier point est à relativiser car des implémentations rapides (temps réel) de l'algorithme GL existent ZHU et al. (2007).

4.4 Vers un modèle d'attaques

Nous avons jusqu'à présent supposé les phases dans les trames d'attaque connues. Pour des applications réalistes, il est nécessaire de reconstruire celles-ci. Nous présentons dans cette section un modèle pour la reconstruction des phases d'attaque. Il est important d'estimer avec précision ces phases pour plusieurs raisons :

— L'attaque est primordiale au niveau perceptif car elle contribue au timbre et à la qualité audio IVERSON et KRUMHANSL (1993).

— L'algorithme 2 de déroulé de phase repose sur une relation récursive et a besoin d'un point de départ, fourni au niveau de l'attaque.

— La cohérence de phase au niveau de l'attaque est ensuite préservée durant le déroulé, ce qui garantit la cohérence entre les différents partiels qui composent la source.

Le problème de la reconstruction des phases d'attaque peut être lié à la problématique de la cohérence verticale des phases. Cette propriété a été étudiée en acoustique musicale GALEMBO et al. (2001); CHAIGNE et KERGOMARD (2008) ainsi qu'en synthèse de signaux étirés temporellement, comme dans l'algorithme du vocoder de phase. Dans LAROCHE et DOLSON (1997) et LAROCHE et DOLSON (1999), les auteurs présentent une technique (dite *phase locking*) qui permet de conserver la cohérence verticale entre partiels lors de la reconstruction de phase dans l'algorithme du vocoder. Ces travaux reposent sur une première approche PUCKETTE (1995) dont le but était similaire. Néanmoins, les auteurs de ces études ne s'appuient que sur une évaluation perceptive pour juger de la réduction de *phasiness* et mentionnent la difficulté de trouver un indicateur adapté pour quantifier ce phénomène.

Dans le chapitre 6, nous proposerons un modèle permettant d'estimer les phases dans les trames d'attaque utilisant la propriété de redondance des signaux audio. Dans cette section, nous proposons une technique de reconstruction des phases d'attaque basée sur un modèle d'impulsion, ce qui conduit à une méthode très similaire à celle présentée dans ce chapitre.

4.4.1 Modèle d'impulsion

Phase d'une impulsion

Nous modélisons les transitoires d'attaque dans le domaine TF par des impulsions. Bien que de tels signaux ne modélisent pas parfaitement les attaques, il fournissent des équations de déroulé vertical (à travers les fréquences) Sugiyama et Miyahara (2013b) qui pourront ensuite être affinées. Une impulsion d'amplitude $A > 0$ centrée en un temps d'attaque n_0 est, $\forall n \in \mathbb{Z}$:

$$x(n) = A\delta_{n-n_0}, \tag{4.18}$$

où δ vaut 1 si $n = n_0$, et 0 sinon. Sa TFCT est nulle excepté au sein des trames d'attaque qui contiennent n_0 :

$$X(f,t) = Aw(n_0 - St)e^{-2i\pi\frac{f}{F}(n_0-St)}. \tag{4.19}$$

Nous pouvons alors obtenir une relation entre la phase de canaux fréquentiels successifs au sein d'une trame d'attaque :

$$\phi(f,t) = \phi(f-1,t) - \frac{2\pi}{F}(n_0 - St). \tag{4.20}$$

Cette relation est similaire à l'équation de déroulé horizontal (4.5). Cette similarité était prévisible car l'impulsion est le dual de la sinusoïde dans le plan TF. En pratique, les signaux dans les trames d'attaque ne suivent pas exactement ce modèle d'impulsion, aussi on estimera le temps d'attaque n_0 dans chaque canal fréquentiel : on reprend en effet l'analogie entre le modèle sinusoïdal (fréquence dépendant du temps) et le modèle d'impulsion (temps d'attaque dépendant du temps).

Estimation du temps d'attaque

Observons l'amplitude de la TFCT dans le canal fréquentiel f :

$$|X(f,t)| = Aw(n_0 - St). \tag{4.21}$$

À un facteur d'amplitude près, l'amplitude de la TFCT à la fréquence f le long des trames d'attaque est une version sous-échantillonnée de la fenêtre d'analyse, décalée de n_0. Nous pouvons alors estimer par moindres carrés la valeur de ce paramètre. Alternativement, on peut s'inspirer du modèle sinusoïdal, et estimer ce paramètre n_0 par une technique similaire à la QIFFT : on effectue une interpolation parabolique autour du maximum d'amplitude correspondant à la trame d'attaque.

Un exemple

Nous testons cette méthode sur un signal composé de deux impulsions d'amplitudes différentes. Les résultats de reconstruction de phase sont présentés sur la figure 4.12.

La phase est reconstruite avec une grande précision dans le domaine TF, indifféremment des amplitudes et des temps d'attaque des impulsions. Nous observons une reconstruction parfaite (plus de 270 dB de SDR). Comparativement, 100 itérations de l'algorithme GL donnent une moyenne de -3.6 dB de SDR sur 30 initialisations aléatoires.

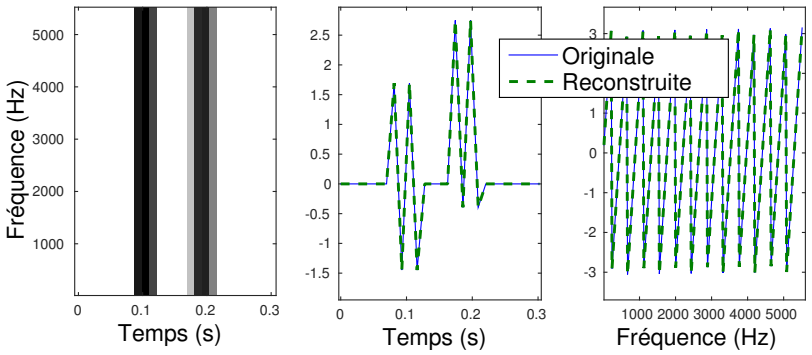


FIGURE 4.12 – Mélange d'impulsions : spectrogramme (gauche) et reconstruction de phase par déroulé linéaire en fonction du temps dans la bande de fréquences à 1800 Hz (centre) et en fonction des fréquences dans la première trame d'attaque (droite).

4.4.2 Validation expérimentale

Instruments à hauteur définie

Dans cette expérience, nous proposons de reconstruire les phases dans les trames d'attaque par différentes méthodes. Nous testons l'équation de déroulé qui provient du modèle d'impulsion (4.20), dont le paramètre de temps d'attaque n_0 peut être estimé par moindre carrés (**LS**), ou par interpolation parabolique (**QI**). Nous testons également des phases aléatoires (**Rand**, pas de cohérence verticale), des phases nulles (**Null**, partiels en phase) et des phases de partiels alternées entre 0 et π (**Alt**, partiels en opposition de phase). Ces choix sont justifiés par l'observation des relations de phases entre partiels de piano en acoustique musicale GALEMBO et al. (2001); CHAIGNE et KERGOMARD (2008). Nous testons enfin des phases d'attaque Oracle (supposées connues). Ces phases d'attaque sont fournies à l'algorithme 2, qui achève la reconstruction des phases par déroulé horizontal. On teste enfin 200 itérations de l'algorithme de Griffin et Lim (**GL**). Les amplitudes sont supposées connues.

Les signaux sont composés de deux notes de piano ou de guitare (*cf.* chapitre 3). Les résultats présentés dans le tableau 4.4 montrent que toutes ces approches fournissent de meilleurs résultats que l'algorithme GL sur ces signaux. L'estimation de phases d'attaque utilisant le modèle d'impulsion (**LS** et **QI**) conduit aux meilleurs résultats. En particulier, nous observons perceptivement que ces approches conduisent à une attaque nette et percussive, alors que des phases aléatoires conduisent à des attaques floues et mal définies.

Partiels	Déroulé linéaire						GL
Attaques	Oracle	LS	QI	Rand	Null	Alt	GL
Notes de piano	6.18	1.56	**3.28**	0.84	0.82	0.87	−0.58
Notes de guitare	3.89	**2.96**	2.50	2.62	2.64	2.64	−4.61

TABLEAU 4.4 – Performance de reconstruction (SDR en dB) de différentes méthodes de reconstruction de phase.

Sons percussifs

Nous testons enfin ce modèle d'impulsion sur trois signaux percussifs (grosse caisse, caisse claire et cymbale Charleston en position fermée), dont les spectrogrammes sont illustrés sur la figure 4.13.

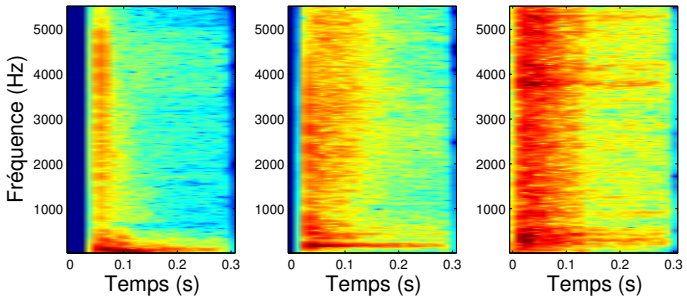


FIGURE 4.13 – Spectrogrammes de signaux percussifs : grosse caisse (gauche), caisse claire (centre) et cymbale Charleston fermée (droite).

Partiels	Griffin	Déroulé linéaire			
Attaques	Lim	LS	QI	Rand	Null
Grosse caisse	**18.8**	11.6	15.5	0.7	9.7
Caisse claire	**11.5**	8.6	1.4	3.5	4.2
Cymbale (fermée)	**9.0**	1.8	0.6	0.3	0.8

TABLEAU 4.5 – Performance de reconstruction de signaux percussifs (SDR in dB).

Les phases d'attaque sont reconstruites avec les mêmes méthodes que présentées dans le paragraphe précédent (à l'exception de l'alternance entre 0 et π qui ne fait pas sens ici). Le déroulé horizontal est ensuite appliqué pour restaurer le reste des phases. Les résultats sont présentés dans le tableau 4.5.

L'algorithme GL donne de meilleurs résultats que notre approche. Parmi les approches utilisant le déroulé linéaire, les signaux de grosse caisse et de caisse claire sont mieux reconstruits lorsque les phases d'attaque sont calculées par le modèle d'impulsion. L'estimation de la phase de la cymbale n'est pas satisfaisante, probablement car ce type de sons comporte du bruit et est assez mal modélisé par des mélanges de sinusoïdes et d'impulsions.

Il semble donc qu'un travail plus approfondi de modélisation des signaux percussifs soit nécessaire pour aboutir à une méthode de reconstruction de phase adaptée à ces signaux.

4.5 Conclusion

La technique de reconstruction de phase introduite dans ce chapitre apparaît comme un outil efficace et prometteur pour cette tâche, comparativement à la méthode de Griffin et Lim basée sur la consistance de la TFCT. Les expériences ont montré le potentiel de cette méthode, notamment dans le cadre de la restauration audio, où de meilleurs résultats qu'avec la méthode temporelle ont été obtenus dans le cas de musiques polyphoniques. L'étude de l'influence des paramètres de la TFCT a mis en évidence deux phénomènes perceptifs qui surviennent lors de l'application de cette méthode : la réverbération due à une résolution temporelle trop faible, et le bruit musical, dû à une résolution fréquentielle trop faible. Un compromis peut être obtenu en choisissant convenablement les paramètres de la TFCT.

Nous proposons, dans le chapitre suivant, de nous intéresser à l'intégration de cette contrainte de phase dans des modèles de mélanges pour la séparation de sources.

Enfin, le modèle d'impulsion qui a été introduit à la fin de ce chapitre n'est pas très satisfaisant pour traiter des signaux complexes, mais peut servir de point de départ à l'élaboration de modèles plus sophistiqués comme des mélanges d'impulsions au sein d'une même trame d'attaque.

Chapitre 5

Déroulé de phase et séparation de sources

Sommaire

Dans ce chapitre, nous intégrons la méthode de déroulé de phase dans le cadre de la séparation de sources. Nous considérons le problème de séparation de sources comme un problème d'optimisation, qui consiste à minimiser une fonction de coût traduisant l'écart entre les mélanges observé et estimé, sous contrainte que le module des sources estimées est fixé à une valeur objectif. Notre idée consiste alors à incorporer dans ce contexte une information à priori sur la phase, calculée par la méthode de déroulé linéaire.

Les aspects techniques de cette procédure itérative ont fait l'objet d'un rapport technique disponible en ligne Magron et al. (2016a). Un article reprenant certaines expériences du chapitre précédent, ainsi que celles présentées dans ce chapitre a été soumis à publication dans la revue *IEEE Transactions on Audio, Speech and Language Processing* Magron et al. (2017b).

Dans la section 5.1, nous proposons une formulation de ce problème de séparation de sources. Dans la section 5.2, nous obtenons donc un algorithme de séparation de sources. Nous évaluons expérimentalement cette technique dans la section 5.3. Nous proposons également une procédure contrainte alternative dans la section 5.4. Enfin, la section 5.5 présente quelques remarques conclusives.

5.1 Position du problème

On raisonne dans cette section (et dans la suivante) dans un point TF (f,t), et on s'affranchit des indices pour plus de clarté. Nous cherchons à obtenir un estimateur $\hat{X}_k$ des K composantes complexes X_k à partir de leur somme X, en supposant que l'on a une estimation de leurs modules V_1,...,V_K (qui peut être obtenue via une NMF préalable, par exemple). Ce problème peut être résolu en minimisant la fonction de coût :

$$|E| = |X - \sum_k \hat{X}_k|. \tag{5.1}$$

Cette fonction possède en général plusieurs zéros. Par exemple, si $K = 2$, considérons deux complexes $X_1 = V_1 e^{i\theta_1}$ et $X_2 = V_2 e^{i\theta_2}$ dont la somme vaut exactement $X = V e^{i\theta}$. Leurs symétriques par rapport à X sont $\bar{X}_1 e^{2\theta}$ et $\bar{X}_2 e^{2\theta}$, où $\bar{z}$ désigne le complexe conjugué de z. Leur somme est :

$$(\bar{X}_1 + \bar{X}_2)e^{2\theta} = \bar{X}e^{2\theta} = V e^{-\theta} e^{2\theta} = X. \tag{5.2}$$

Ainsi, à partir d'un couple de solutions, on peut en identifier aisément un autre, ce qui motive la recherche d'un "bon" minimum. En outre, lorsque $K \geq 3$, il y a une infinité de solutions à ce problème. La stratégie la plus répandue dans la littérature pour obtenir (de façon non-itérative) un zéro de $|E|$ consiste à appliquer un filtrage de type Wiener Liutkus et Badeau (2015). Les estimées sont :

$$\hat{X}_k = \frac{V_k^2}{\sum_l V_l^2} \odot X, \tag{5.3}$$

et l'erreur vaut alors :

$$|E| = |X - \sum_k \hat{X}_k| = |X| \left| 1 - \sum_k \frac{V_k^2}{\sum_l V_l^2} \right| = |X||1 - 1| = 0. \tag{5.4}$$

Cependant, les estimées par filtrage de Wiener ne sont pas des solutions du problème car $|\hat{X}_k| \neq V_k$, et ne conduisent pas toujours à des résultats satisfaisants du point de vue de

la qualité audio lorsque les sources se recouvrent dans le domaine TF (*cf.* chapitre 3). Nous proposons donc d'obtenir de nouvelles estimées des sources par une procédure itérative visant à minimiser $|E|$. Une initialisation par déroulé linéaire confère ainsi à l'estimée obtenue une propriété de régularité temporelle.

5.2 Procédure itérative d'estimation des composantes

Dans cette section, nous présentons l'algorithme que nous avons mis au point pour estimer les composantes complexes à partir de leur mélange.

5.2.1 Exemple dans le cas de deux nombres complexes

Afin d'estimer les phases de chaque source dans le mélange, illustrons notre technique avec le cas particulier de deux nombres complexes X_1 et X_2. Leurs modules V_1 et V_2 sont fixés. La somme X des deux est connue, et on cherche donc à estimer leur argument, ou, formulé différemment, on cherche ces deux complexes avec une contrainte sur leurs modules.

La procédure que nous utilisons est inspirée de Gunawan et Sen (2010), dont nous rappelons le principe. À l'itération (it), on a une estimation de chacun des complexes, $\hat{X}_1^{(it)}$ et $\hat{X}_2^{(it)}$. On peut donc calculer l'erreur entre la somme des deux estimées et le mélange X, soit $E^{(it)} = X - \hat{X}_1^{(it)} - \hat{X}_2^{(it)}$, erreur que l'on redistribue sur les composantes. On normalise ensuite le module de chaque estimée à la valeur objectif, puis on réitère jusqu'à convergence. Ainsi, on effectue à l'itération (it) les opérations suivantes :

1. Distribution de l'erreur : $Y_1^{(it)} = \hat{X}_1^{(it-1)} + \dfrac{E^{(it-1)}}{2}$ et $Y_2^{(it)} = \hat{X}_2^{(it-1)} + \dfrac{E^{(it-1)}}{2}$;
2. Normalisation à la valeur objectif : $\hat{X}_1^{(it)} = \dfrac{Y_1^{(it)}}{|Y_1^{(it)}|} V_1$ et $\hat{X}_2^{(it)} = \dfrac{Y_2^{(it)}}{|Y_2^{(it)}|} V_2$;
3. Calcul de l'erreur : $E^{(it)} = X - \hat{X}_1^{(it)} - \hat{X}_2^{(it)}$;
4. Retour à l'étape 1 jusqu'à convergence.

Cette méthode itérative est illustrée par la figure 5.1. Notons qu'une approche similaire a été utilisée dans Mowlaee et al. (2012) pour l'estimation des phases dans un mélange de deux sources uniquement.

5.2.2 Procédure itérative dans le cas général

Nous étendons cette méthode au cas de la somme de K sources composant un mélange. On a :

$$Y_k^{(it+1)} = \hat{X}_k^{(it)} + \lambda_k E^{(it)}, \tag{5.5}$$

ce qui permet d'ajuster la distribution de l'erreur sur la composante k via un poids positif λ_k. Nous souhaitons par ailleurs que les Y_k somment à X, ce qui revient à imposer la condition suivante :

$$\sum_k \lambda_k = 1. \tag{5.6}$$

Par ailleurs, il est souhaitable que le k-ième poids soit d'autant plus grand que la composante correspondante possède d'énergie : une composante de grande énergie est en effet majoritairement responsable de l'erreur de reconstruction. Ces deux conditions nous conduisent à

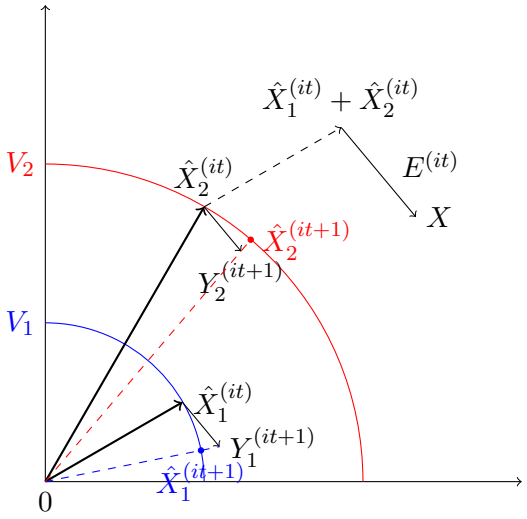


FIGURE 5.1 – Illustration de la procédure itérative consistant à estimer deux nombres complexes dont le module et la somme sont connus.

proposer la définition suivante pour les poids :

$$\lambda_k = \frac{V_k^2}{\sum_l V_l^2}, \tag{5.7}$$

qui correspond au gain de Wiener. D'autres définitions sont possibles, tant qu'elles respectent (5.6). Nous choisissons (5.7) car elle a donné de bons résultats expérimentaux. L'Algorithme 3 détaille cette procédure (dans le cas général où l'on peut utiliser n'importe quelle définition pour les poids λ_k).

5.2.3 Décroissance de la fonction de coût

Nous prouvons ci-après que cette procédure fait décroître la fonction de coût au cours des itérations. On a :

$$\begin{aligned}
\left|E^{(it+1)}\right| &= \left|X - \sum_k \hat{X}_k^{(it+1)}\right| \\
&= \left|\sum_k Y_k^{(it+1)} - \hat{X}_k^{(it+1)}\right| \\
&= \left|\sum_k \hat{X}_k^{(it)} + \lambda_k E^{(it)} - \frac{\hat{X}_k^{(it)} + \lambda_k E^{(it)}}{|\hat{X}_k^{(it)} + \lambda_k E^{(it)}|} V_k\right| \\
&= \left|\sum_k (\hat{X}_k^{(it)} + \lambda_k E^{(it)}) \left(1 - \frac{V_k}{|\hat{X}_k^{(it)} + \lambda_k E^{(it)}|}\right)\right|.
\end{aligned}$$

Algorithme 3 Estimation des sources complexes à partir de leur mélange.

Entrées :

Mélange $X \in \mathbb{C}$,

Modules $V_k \in \mathbb{R}_+$, poids λ_k et valeurs initiales $\hat{X}_k \in \mathbb{C}$, $\forall k \in [\![1, K]\!]$,

Nombre d'itérations N_{it}.

Erreur initiale : $E = X - \sum_k \hat{X}_k$.

pour $it = 1$ à N_{it} **faire**

  pour $k = 1$ à K **faire**

    $Y_k \leftarrow \hat{X}_k + \lambda_k E$,

    $\hat{X}_k \leftarrow \dfrac{Y_k}{|Y_k|} V_k$.

  fin pour

  $E \leftarrow X - \sum_k \hat{X}_k$.

fin pour

Sorties :

$\forall k \in [\![1, K]\!]$, $\hat{X}_k \in \mathbb{C}$.

En appliquant l'inégalité triangulaire, il vient :

$$\begin{aligned}\left|E^{(it+1)}\right| &\leq \sum_k |\hat{X}_k^{(it)} + \lambda_k E^{(it)}| \left|1 - \frac{V_k}{|\hat{X}_k^{(it)} + \lambda_k E^{(it)}|}\right| \\ &\leq \sum_k \left||\hat{X}_k^{(it)} + \lambda_k E^{(it)}| - V_k\right| \\ &\leq \sum_k \left||\hat{X}_k^{(it)} + \lambda_k E^{(it)}| - |\hat{X}_k^{(it)}|\right|.\end{aligned}$$

On utilise alors une propriété des nombres complexes (aisément démontrable en utilisant l'inégalité triangulaire) : $\forall (a, b) \in \mathbb{C}^2$,

$$||a| - |b|| \leq |a - b|. \tag{5.8}$$

En prenant $a = \hat{X}_k^{(it)} + \lambda_k E^{(it)}$ et $b = \hat{X}_k^{(it)}$, on obtient l'inégalité suivante :

$$\left||\hat{X}_k^{(it)} + \lambda_k E^{(it)}| - |\hat{X}_k^{(it)}|\right| \leq \left|\lambda_k E^{(it)}\right|. \tag{5.9}$$

En injectant cette inégalité dans le calcul précédent, on obtient :

$$\left|E^{(it+1)}\right| \leq \sum_k \left|\lambda_k E^{(it)}\right| \leq \left|E^{(it)}\right| \sum_k \lambda_k. \tag{5.10}$$

Or, les λ_k somment à l'unité, comme on l'a vu en (5.6), donc :

$$\left|E^{(it+1)}\right| \leq \left|E^{(it)}\right|, \tag{5.11}$$

ce qui montre la décroissance de la fonction de coût au cours des itérations (et donc sa convergence, vu qu'il s'agit d'une suite à termes positifs).

Remarque : Il est possible d'obtenir l'algorithme 3 en utilisant la méthode de la fonction auxiliaire, similairement à Kameoka et al. (2009). Cette méthode permet d'aboutir naturellement à la procédure itérative, avec une garantie de convergence. Nous avons préféré introduire cette procédure ici de façon intuitive, afin de ne pas surcharger ce chapitre de détails mathématiques. Ceux-ci sont fournis dans l'Annexe B de ce manuscrit.

5.2.4 Initialisation de l'algorithme

D'après les remarques de la section 5.1, le choix de l'initialisation de l'algorithme conditionne fortement la qualité des estimations ainsi que la vitesse de convergence.

Une idée consiste à initialiser les composantes en leur donnant la phase du mélange. Nous montrons dans l'annexe B que cela est un point fixe de la procédure : après une ou deux itérations (selon le choix des poids λ_k), les composantes ne sont plus modifiées, et sont alors égales aux composantes initiales à un déphasage de π près, ce qui n'est pas le but recherché.

Cela nous incite à initialiser l'algorithme avec notre technique de déroulé de phase plutôt qu'en donnant la phase du mélange aux composantes : on s'attend à ce que celle-ci soit assez proche d'un minimum local, et permette non seulement une convergence rapide de l'algorithme, mais également l'obtention d'une solution qui bénéficie de la propriété de régularité temporelle issue du modèle sinusoïdal.

5.2.5 Protocole de séparation de sources

Notre approche repose sur l'utilisation du déroulé linéaire de phase introduit au chapitre précédent. Le déroulé de phase étant récursif, il est initialisé aux trames d'attaque (*cf.* chapitre 4 section 4.1.5).

Nous devons donc faire certaines hypothèses sur les phases des sources dans ces trames. Tout d'abord, nous supposons connues leurs positions. En pratique, on utilise la boîte à outils MATLAB Tempogram Grosche et Müller (2011) sur chaque spectrogramme de source pour les déterminer. On obtient alors la fonction indicatrice des trames d'attaque :

$$\mathbb{1}_k(t) = \begin{cases} 1 & \text{si } t \in \Omega_k \\ 0 & \text{sinon,} \end{cases} \tag{5.12}$$

où Ω_k désigne l'ensemble des trames d'attaque de la source k. La fonction indicatrice du complémentaire est donc $\bar{\mathbb{1}}_k = 1 - \mathbb{1}_k$. Dans les expériences, les phases dans les trames d'attaque seront supposées connues, à l'exception de l'expérience décrite dans la section 5.3.4 où elles seront égales à la phase du mélange, pour une comparaison équitable avec les autres méthodes. Nous effectuerons néanmoins dans la section 5.3.5 une expérience visant à comparer ces deux cas de figure afin de déterminer le potentiel d'amélioration d'estimation des phases d'attaque.

Nous présentons dans l'Algorithme 4 la procédure complète de séparation de sources que nous proposons. Notons que dans cette procédure, l'optimisation est effectuée trame par trame, avant de procéder à l'initialisation de la trame suivante par déroulé linéaire. Nous verrons dans l'expérience 5.3.2 l'intérêt de cette approche par rapport à une application de la procédure d'optimisation directement sur toute la TFCT.

Algorithme 4 Estimation de sources complexes dans un mélange à partir de leurs spectrogrammes en utilisant le déroulé linéaire de phase

Entrées :

Mélange $X \in \mathbb{C}^{F \times T}$,

Spectrogrammes $V_k \in \mathbb{R}_+^{F \times T}$, $\forall k \in [\![1, K]\!]$,

Indicatrices des trames d'attaque $\mathbb{1}_k$, $\forall k \in [\![1, K]\!]$,

Phases d'attaque $\phi_k^o(f,t)$, $\forall t \in \Omega_k$,

Nombre d'itérations N_{it}.

Poids $\lambda_k = \dfrac{V_k^2}{\sum_l V_l^2}$.

pour $t = 1$ à $T-1$ **faire**

  pour $k = 1$ à K **faire**

    si $\mathbb{1}_k(t) = 1$ **alors**

      Phase d'attaque : $\forall f$, $\phi_k(f,t) = \phi_k^o(f,t)$.

    sinon

      Déroulé linéaire (cf. Algorithme 2).

    fin si

    $\forall f$, $\hat{X}_k^{(0)}(f,t) = V_k(f,t)e^{i\phi_k(f,t)}$.

  fin pour

  Calculer $\forall f : E^{(0)}(f,t) = X(f,t) - \sum_k \hat{X}_k^{(0)}(f,t)$.

  pour $it = 1$ à N_{it} **faire**

    Distribution de l'erreur $\forall k, f : Y_k^{(it)}(f,t) = \hat{X}_k^{(it-1)}(f,t) + \lambda_k(t)E^{(it-1)}(f,t)$.

    Normalisation $\forall k, f : \hat{X}_k^{(it)}(f,t) = \dfrac{Y_k^{(it)}(f,t)}{|Y_k^{(it)}(f,t)|}V_k(f,t)$.

    Erreur $\forall f : E^{(it)}(f,t) = X(f,t) - \sum_k \hat{X}_k^{(it)}(f,t)$.

  fin pour

  Composante complexe $\forall k, f : \hat{X}_k(f,t) = \hat{X}_k^{(N_{it})}(f,t)$.

fin pour

Sortie :

$\forall k \in [\![1, K]\!]$, $\hat{X}_k \in \mathbb{C}^{F \times T}$.

5.3 Résultats expérimentaux

5.3.1 Protocole

Les signaux sont échantillonnées à 44100 Hz et la TFCT est calculée avec une fenêtre de Hann de longueur 4096 échantillons (soit 92 ms), 75 % de recouvrement et pas de bourrage de zéros.

Les données utilisées proviennent (à l'exception de certains mélanges de notes de piano issues de la base MAPS Emiya et al. (2010)) de la base DSD100 Ono et al. (2015), pour lesquels nous disposons des sources séparées :

— `bass` : il s'agit de la partie de basse électrique ;

— `drums` : il s'agit des percussions (batterie notamment) ;

— `vocals` : il s'agit de la voix chantée ;

— `other` : ce sont tous les autres instruments d'accompagnement (guitare, piano, sons électroniques etc.).

On peut calculer le spectrogramme de chaque source isolément. Si on utilise ces valeurs dans les différents algorithmes, on parlera de scénario Oracle. Alternativement, dans l'expérience décrite à la section 5.3.4, on considérera des spectrogrammes d'amplitude estimés, afin de tester l'efficacité des différentes méthodes dans un cadre plus réalistes où les V_k ne sont plus parfaitement connus. Ces estimés sont obtenus par une KLNMF effectuée sur chaque spectrogramme de source isolée. Chaque KLNMF utilise 50 itérations de règles de mises à jour multiplicatives et un rang de factorisation égal à 10. Notons que ce n'est pas un scénario "aveugle", puisqu'on estime les spectrogrammes sur chaque source séparée, mais il nous informe néanmoins sur la performance des méthodes lorsque les spectrogrammes ne sont plus égaux à la vérité terrain.

À partir de ces spectrogrammes, on applique les méthodes d'estimation des composantes complexes suivantes :

— Le filtrage de Wiener Févotte et al. (2009), noté **Wiener** ;

— Le filtrage de Wiener consistant Le Roux et Vincent (2013), noté **W-Cons** ;

— Le déroulé horizontal, effectué sur chaque source séparément, sans tenir compte de la phase du mélange, noté **Unwrap** ;

— L'algorithme 3 d'estimation des composantes initialisé par la méthode de déroulé, ce qui correspond donc à l'algorithme 4, noté **Iter**.

La qualité de la séparation de source est mesurée par les SDR, SIR et SAR, calculés par la boîte à outils BSS Eval Vincent et al. (2006).

Le filtrage de Wiener consistant (introduit dans le chapitre 2, section 2.1.4) dépend d'un paramètre γ qui ajuste l'importance relative du filtrage de Wiener et de la contrainte de consistance. On apprend le paramètre γ optimal (au sens du maximum de SDR, SIR et SAR) sur la base de développement (50 morceaux issus de la base DSD100 différents de ceux utilisés pour les tests). L'influence du paramètre γ sur la qualité de la séparation est illustrée sur la figure 5.2. On choisit la valeur $\gamma = 4$ pour l'expérience de séparation de sources qui correspond à une valeur optimale pour l'utilisation du filtrage de Wiener consistant sur ce jeu de données, aussi bien pour le scénario Oracle que pour le cas non-Oracle.

5.3.2 Impact du caractère séquentiel de l'algorithme

Nous proposons tout d'abord une expérience pour justifier l'intérêt de la procédure telle que décrite dans l'Algorithme 4. En effet, dans cette procédure, on agit séquentiellement

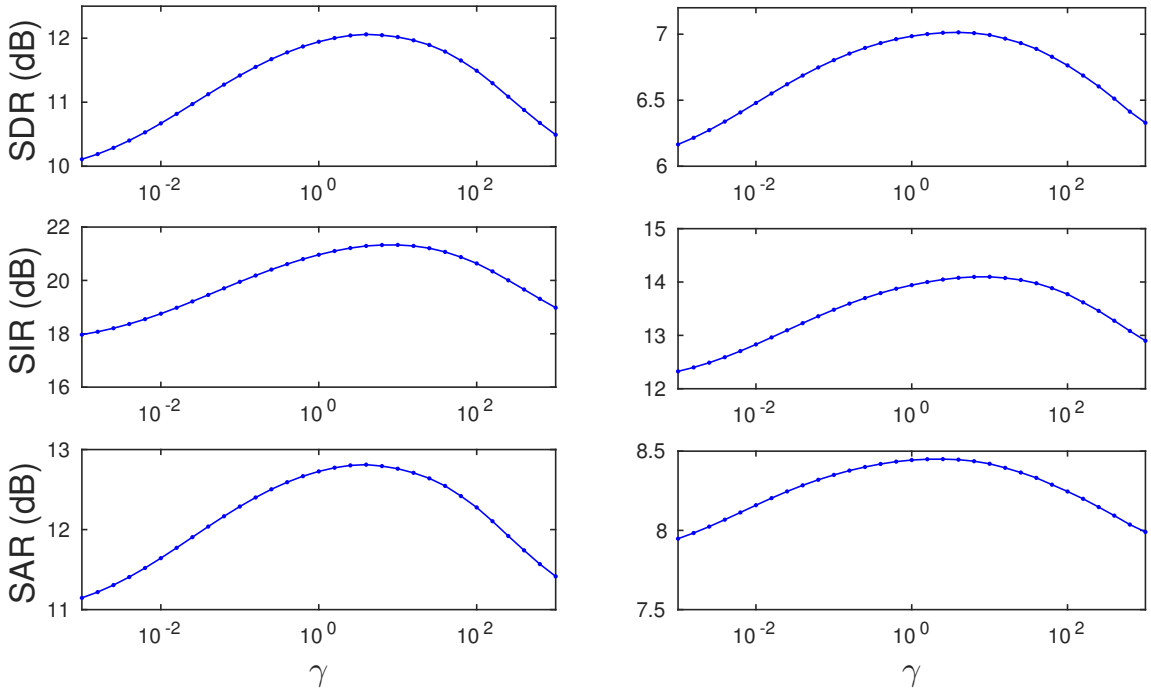


FIGURE 5.2 – Influence du paramètre de concentration γ sur la qualité de la séparation de sources pour le filtrage de Wiener consistant sur la base d'apprentissage de DSD100, dans le cas Oracle (à gauche) et non-Oracle (à droite).

	SDR	SIR	SAR	Temps (s)
Séquentiel	**14.0**	**27.0**	**14.2**	3.8
Direct	12.7	24.7	13.0	**3.6**

TABLEAU 5.1 – Performance de la séparation de sources sur la base DSD100 (SDR, SIR et SAR en dB) pour différentes techniques d'utilisation de la procédure itérative.

sur les trames : ainsi, l'initialisation de l'Algorithme 3 dans une trame donnée est faite par déroulé linéaire à partir de la phase obtenue après application de l'Algorithme 3 dans la trame précédente. On pourrait penser qu'effectuer un déroulé sur toute la TFCT, puis appliquer l'Algorithme 3 matriciellement serait moins gourmand en temps de calcul.

On considère donc 10 morceaux de musique issus de la base DSD100 (extraits de 10 secondes) et on applique deux méthodes de reconstruction des composantes complexes : une méthode *séquentielle*, c'est-à-dire telle que présentée dans l'Algorithme 4, et une méthode *directe*, c'est-à-dire en utilisant un déroulé complet (Algorithme 2) puis une application de l'Algorithme 3 matriciellement. Dans les deux cas, les phases dans les trames d'attaque ainsi que les amplitudes des sources sont supposées connues, et la procédure itérative utilise 10 itérations. Les résultats sont présentés dans le tableau 5.1.

La méthode séquentielle fournit des résultats supérieurs à la méthode directe. En effet, il est préférable d'estimer convenablement les phases dans une trame donnée avant de procéder au déroulé pour initialiser la trame suivante : la procédure itérative bénéficie alors d'une meilleure initialisation, ce qui conduit à de meilleurs résultats. Bien que la méthode directe soit très légèrement plus rapide, la perte de qualité nous incite à conserver la méthode telle que décrite dans l'Algorithme 4 pour la suite des expériences.

Initialisation	SDR	SIR	SAR
Aléatoire	10.4	20.6	10.9
Déroulé	**14.0**	**27.0**	**14.2**

TABLEAU 5.2 – Performance de la séparation de sources sur la base DSD100 (SDR, SIR et SAR en dB) pour différentes initialisations de l'algorithme 3.

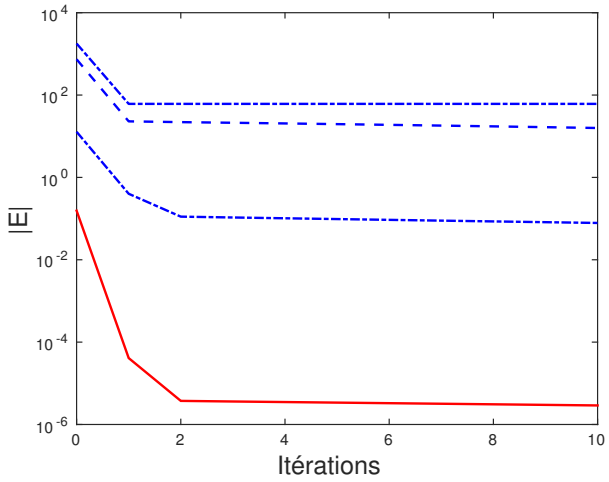


FIGURE 5.3 – Comparaison de l'erreur $|E|$ calculée au cours des itérations au niveau d'un point TF où deux sources (notes de piano C4 et G4) se recouvrent : les courbes en pointillés correspondent à une initialisation aléatoire (pour 30 initialisations différentes, les 3 courbes sont donc le maximum, le minimum et la valeur moyenne de l'erreur), et la courbe en traits pleins correspond à une initialisation par déroulé linéaire.

5.3.3 Influence de l'initialisation

Nous montrons dans cette expérience l'intérêt de l'initialisation par déroulé linéaire dans l'Algorithme 3. On considère 10 morceaux extraits de la base DSD100. Les phases d'attaque sont supposées connues et les phases des partiels sont estimées par 10 itérations de l'algorithme 4 à partir des amplitudes dans le scénario Oracle. Les phases des composantes complexes peuvent être initialisées soit avec des valeurs aléatoires, soit par déroulé linéaire. Les résultats de la séparation de sources sont fournis dans le tableau 5.2.

Initialiser l'algorithme par déroulé linéaire améliore significativement les résultats (gain d'environ 3.5 dB en SDR et SAR, et d'environ 6.5 dB en SIR) par rapport à une initialisation aléatoire.

Pour illustrer ce résultat, on effectue la séparation sur un mélange de notes de piano issues de la base MAPS, C4 et G4. On trace sur la figure 5.3 l'erreur au cours des itérations avec ces deux approches en un point TF où il y a recouvrement. Avec notre approche, l'erreur converge vers une valeur nettement plus basse.

Enfin, afin d'illustrer la différence entre les deux approches après application de l'algorithme, on trace sur la figure 5.4 les parties réelles des signaux originaux et reconstruits par notre algorithme avec différentes initialisations. On constate que l'utilisation du déroulé linéaire conduit à une reconstruction quasi-parfaite du partiel.

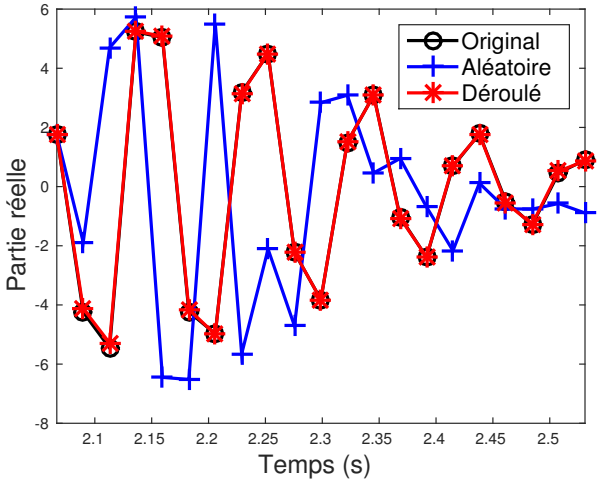


FIGURE 5.4 – Reconstruction d'un partiel de piano (partie réelle) de la note C4 dans la bande de fréquences à 796 Hz, sur une fenêtre de temps où les deux sources (C4 et G4) se recouvrent par application de notre algorithme avec différentes initialisations.

5.3.4 Test sur la base DSD100

On considère dans cette expérience 50 morceaux issus de la base DSD100 (base de test) et on applique les différentes méthodes de séparation de sources présentées dans la section 5.3.1. L'algorithme d'estimation des phases à partir du mélange utilise $N_{it} = 10$ itérations. Les résultats sont illustrés sur la figure 5.5.

Nous constatons que la méthode basée sur l'algorithme 4 conduit à une amélioration des performances par rapport au filtrage de Wiener, notamment en SIR, et à la méthode de déroulé linéaire appliqué isolément sur les sources. Cette dernière mène à des performances moindres du fait des erreurs dues à l'estimation de fréquence qui sont propagées à travers les trames.

L'algorithme que nous proposons mène à des résultats similaires au filtrage de Wiener consistant. Les différences entre ces deux méthodes ne sont pas statistiquement significatives en matière de SDR et SAR, mais notre approche conduit à une légère augmentation de SIR. Dans le cas non-Oracle, notre approche donne des résultats moins bons que le filtrage de Wiener consistant en SDR et SAR, et des résultats similaires pour le rejet d'interférences. Ce résultat est à relativiser car notre méthode présente deux avantages : tout d'abord, son temps de calcul est nettement moins élevé (d'un rapport 7). Par ailleurs, les résultats présentés ici avec le filtrage de Wiener consistant sont optimaux car on a utilisé un paramètre γ appris au préalable sur une base de développement. Notre méthode n'utilise pas de tel paramètre. Les résultats du filtrage de Wiener consistant sont très dépendant du choix de ce paramètre, et celui-ci peut varier significativement selon la base de données utilisée (pour les données de l'article originel LE ROUX et VINCENT (2013), la valeur optimale est $\gamma = 10^6$ contre 4 ici).

On illustre ce résultat sur un mélange simple, constitué de deux notes de piano (C4 et G4) qui se recouvrent dans le plan TF. On sépare les sources par différentes méthodes et on trace la partie réelle de la première source (C4) estimée dans une bande de fréquences et sur un intervalle de temps où le recouvrement est observé. Le résultat présenté sur la figure 5.6 montre une reconstruction du partiel quasi-parfaite avec la méthode basée sur l'algorithme 4. Le filtrage de Wiener conduit quant à lui à utiliser la phase du mélange dans laquelle les battements sont marqués, ce qui mène à une nette dégradation du signal reconstruit. Enfin, la technique de déroulé linéaire seule n'est pas parfaite mais est néanmoins relativement proche

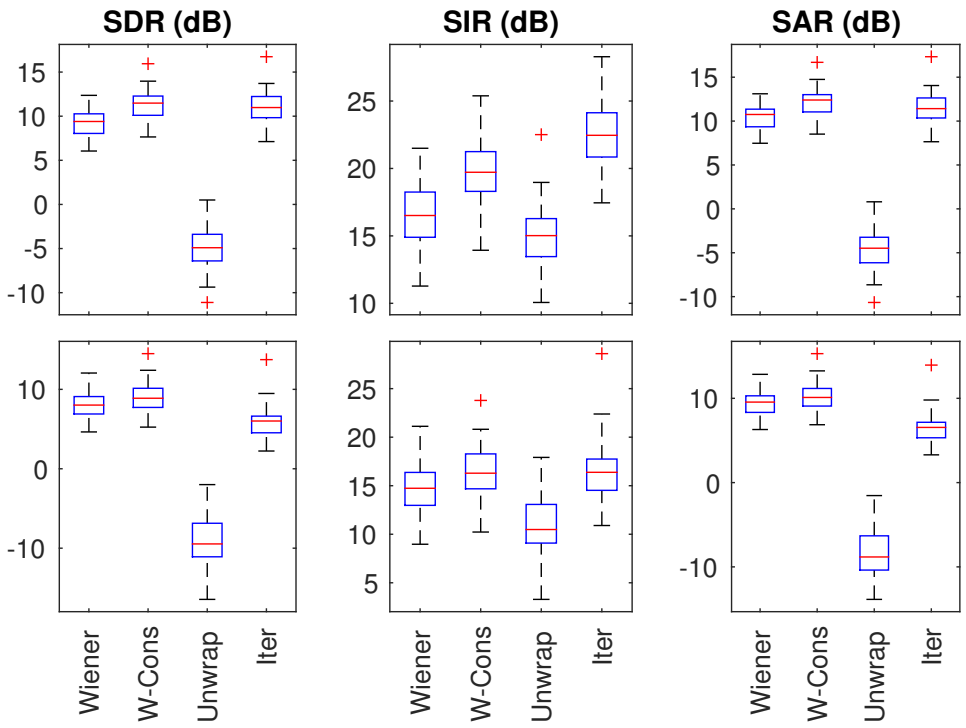


FIGURE 5.5 – Performance de la séparation de sources (SDR, SIR et SAR en dB) sur des morceaux de musiques de la base DSD100. Amplitudes Oracle (en haut) et estimées (en bas).

du résultat original (en comparaison avec le filtrage de Wiener), et permet notamment de s'affranchir des battements. Le filtrage de Wiener consistant mène à un résultat similaire au filtrage de Wiener.

Une évaluation subjective informelle de notre part montre que ces conclusions sont également constatées perceptivement. Le filtrage de Wiener crée des interférences entre sources à cause du recouvrement TF. Le déroulé seul élimine ces interférences, mais au prix de la création d'artéfacts qui dégradent la qualité globale des signaux. L'approche utilisant le déroulé linéaire ainsi que la phase du mélange s'affranchit de ces problèmes et conduit à des signaux proches de la reconstruction parfaite, dans lesquels ni interférences ni artéfacts ne sont audibles dans le cas Oracle.

5.3.5 Potentiel d'amélioration dans les trames d'attaque

Les phases dans les trames d'attaque étaient dans les expériences précédentes obtenues en prenant la phase du mélange ou en les supposant connues. Nous proposons donc d'appliquer l'algorithme 3 dans ces deux cas. La comparaison entre ces deux approches permettra une estimation de l'impact de la phase d'attaque sur la reconstruction du reste des phases, et pourra nous renseigner sur le potentiel d'amélioration d'estimation des phases d'attaque. Les amplitudes sont ici supposées connues.

Les résultats présentés dans le tableau 5.3 montrent qu'en supposant une connaissance parfaite des phases dans les trames d'attaque, on peut améliorer les résultats d'environ 2 dB en SDR et SAR et de 3 dB en SIR. Il existe donc une marge de progression possible pour

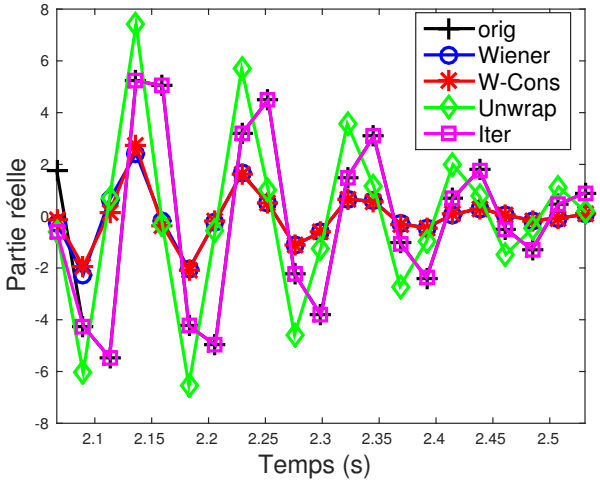


FIGURE 5.6 – Reconstruction d'un partiel de piano (partie réelle) de la note C4 dans le canal fréquentiel à 796 Hz, sur une fenêtre de temps où les deux sources (C4 et G4) se recouvrent. Plusieurs algorithmes de reconstruction des partiels sont comparés dans le cas Oracle.

Phases d'attaque	SDR	SIR	SAR
Mélange	11.2	22.3	11.7
Oracle	**13.2**	**25.4**	**13.6**

TABLEAU 5.3 – Comparaison des performances de séparation de sources selon la technique utilisée pour les phases d'attaque.

dépasser les résultats obtenus en utilisant la phase du mélange au niveau des attaques. Cela traduit l'importance de ces phases d'attaque, et montre qu'il peut être intéressant de travailler à les reconstruire plus proprement.

5.4 Algorithme contraint par le déroulé de phase

5.4.1 Principe

L'approche présentée dans ce chapitre consiste à minimiser une fonction de coût $|E|$ qui mesure l'écart entre le mélange et la somme des sources estimées. L'information à priori sur la phase est introduite par le biais de l'initialisation. Dans cette section, nous proposons d'ajouter à cette fonction de coût une contrainte sur la phase, pour que celle-ci reste relativement "proche" du candidat que nous proposons (la composante obtenue par déroulé linéaire). C'est ce type d'approche qui est proposé dans les algorithmes de NMF complexes à phase contrainte, comme LE ROUX et al. (2009) ou BRONSON et DEPALLE (2014).

Cette approche est posée comme un problème d'optimisation, qui consiste à minimiser la fonction :

$$\tilde{\mathcal{C}}(\theta) = |X - \sum_k \hat{X}_k|^2 + \sigma \sum_k |\hat{X}_k - \tilde{X}_k|^2 \tag{5.13}$$

sous la contrainte $|\hat{X}_k| = V_k$. Cette fonction objectif comporte un terme d'attache aux données $|E|^2$ et un terme d'attache à l'à priori, l'importance relative des deux étant réglée par le paramètre σ. L'à priori est $\tilde{X}_k$, dont la phase est obtenue par déroulé linéaire.

Pour minimiser cette fonction, on applique la méthode de la fonction auxiliaire de la même façon que cela a été fait pour l'algorithme 3 (pour rappel, les détails de ce calcul sont présentés

Algorithme 5 Estimation des sources complexes à partir de leur mélange, avec contrainte sur la phase.

Entrées :

Mélange $X \in \mathbb{C}$,

Modules $V_k \in \mathbb{R}_+$, poids λ_k et valeurs initiales $\hat{X}_k \in \mathbb{C}$, $\forall k \in [\![1, K]\!]$,

Candidat $\tilde{X}_k$, $\forall k \in [\![1, K]\!]$ et paramètre $\sigma \geq 0$,

Nombre d'itérations N_{it}.

Erreur initiale : $E = X - \sum_k \hat{X}_k$.

pour $it = 1$ à N_{it} **faire**

  pour $k = 1$ à K **faire**

    $Y_k \leftarrow \hat{X}_k + \lambda_k E$,

    $\hat{X}_k \leftarrow \dfrac{Y_k + \sigma\lambda_k \tilde{X}_k}{|Y_k + \sigma\lambda_k \tilde{X}_k|} V_k$.

  fin pour

  $E \leftarrow X - \sum_k \hat{X}_k$.

fin pour

Sorties :

$\forall k \in [\![1, K]\!]$, $\hat{X}_k \in \mathbb{C}$.

dans l'annexe B). La fonction auxiliaire est obtenue en utilisant l'inégalité de Jensen pour la fonction convexe $|.|^2$ ce qui requiert d'introduire des poids λ_k. Les mises à jour des variables auxiliaires Y_k sont inchangées, puisque le terme supplémentaire dans (5.13) ne dépend pas de ces variables. Enfin, la mise à jour des $\hat{X}_k$ ne se fait plus dans la direction des Y_k mais des $Y_k + \sigma\lambda_k \tilde{X}_k$.

Les règles de mise à jour pour l'estimation de ce modèle sont résumées dans l'Algorithme 5.

Les algorithmes 3 et 5 sont similaires, mais dans le 5, les composantes sont contraintes à rester proche du candidat $\tilde{X}_k$.

5.4.2 Résultats expérimentaux

Nous comparons ici expérimentalement les performances méthodes de séparation basées sur les algorithmes 3 et 5. On considère 10 morceaux extraits de la base DSD100, et on fait varier le paramètre σ. L'algorithme 5 peut être initialisé :

— Soit avec des composantes à phases aléatoires ;

— Soit avec des composantes dont les phases sont obtenues par déroulé linéaire.

On tire un certain nombre de conclusions à partir des résultats présentés sur la figure 5.7 :

— Si on initialise l'algorithme par déroulé linéaire, augmenter la valeur de σ à partir de 0 n'améliore pas les résultats. En d'autres termes, la performance maximale est obtenue pour une contrainte de phase nulle. Comme l'algorithme est initialisé convenablement, mieux vaut favoriser le terme d'attache aux données plutôt que forcer les solutions à rester trop proches de cette valeur initiale.

— Lorsque le poids σ est grand, on obtient avec cet algorithme (pour les deux initialisations) le même résultat qu'avec la méthode **Unwrap**. Le terme dans la fonction de coût (5.13) qui mesure l'écart entre modèle et mélange devient alors négligeable : les estimées des sources sont alors simplement données par les $\tilde{X}_k$.

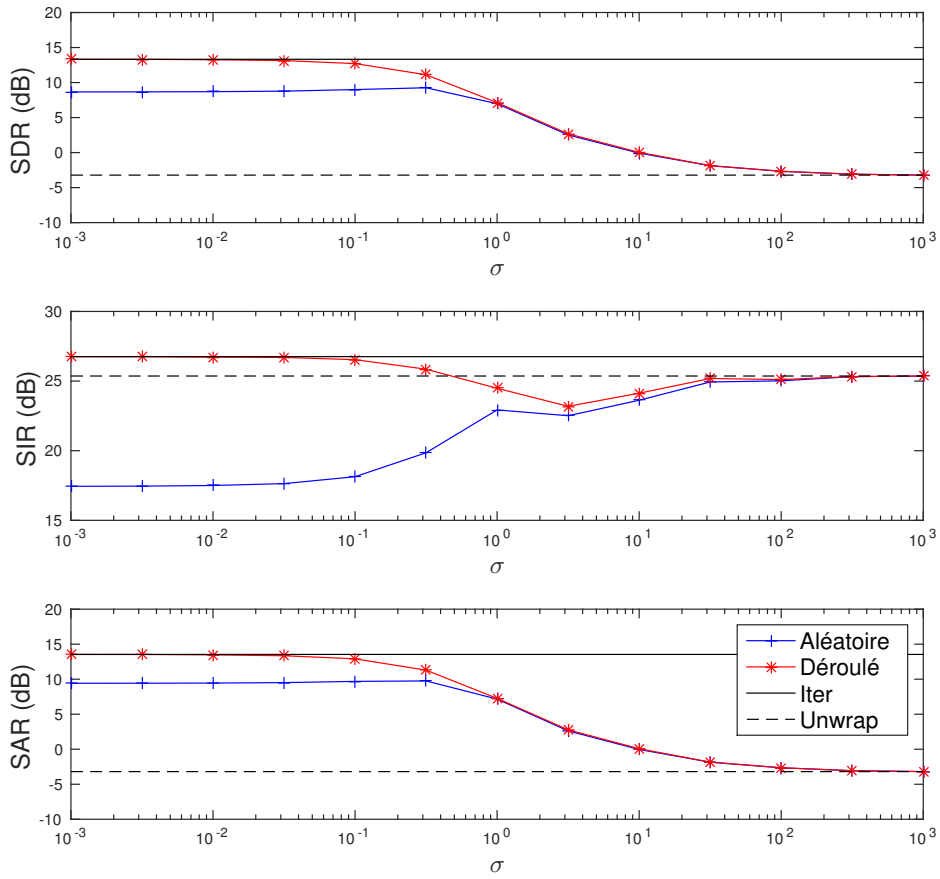


FIGURE 5.7 – Séparation de sources (SDR, SIR et SAR en dB) pour différentes initialisations (Aléatoire ou par Déroulé) de l'algorithme 5 sur le jeu de données E, et comparaison avec les approches **Iter** et **Unwrap**.

— Avec une initialisation aléatoire, on constate qu'en augmentant la valeur de σ, les indicateurs augmentent jusqu'à un certain seuil à partir duquel ils redescendent (pour le SDR et le SAR), ou alors continuent globalement d'augmenter (pour le SIR). Ceci peut s'expliquer par le fait qu'une initialisation aléatoire ne donne pas de très bons résultats pour l'algorithme non contraint, et que forcer les composantes estimées à être proches des $\tilde{X}_k$ augmentent nécessairement la qualité des résultats. Néanmoins, au-delà d'une certaine valeur de σ, la contrainte devient trop importante par rapport à l'erreur de reconstruction, ce qui se traduit par une chute des indicateurs SDR et SAR, même si le SIR continue de croître (ce qui s'interprète aisément, compte-tenu du fait qu'en négligeant le terme d'attache aux données, on ne tient plus compte du mélange, ce qui implique une suppression des interférences).

— L'initialisation avec déroulé linéaire, couplée à une contrainte nulle, correspond à la méthode **Iter** et donne donc les mêmes résultats. À partir de cette valeur, une augmentation de σ conduit à une augmentation des SDR et SAR qui reste négligeable (non visible sur la figure).

En fin de compte, on peut constater que la borne supérieure de résultats pour l'algorithme avec contrainte est donnée par la méthode **Iter**, valeur atteinte lorsqu'on initialise l'algorithme

avec le déroulé linéaire et que l'on fixe une contrainte nulle. En d'autres termes, cela revient à utiliser directement la méthode **Iter** : contraindre les phases n'apporte pas d'amélioration significative par rapport à une initialisation bien choisie.

5.5 Conclusion

Les expériences présentées dans ce chapitre ont montré l'intérêt de la technique de déroulé linéaire de phase pour la séparation de sources. Nous avons introduit une procédure itérative pour résoudre ce problème, qui permet, via son initialisation, d'introduire un à priori sur la phase. Cette méthode a été expérimentalement testée sur une base de musiques réalistes et nous avons obtenu des résultats comparables au filtrage de Wiener consistant avec un net bénéfice en temps de calcul.

La phase au niveau des trames d'attaque était dans ces expériences soit supposée connue, soit égale à la phase du mélange. Néanmoins, l'écart entre les deux identifié dans la section 5.3.5 suggère qu'il est possible d'améliorer les performances en travaillant à la reconstruction de phases d'attaque, ce qui fera l'objet du chapitre 6.

En outre, la procédure introduite ici consiste à fixer à chaque itération les amplitudes à une valeur objectif, qui n'est pas forcément réaliste. Il pourrait donc être préférable de mettre au point un modèle dans lequel on s'accorde une certaine incertitude sur les amplitudes, et dans lequel le mélange estimé est bien égal au mélange observé : cela fera l'objet du chapitre 8. Enfin, il sera intéressant de mettre en place un cadre de séparation plus réaliste dans lequel les amplitudes et les phases seront estimées conjointement à partir du mélange : c'est ce que nous proposons au chapitre 7.

Chapitre 6

Modèle de phase d'attaque basé sur la répétition d'évènements audio

Sommaire

Comme on l'a indiqué dans l'introduction de ce manuscrit, la majorité des techniques de séparation de sources reposent sur le caractère répétitif des évènements qui constituent les données. En audio, on peut voir une source comme étant constituée de la répétition d'un évènement élémentaire (comme une note). La NMF repose sur le postulat de l'existence d'un spectre associé à chaque évènement audio, qui est activé à différents instants avec un gain variable. Néanmoins, les répétitions de phase ne sont guère exploitées. Or, comme c'est l'évènement audio (un signal temporel) qui est redondant, on peut intuitivement penser qu'il existe une forme de redondance qui se retrouve dans sa TFCT, aussi bien au niveau du spectrogramme d'amplitude que de la phase.

Nous proposons donc d'exploiter ces redondances au niveau des attaques des sources musicales pour reconstruire les phases dans les trames d'attaque dans le domaine TF. On postule l'existence d'une *phase de référence* et on suppose que pour une source donnée, la phase au niveau des trames d'attaque est égale à cette phase de référence, à laquelle est ajouté un décalage qui est une fonction linéaire de la fréquence. Ce modèle est testé sur divers signaux afin d'en attester la validité. Nous proposons également de le combiner à l'algorithme de déroulé linéaire afin de restaurer complètement la phase d'un signal, dans un cadre supervisé (les phases de référence sont préalablement apprises à partir d'une base de données externe).

Nous proposons ensuite d'intégrer cette propriété dans un modèle de mélange de sources complexes au niveau des attaques. Ce modèle de mélange est estimé par deux algorithmes, la contrainte de phase pouvant être stricte ou relaxée. Une fois la restauration des phases complétée par l'application de l'algorithme de déroulé linéaire, nous obtenons alors une procédure complète de reconstruction de phase dans le cadre de la séparation de sources.

Les contributions de ce chapitre ont fait l'objet d'une publication à la conférence WASPAA 2015 Magron et al. (2015a).

La section 6.1 présente ce modèle de phase, qui est expérimentalement validé dans la section 6.2. Cette propriété est intégrée à un modèle de mélange dans la section 6.3, dont l'estimation des paramètres est décrite dans la section 6.4. Des expériences de séparation de sources sont conduites dans la section 6.5, et nous concluons dans la section 6.6.

6.1 Modèle de phase d'évènements audio répétés

Considérons un signal réel $x(n)$. On rappelle l'expression de sa TFCT, pour chaque canal fréquentiel $f \in [\![0, F-1]\!]$ et trame $t \in [\![0, T-1]\!]$:

$$X(f,t) = \sum_{n=0}^{N_w-1} x(n+tS)w(n)e^{-2i\pi\frac{f}{F}n}, \tag{6.1}$$

où w est une fenêtre d'analyse de taille N_w, et S est le décalage temporel (en échantillons) entre deux trames successives. On note $x_w^t(n) = x(n+tS)w(n)$ le signal fenêtré correspondant à la trame d'analyse t, et on a donc :

$$X(f,t) = \sum_{n=0}^{N_w-1} x_w^t(n)e^{-2i\pi\frac{f}{F}n}. \tag{6.2}$$

Supposons que le signal x représente un évènement audio élémentaire qui est activé plusieurs fois (par exemple une note de piano) : les indices de trames t_m, $m \in [\![0, M-1]\!]$, correspondent aux M trames d'attaque de cette source dans le domaine TF. L'idée qui est au coeur du modèle de ce chapitre consiste à supposer que les signaux fenêtrés $x_w^{t_m}(n)$ dans les trames

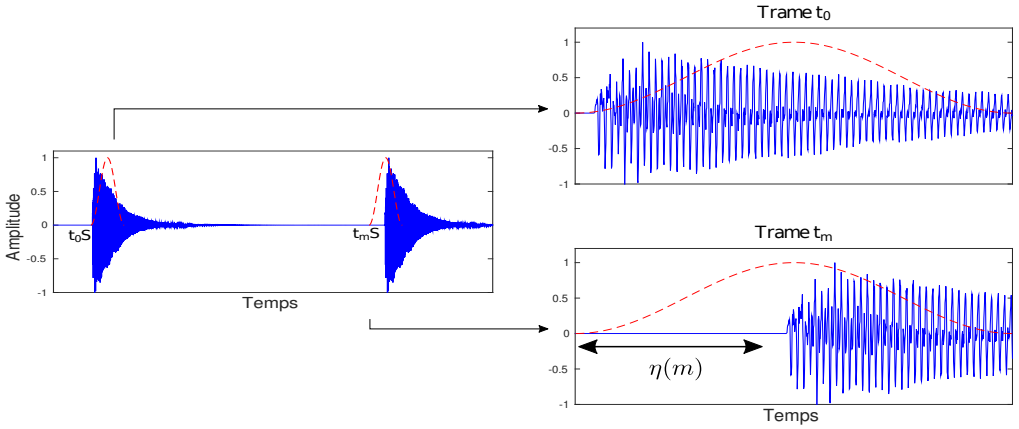


FIGURE 6.1 – Répétition d'un évènement audio (traits pleins) et positionnement de la fenêtre d'analyse (pointillés) : on suppose que les signaux fenêtrés sont identiques à un délai et facteur d'amplitude près.

d'attaque sont égaux, à un délai η et un facteur d'amplitude ρ près. Ainsi, en prenant comme référence le signal fenêtré dans la première trame d'attaque, on a :

$$x_w^{t_m}(n) \approx \rho(m) x_w^{t_0}(n - \eta(m)), \tag{6.3}$$

comme c'est illustré sur la figure 6.1. Le délai de $\eta(m)$ échantillons [1] dépend donc du positionnement du signal par rapport à la fenêtre d'analyse. Cette hypothèse n'est vérifiée que pour une fenêtre d'analyse rectangulaire et lorsque la totalité du signal se trouve à l'intérieur de la fenêtre : cela ne correspond pas à des cas d'application pratiques, puisqu'on utilise généralement une fenêtre plus sophistiquée (afin d'éviter le phénomène de fuite spectrale), et que les signaux sont tronqués par l'opération de fenêtrage. En outre, ce modèle repose sur l'hypothèse que la forme d'onde des signaux correspondants aux diverses occurrences d'un même évènement sonore soient identique (au facteur d'amplitude près) : pour plus de réalisme, il faudrait considérer des variations de forme d'onde qui correspondent aux différentes nuances du jeu du musicien. Nous étudierons expérimentalement la validité de ce modèle lorsque l'on considère des occurrences d'une source qui comportent des variations de nuance.

La TFCT de x dans la trame t_m est donc :

$$\begin{aligned} X(f, t_m) &= \sum_{n=0}^{N_w-1} x_w^{t_m}(n) e^{-2i\pi \frac{f}{F} n} \\ &\approx \sum_{n=0}^{N_w-1} \rho(m) x_w^{t_0}(n - \eta(m)) e^{-2i\pi \frac{f}{F} n} \\ &\approx \rho(m) e^{2i\pi \frac{f}{F} \eta(m)} X(f, t_0). \end{aligned}$$

On pose $\lambda(m) = \dfrac{2\pi\eta(m)}{F}$, ce qui conduit à :

$$X(f, t_m) \approx \rho(m) e^{i\lambda(m)f} X(f, t_0). \tag{6.4}$$

1. Cette écriture suppose que le délai $\eta(m)$ soit entier. En effet, lorsque celui-ci devient fractionnaire, l'expression (6.3) n'a plus de sens. Néanmoins, on introduit ultérieurement la notation $\lambda(m)$ dans le domaine TF, qui permet d'étendre la notation ç des valeurs continues du délai.

En notant X sous forme polaire (module V et argument ϕ), (6.4) devient :

$$V(f,t_m)e^{i\phi(f,t_m)} \approx \rho(m)V(f,t_0)e^{i(\phi(f,t_0)+\lambda(m)f)}. \tag{6.5}$$

La relation (6.5) contient à la fois une information sur les modules et sur les phases. Si on note $W(f) = V(f,t_0)$ et $H(t_m) = \rho(m)$, on obtient $V(f,t_m) \approx W(f)H(t_m)$. En généralisant ce résultat à toutes les trames, on retrouve le modèle NMF sur les amplitudes. Ainsi, ce modèle d'évènements audio répétés conduit naturellement au modèle NMF traditionnel sur les spectrogrammes d'amplitude. Ce qui nous intéresse ici est la relation sur les phases :

$$\phi(f,t_m) \approx \psi(f) + \lambda(m)f, \tag{6.6}$$

avec $\psi(f) = \phi(f,t_0)$, à laquelle on se réfère sous le nom de *phase de référence*. L'équation (6.6) traduit donc le modèle de phase que nous proposons : dans une trame d'attaque, la phase de la TFCT X est égale à une phase de référence (ne dépendant pas de l'indice de trame) à laquelle est ajouté un délai linéaire en fréquence. La forme de la phase de référence $\psi(f)$ caractérise alors le timbre de l'instrument (et notamment les relations de phase entre partiels), alors que le délai λ informe sur le positionnement temporel du signal au sein de la fenêtre d'analyse.

6.2 Validation expérimentale préliminaire

On cherche dans cette partie à mettre en évidence la phase de référence introduite précédemment. On suppose connues les trames d'attaque t_m, $m \in [\![0,1]\!]$, bien que l'on pourrait par la suite les estimer (à partir par exemple de la MATLAB Tempogram Toolbox Grosche et Müller (2011)).

6.2.1 Précision du modèle de phase

On considère dans cette expérience des signaux constitués d'une source (un évènement audio) qui est activée à deux reprises. Ces signaux sont :

— des notes de piano tirées aléatoirement de la base MAPS Emiya et al. (2010),

— des notes de guitare électrique tirées aléatoirement de la base IDMT-SMT-GUITAR Kehling et al. (2014).

Ces sources peuvent être activées soit à l'identique, c'est-à-dire sans variation de forme d'onde, soit avec une certaine nuance (afin de reproduire de façon plus réaliste les variations de jeu de l'instrumentiste). Dans le cas des notes de piano, les nuances possibles sont "mezzo-forte","forte" et "piano". Pour les notes de guitare, les nuances sont obtenues en considérant des notes qui sont jouées à différents endroits du manche (en effet, avec la guitare, on peut produire une même note de plusieurs façons différentes, selon la case et la corde choisie).

On calcule la différence de phase $\Delta\phi(f) = \phi(f,t_1) - \phi(f,t_0)$ entre les trames d'attaque. On s'attend à ce que cet écart de phase soit linéaire en f pour qu'il respecte le modèle (6.6). On calcule donc une approximation de ce décalage entre phases d'attaque par régression linéaire, et on illustre cette expérience sur la figure 6.2 dans le cas où il n'y a pas de variation de forme d'onde entre les deux occurrences de notes.

Afin de mesurer la pertinence de ce modèle, nous calculons (sur 25 signaux dans chaque cas de figure) l'erreur moyennée sur les fréquences (en radians) entre décalages de phases d'attaque observés et estimés par le modèle, et nous présentons les résulats dans le tableau 6.1.

La première ligne de ce tableau confirme ce que l'on constate visuellement sur la figure 6.2 : lorsqu'une source est activée "à l'identique", c'est-à-dire sans variation de forme d'onde, le

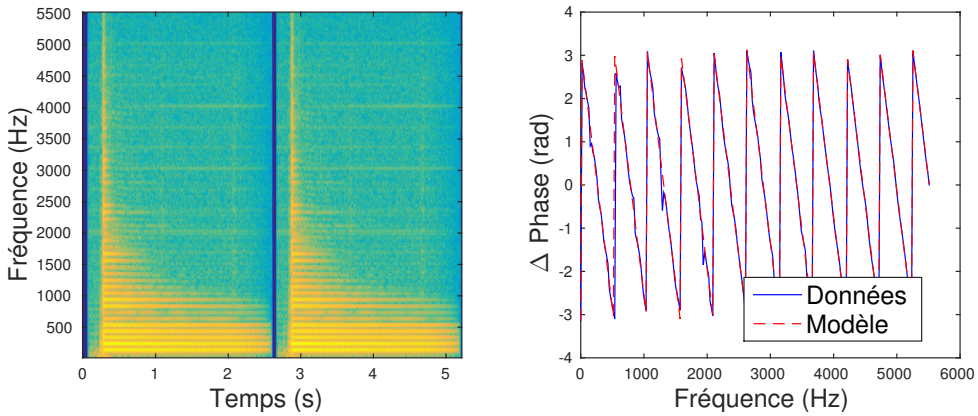


FIGURE 6.2 – Spectrogramme comportant deux occurrences d'une note de guitare sans variation de forme d'onde (gauche) et décalages de phase entre attaques (droite).

	Piano	Guitare
Sans nuance	0.21	0.15
Avec nuance	1.48	1.49

TABLEAU 6.1 – Erreur moyenne (en radians) entre décalage de phases d'attaque observé et estimé par régression linéaire pour plusieurs types de données.

modèle est relativement précis. Lorsque les notes sont activées avec une certaine nuance variable, il reste possible d'exploiter une phase de référence pour caractériser la phase dans les trames d'attaque, même si ce modèle perd nettement en précision par rapport au cas de sources identiques.

6.2.2 Combinaison avec le déroulé linéaire

Nous combinons à présent le modèle de phase (6.6) et le déroulé linéaire introduit au chapitre 4 pour achever la reconstruction des phases à partir de celle des attaques.

Phase apprise à partir d'une première occurrence de la note

Nous considérons des mélanges constitués de deux occurrences d'une note (piano ou guitare) avec ou sans nuance. Nous reconstruisons la phase au niveau de la deuxième activation à partir de la première. Après avoir reconstruit la phase dans la deuxième trame d'attaque, nous achevons la reconstruction de phase de la deuxième occurrence de note par déroulé linéaire. Nous effectuons également une reconstruction par l'algorithme de Griffin et Lim (GL) qui utilise 200 itérations, et on mesure la qualité de la reconstruction par le SDR VINCENT et al. (2006). Les résultats sont présentés dans le tableau 6.2.

Globalement, notre modèle de phase améliore la qualité de reconstruction du signal par rapport à l'algorithme GL. Cela montre l'importance de cette phase d'attaque, qui est non seulement indispensable pour conduire à un résultat perceptif satisfaisant (attaque nette) mais également pour initialiser un déroulé pertinent des phases des partiels.

Base de phases de référence

En pratique, comme toutes les phases sont inconnues, nous n'avons pas à disposition la phase de référence. Une possibilité est alors d'utiliser une base de données de phases d'attaque.

	Griffin Lim	Modèle + Déroulé
Piano (sans nuance)	−3.4	**−0.5**
Guitare (sans nuance)	−2.6	**−0.3**
Piano (avec nuance)	−6.1	**−5.1**
Guitare (avec nuance)	−1.0	**0.3**

TABLEAU 6.2 – Qaulité de reconstruction de phase (SDR en dB) d'une deuxième activation d'une source à partir de la première via un modèle de phase d'attaque combiné au déroulé linéaire.

	Griffin Lim	Modèle+Déroulé	Oracle+Déroulé
Morceaux de piano	−4.4	−3.3	−0.1
Morceaux de guitare	−4.3	−3.7	−0.6

TABLEAU 6.3 – Qualité de la reconstruction de signal (SDR en dB) en utilisant une base de données de phases d'attaque combinée au déroulé linéaire.

Nous considérons donc deux ensembles de notes (piano et guitare) à partir desquels nous fabriquons deux dictionnaires de phases de référence. Les dictionnaires sont construit avec une seule nuance par note.

On considère ensuite deux jeux de données : des morceaux de piano tirés de la base MAPS et des morceaux de guitare tirés de la base IDMT-SMT-GUITAR (ces bases comportent en effet aussi bien des notes isolées que des morceaux de musique, qui comportent donc plusieurs nuances des notes). Les trames d'attaque sont détectées grâce à la boîte à outils MATLAB Tempogram toolbox. Le délai λ est estimé grâce à un calcul sur les amplitudes similaire à celui conduit dans le chapitre 4 (section 4.4.1). Il est alors appliqué à la phase de référence contenue dans la base afin de reconstruire la phase d'attaque du signal. Les phases des partiels sont ensuite reconstruites par déroulé linéaire. A titre de comparaison, on reconstruit également les phases par déroulé linéaire à partir de phases d'attaque connues (cas Oracle). Nous testons enfin l'algorithme de Griffin Lim qui utilise 200 itérations.

Les résultats présentés dans le tableau 6.3 montrent l'intérêt d'utiliser une base de phases de références pour la reconstruction des phases d'attaque. Ce modèle améliore en effet légèrement les résultats par rapport à l'algorithme GL, même si sa performance reste éloignée de la performance Oracle.

6.3 Modèle de mélange de sources

Considérons à présent la TFCT X d'un mélange de K sources X_k, dont les amplitudes et les phases sont notées V_k et ϕ_k respectivement. On note t_m, $m \in [\![0, M-1]\!]$ les indices des M trames d'attaque de X. On peut extraire de X la matrice des attaques $Y \in \mathbb{C}^{F \times M}$:

$$Y(f,m) = X(f,t_m) = \sum_{k=1}^{K} V_k(f,t_m) e^{i\phi_k(f,t_m)}. \tag{6.7}$$

On introduit le modèle de phase (6.6) pour chaque source complexe :

$$\phi_k(f,t_m) = \psi_k(f) + \lambda_k(m) f. \tag{6.8}$$

On aboutit alors au modèle de mélange suivant : $\forall (f,m) \in [\![0, F-1]\!] \times [\![0, M-1]\!]$,

$$\hat{Y}(f,m) = \sum_{k=1}^{K} V_k(f,t_m) e^{i\psi_k(f)} e^{i\lambda_k(m)f}. \tag{6.9}$$

Remarque : Il serait plus rigoureux de considérer des ensembles de trames d'attaque différents pour chaque source, car les paramètres λ_k et ψ_k n'ont de sens que dans les trames d'attaque pour la source k. Néanmoins, dans ce chapitre, on considère des données pour lesquelles une trame d'attaque du mélange est une trame d'attaque pour toutes les sources, ou bien telle que si une trame d'attaque pour une source ne l'est pas pour une autre, alors cette autre source a une énergie nulle dans cette trame. Ainsi, le fait de considérer les amplitudes dans le modèle de mélange permet de rendre compte de cette hypothèse. Pour des données réalistes plus complexes, il est nécessaire de distinguer les trames d'attaque selon les sources : c'est ce que nous proposons dans le chapitre 7.

Ce modèle basé sur la répétition de phases permet de réduire la dimension des données : les phases des sources sont initialement représentées par KFM paramètres alors que notre modèle utilise $K(F+M-1)$ paramètres. Si $M > 1$ (ce qui est en pratique le cas, puisque on suppose qu'il y a plusieurs activations des sources), on observe bien que $K(F+M-1) < KFM$. Pour un mélange de 3 sources avec 5 trames d'attaque et une transformée de Fourier à 512 canaux, ce modèle utilise presque cinq fois moins de paramètres que les données originales.

Ce modèle peut se réécrire sous la forme matricielle suivante :

$$\hat{Y} = \sum_{k=1}^{K} V_k \odot (\Psi_k \Lambda_k), \tag{6.10}$$

où $\Psi_k = \texttt{diag}(e^{i\psi_k(0)}, ..., e^{i\psi_k(F-1)})$ est une matrice diagonale de dimensions $F \times F$ contenant les phases de référence, et Λ_k est une matrice de Vandermonde de dimensions $F \times M$ contenant les paramètres de délai :

$$\Lambda_k = \begin{pmatrix} 1 & \dots & 1 \\ e^{i\lambda_k(0)} & \dots & e^{i\lambda_k(M-1)} \\ \vdots & & \vdots \\ e^{i\lambda_k(0)(F-1)} & \dots & e^{i\lambda_k(M-1)(F-1)} \end{pmatrix} \tag{6.11}$$

6.4 Estimation des phases des composantes

Dans cette partie, nous estimons les paramètres du modèle (6.9). On suppose que les amplitudes des sources V_k dans les trames d'attaque sont connues. On cherche donc à estimer les paramètres $\psi_k(f)$ et $\lambda_k(m)$, $\forall(f, m, k)$. Une première méthode consiste à minimiser une fonction de coût représentant l'erreur entre les données et le modèle (contrainte *stricte*).

Lorsque les données ne correspondent plus exactement au modèle, une contrainte stricte peut être trop forte pour estimer les paramètres. On s'inspire donc de Rigaud et al. (2013) pour proposer une méthode alternative d'estimation, qui repose sur une contrainte *relaxée*.

6.4.1 Contrainte stricte

On considère la fonction de coût suivante :

$$\mathcal{C}_s = \sum_{f,m} \left| Y(f,m) - \sum_{k=1}^{K} V_k(f,t_m) e^{i\psi_k(f)} e^{i\lambda_k(m)f} \right|^2. \tag{6.12}$$

Nous minimisons cette fonction par rapport à chacune des variables successivement, ce qui conduit à une procédure itérative. Cette méthode est connue sous le nom de méthode de relaxation.

Estimation de $\psi_k(f)$

Réécrivons $\mathcal{C}_s$ en isolant les termes dépendant de $\psi_k(f)$:

$$\mathcal{C}_s = \sum_m \left| Y(f,m) - \sum_{l\neq k} V_l(f,t_m)e^{i\psi_l(f)}e^{i\lambda_l(m)f} - V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f} \right|^2 + \sum_m \sum_{p\neq f} \left| Y(p,m) - \sum_l V_l(p,t_m)e^{i\psi_l(p)}e^{i\lambda_l(m)p} \right|^2. \quad (6.13)$$

On pose

$$B_k(f,m) = Y(f,m) - \sum_{l\neq k} V_l(f,t_m)e^{i\psi_l(f)}e^{i\lambda_l(m)f}, \quad (6.14)$$

et on note $\overset{c}{=}$ l'égalité à une constante additive près, ne dépendant pas de la variable considérée, ici $\psi_k(f)$. On peut alors écrire :

$$\mathcal{C}_s \overset{c}{=} \sum_m \left| B_k(f,m) - V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f} \right|^2, \quad (6.15)$$

que l'on développe :

$$\begin{aligned}
\mathcal{C}_s &\overset{c}{=} \sum_m \left| B_k(f,m) - V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f} \right|^2 \\
&\overset{c}{=} \sum_m |B_k(f,m)|^2 + |V_k(f,t_m)|^2 - 2\Re(\overline{B}_k(f,m)V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f}) \\
&\overset{c}{=} -2\Re\left(\sum_m \overline{B}_k(f,m)V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f} \right) \\
&\overset{c}{=} -2\Re\left(e^{i\psi_k(f)} \sum_m \overline{B}_k(f,m)V_k(f,t_m)e^{i\lambda_k(m)f} \right),
\end{aligned}$$

où $\Re$ désigne la partie réelle. On pose $z_s = \sum_m B_k(f,m)V_k(f,t_m)e^{-i\lambda_k(m)f}$ (on enlève les indices k et f pour plus de clarté), que l'on note en coordonnées polaires $z_s = |z_s|e^{i\xi_s}$. On a :

$$\begin{aligned}
\mathcal{C}_s &\overset{c}{=} -2|z_s|\Re\left(e^{i\psi_k(f)}e^{-i\xi_s}\right) \\
&\overset{c}{=} -2|z_s|\cos(\psi_k(f) - \xi_s).
\end{aligned}$$

La fonction de coût, ainsi écrite, est aisée à minimiser par rapport à $\psi_k(f)$. En effet, minimiser $\mathcal{C}_s$ revient à maximiser le cosinus de $\psi_k(f) - \xi_s$, ce qui conduit à $\psi_k(f) = \xi_s = \angle(z_s)$, soit :

$$\psi_k(f) = \angle\left(\sum_m B_k(f,m)V_k(f,t_m)e^{-i\lambda_k(m)f} \right). \quad (6.16)$$

Remarque : Il est possible d'obtenir ce résultat en passant par la dérivée de $\mathcal{C}_s$ par rapport à $\psi_k(f)$ (ce que nous avons proposé dans Magron et al. (2015a)), qui est :

$$\frac{\partial \mathcal{C}_s}{\partial \psi_k(f)} = i\sum_m B_k(f,m)V_k(f,t_m)e^{-i\lambda_k(m)}e^{-i\psi_k(f)} - \overline{B}_k(f,m)V_k(f,t_m)e^{i\lambda_k(m)}e^{i\psi_k(f)}. \quad (6.17)$$

Annuler cette dérivée partielle conduit à l'estimation suivante pour $\psi_k(f)$:

$$\psi_k(f) = \pm\angle\left(\sum_m B_k(f,m)V_k(f,t_m)e^{-i\lambda_k(m)f}\right). \tag{6.18}$$

Pour lever l'ambiguïté sur le signe de l'argument, on peut examiner le signe de la dérivée seconde. Pour la valeur donnée par (6.16), celle-ci est positive, ce qui confirme que cette solution est bien un minimum de $\mathcal{C}_s$.

Estimation de $\lambda_k(m)$

On écrit à présent $\mathcal{C}_s$ de façon similaire à (6.15) en isolant les termes dépendant de $\lambda_k(m)$:

$$\mathcal{C}_s \overset{c}{=} \sum_f \left|B_k(f,m) - V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f}\right|^2. \tag{6.19}$$

Minimiser $\mathcal{C}_s$ par rapport à $\lambda_k(m)$ revient donc à minimiser la quantité suivante :

$$\tilde{\mathcal{C}_s}(k,m) = \sum_f \left|B_k(f,m)e^{-i\psi_k(f)} - V_k(f,t_m)e^{i\lambda_k(m)f}\right|^2. \tag{6.20}$$

On note :

$$\begin{aligned}
\beta_k(f,m) &= B_k(f,m)e^{-i\psi_k(f)},\\
\underline{\beta}_k(m) &= [\beta_k(0,m),...,\beta_k(F-1,m)]^T,\\
\underline{\Lambda}_k(m) &= [1, e^{i\lambda_k(m)},...,e^{i\lambda_k(m)(F-1)}]^T,\\
\underline{V}_k(m) &= [V_k(0,t_m),...,V_k(F-1,t_m)]^T.
\end{aligned}$$

La fonction (6.20) se reformule alors comme suit :

$$\tilde{\mathcal{C}_s}(k,m) = ||\underline{V}_k(m) \odot \underline{\Lambda}_k(m) - \underline{\beta}_k(m)||^2, \tag{6.21}$$

où $||.||$ désigne la norme euclidienne pour les vecteurs. La solution à ce problème peut être obtenue en s'inspirant de l'algorithme ESPRIT Hua et al. (2004). En effet, lorsque la fonction de coût $\tilde{\mathcal{C}_s}$ est nulle, on remarque que (on s'affranchit des indices k et m pour plus de lisibilité) :

$$\underline{\beta}_\downarrow^H \underline{\beta}_\uparrow = (\underline{V} \odot \underline{\Lambda})_\downarrow^H (\underline{V} \odot \underline{\Lambda})_\uparrow = \underline{V}_\downarrow^H \underline{V}_\uparrow e^{i\lambda}, \tag{6.22}$$

où $.^H$ désigne le transposé Hermitien, et la notation $\underline{v}_\downarrow$ (respectivement $\underline{v}_\uparrow$) désigne le vecteur obtenu en retirant le dernier élément (respectivement le premier) d'un vecteur $\underline{v}$. Ainsi :

$$e^{i\lambda} = \frac{\underline{\beta}_\downarrow^H \underline{\beta}_\uparrow}{\underline{V}_\downarrow^H \underline{V}_\uparrow}. \tag{6.23}$$

L'estimation de $\lambda_k(m)$ est donc :

$$\lambda_k(m) = \angle\left(\underline{\beta}_k(m)_\downarrow^H \underline{\beta}_k(m)_\uparrow\right). \tag{6.24}$$

Algorithme 6 Estimation des paramètres de phase sous contrainte stricte

Entrées : $Y \in \mathbb{C}^{F\times M}$, $V_k \in \mathbb{R}_+^{F\times M}$, $\psi_k^{ini} \in \mathbb{R}^{F\times 1}$, $\lambda_k^{ini} \in \mathbb{R}^{1\times M}$.
Initialisation :
$\psi_k = \psi_k^{ini}$, $\lambda_k = \lambda_k^{ini}$.
$\hat{Y}_k(f,m) = V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f}$.
$\hat{Y}(f,m) = \sum_{k=1}^{K} \hat{Y}_k(f,m)$.
$B_k(f,m) = Y(f,m) - \hat{Y}(f,m) + \hat{Y}_k(f,m)$.
tant que critère non atteint **faire**
  pour $k = 1$ à K, $f = 0$ à $F-1$ et $m = 0$ à $M-1$ **faire**
    Calculer ψ
    $\psi_k(f) = \angle\left(\sum_m B_k(f,m)V_k(f,t_m)e^{-i\lambda_k(m)f}\right)$.
    Calculer $\underline{\beta}$
    $\beta_k(f,m) = B_k(f,m)e^{-i\psi_k(f)}$,
    $\underline{\beta}_k(m) = [\beta_k(0,m), ..., \beta_k(F-1,m)]^T$.
    Calculer λ
    $\lambda_k(m) = \angle\left(\underline{\beta}_k(m)_\downarrow^H \underline{\beta}_k(m)_\uparrow\right)$.
    Calculer $\hat{Y}$
    $\hat{Y}_k(f,m) = V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f}$,
    $\hat{Y}(f,m) = \sum_{k=1}^{K} \hat{Y}_k(f,m)$.
    Calculer B
    $B_k(f,m) = Y(f,m) - \hat{Y}(f,m) + \hat{Y}_k(f,m)$.
  fin pour
fin tant que
Sorties : $\hat{Y}, \hat{Y}_k, \psi, \lambda$.

Algorithme d'estimation strict

Les équations (6.16) et (6.24) fournissent des règles de mises à jour pour estimer itérativement les paramètres de phase. L'Algorithme 6 détaille cette procédure. Cet algorithme prend en entrée la matrice des attaques, obtenue au préalable, grâce par exemple à la boîte à outils MATLAB Tempogram Toolbox Grosche et Müller (2011). L'initialisation pourrait par ailleurs être affinée en prenant pour $\psi_k(f)$ la valeur de la phase du mélange en une trame où l'on est sûr que seule la source k est active (la phase du mélange est alors égale à la phase de la source, prise comme référence dans cette trame).

Notons que l'ordre dans lequel sont effectuées les mises à jour des paramètres dans l'algorithme 6 (et de même plus loin dans l'algorithme 7) est arbitraire. Nous avons obtenu des résultats similaires en inversant cet ordre.

6.4.2 Contrainte relaxée

On considère à présent une contrainte relaxée, ce qui mène à la fonction de coût suivante :

$$\mathcal{C}_r = \sum_{f,m}\left|Y(f,m) - \sum_{k=1}^{K} V_k(f,t_m)e^{i\phi_k(f,t_m)}\right|^2 + \sigma \sum_{f,m,k} V_k(f,t_m)^2\left|e^{i\phi_k(f,t_m)} - e^{i\psi_k(f)}e^{i\lambda_k(m)f}\right|^2. \quad (6.25)$$

Le paramètre σ permet de donner plus ou moins d'importance à la contrainte issue du modèle de phase. Pour $\sigma = 0$, les phases d'attaque ne sont plus contraintes, et lorsque σ devient très grand, les sources ne somment plus au mélange : ainsi, il est nécessaire de choisir convenablement ce paramètre pour obtenir un compromis entre attache aux données et contrainte de phase (c'est l'objet de l'expérience conduite dans la section 6.5.1).

Estimation de $\psi_k(f)$

On applique la même méthode que dans le cas strict, qui consiste à isoler dans la fonction de coût les termes ne dépendant que de $\psi_k(f)$:

$$\mathcal{C}_r \overset{c}{=} \sum_m V_k(f,t_m)^2 \left| e^{i\phi_k(f,t_m)} - e^{i\psi_k(f)} e^{i\lambda_k(m)f} \right|^2 . \tag{6.26}$$

On a alors :

$$\begin{aligned}
\mathcal{C}_r &\overset{c}{=} \sum_m V_k(f,t_m)^2 \left(1 + 1 - 2\Re(e^{-i\phi_k(f,t_m)} e^{i\psi_k(f)} e^{i\lambda_k(m)f})\right) \\
\mathcal{C}_r &\overset{c}{=} -2\Re\left(\sum_m V_k(f,t_m)^2 e^{-i\phi_k(f,t_m)} e^{i\psi_k(f)} e^{i\lambda_k(m)f}\right) \\
\mathcal{C}_r &\overset{c}{=} -2\Re\left(e^{i\psi_k(f)} \sum_m V_k(f,t_m)^2 e^{-i\phi_k(f,t_m)} e^{i\lambda_k(m)f}\right).
\end{aligned}$$

On pose $z_r = \sum_m V_k(f,t_m)^2 e^{-i\phi_k(f,t_m)} e^{i\lambda_k(m)f}$ (on enlève les indices k et f pour plus de clarté), que l'on écrit en coordonnées polaires $z_r = |z_r| e^{i\xi_r}$. On a :

$$\mathcal{C}_r \overset{c}{=} -2|z_r| \cos(\psi_k(f) - \xi_r).$$

Minimiser $\mathcal{C}_r$ par rapport à $\psi_k(f)$ revient à maximiser le cosinus de $\psi_k(f) - \xi_r$, ce qui conduit à $\psi_k(f) = \xi_r = \angle(z_r)$, soit :

$$\psi_k(f) = \angle\left(\sum_m V_k(f,t_m)^2 e^{i\phi_k(f,t_m)} e^{-i\lambda_k(m)f}\right). \tag{6.27}$$

Estimation de $\lambda_k(m)$

Nous estimons $\lambda_k(m)$ en appliquant une méthode similaire au cas de la contrainte stricte. Cela requiert d'introduire un nouveau paramètre γ (similaire au β de la partie précédente) défini comme suit :

$$\gamma_k(f,m) = V_k(f,t_m) e^{i\phi_k(f,t_m)} e^{-i\psi_k(f)}. \tag{6.28}$$

On pose :

$$\underline{\gamma}_k(m) = [\gamma_k(0,m), ..., \gamma_k(F-1,m)]^T. \tag{6.29}$$

L'adaptation de la méthode ESPRIT comme utilisée précédemment mène à l'estimation :

$$\lambda_k(m) = \angle\left(\underline{\gamma}_k(m)_{\downarrow}^H \underline{\gamma}_k(m)_{\uparrow}\right). \tag{6.30}$$

Estimation de $\phi_k(f, t_m)$

Enfin, il faut estimer, en plus des paramètres du modèle de phase, les termes $\phi_k(f, t_m)$. La méthode est la même que pour les termes $\psi_k(f)$: on réécrit la fonction de coût en isolant les termes dépendant uniquement de $\phi_k(f, t_m)$. On note ici, de façon similaire à la partie précédente :

$$B_k(f, m) = Y(f, m) - \sum_{l \neq k} V_l(f, t_m) e^{i\phi_l(f, t_m)}, \tag{6.31}$$

et on écrit donc la fonction de coût de la façon suivante :

$$\mathcal{C}_r \overset{c}{=} \left| B_k(f, t_m) - V_k(f, t_m) e^{i\phi_k(f, t_m)} \right|^2 + \sigma V_k(f, t_m)^2 \left| e^{i\phi_k(f, t_m)} - e^{i\psi_k(f)} e^{i\lambda_k(m) f} \right|^2, \tag{6.32}$$

que l'on réécrit :

$$\begin{aligned} \mathcal{C}_r \overset{c}{=} |B_k(f, t_m)|^2 + |V_k(f, t_m)|^2 - 2\Re(\overline{B}_k(f, t_m) V_k(f, t_m) e^{i\phi_k(f, t_m)}) \\ + \sigma V_k(f, t_m)^2 \left(1 + 1 - 2\Re(e^{i\phi_k(f, t_m)} e^{-i\psi_k(f)} e^{-i\lambda_k(m) f}) \right). \end{aligned} \tag{6.33}$$

En retirant les termes ne dépendant pas de $\phi_k(f, t_m)$, on peut simplifier cette expression :

$$\begin{aligned} \mathcal{C}_r &\overset{c}{=} -2V_k(f, t_m) \Re \left(\overline{B}_k(f, t_m) e^{i\phi_k(f, t_m)} + \sigma V_k(f, t_m) e^{i\phi_k(f, t_m)} e^{-i\psi_k(f)} e^{-i\lambda_k(m) f} \right) \\ &\overset{c}{=} -2V_k(f, t_m) \Re \left(e^{i\phi_k(f, t_m)} (\overline{B}_k(f, t_m) + \sigma V_k(f, t_m) e^{-i\psi_k(f)} e^{-i\lambda_k(m) f}) \right). \end{aligned}$$

On pose $z_\phi = B_k(f, t_m) + \sigma V_k(f, t_m) e^{i\psi_k(f)} e^{i\lambda_k(m) f}$ (on enlève les indices k, f et m pour plus de clarté), que l'on réécrit en coordonnées polaires $z_\phi = |z_\phi| e^{i\xi_\phi}$, et on a :

$$\mathcal{C}_r \overset{c}{=} -2V_k(f, t_m) |z_\phi| \cos(\phi_k(f, t_m) - \xi_\phi). \tag{6.34}$$

En appliquant la même technique que précédemment, on a finalement :

$$\phi_k(f, t_m) = \angle \left(B_k(f, t_m) + \sigma V_k(f, t_m) e^{i\psi_k(f)} e^{i\lambda_k(m) f} \right). \tag{6.35}$$

Algorithme d'estimation relaxé

L'Algorithme 7 détaille la procédure d'estimation itérative des paramètres de phase obtenue grâce aux équations (6.27), (6.30) et (6.35) dans le cas relaxé.

Algorithme 7 Estimation des paramètres de phase sous contrainte relaxée

Entrées $Y \in \mathbb{C}^{F\times M}$, $V_k \in \mathbb{R}_+^{F\times M}$, $\psi_k^{ini} \in \mathbb{R}^{F\times 1}$, $\lambda_k^{ini} \in \mathbb{R}^{1\times M}$, $\phi_k^{ini} \in \mathbb{R}^{F\times T}$, $\sigma \in \mathbb{R}_+$.
Initialisation :
$\phi = \phi_{ini}$, $\psi = \psi_{ini}$, $\lambda = \lambda_{ini}$.
$\hat{Y}_k(f,m) = V_k(f,t_m)e^{i\phi_k(f,t_m)}$.
$\hat{Y}(f,m) = \sum_{k=1}^{K} \hat{Y}_k(f,m)$.
$B_k(f,m) = Y(f,m) - \hat{Y}(f,m) + \hat{Y}_k(f,m)$.
tant que critère non atteint **faire**
  pour $k = 1$ à K, $f = 0$ à $F-1$ et $m = 0$ à $M-1$ **faire**
    Calculer ϕ
    $\phi_k(f,t_m) = \angle(B_k(f,m) + \sigma V_k(f,t_m)e^{i\psi_k(f)}e^{i\lambda_k(m)f})$.
    Calculer ψ
    $\psi_k(f) = \angle\left(\sum_m V_k(f,t_m)^2 e^{i\phi_k(f,t_m)}e^{-i\lambda_k(m)f}\right)$.
    Calculer $\underline{\gamma}$
    $\gamma_k(f,m) = V_k(f,t_m)e^{i\phi_k(f,t_m)}e^{-i\psi_k(f)}$,
    $\underline{\gamma}_k(m) = [\gamma_k(0,m), ..., \gamma_k(F-1,m)]^T$.
    Calculer λ
    $\lambda_k(m) = \angle\left(\underline{\gamma}_k(m)_{\downarrow}^H \underline{\gamma}_k(m)_{\uparrow}\right)$.
    Calculer $\hat{Y}$
    $\hat{Y}_k(f,m) = V_k(f,t_m)e^{i\phi_k(f,t_m)}$,
    $\hat{Y}(f,m) = \sum_{k=1}^{K} \hat{Y}_k(f,m)$.
    Calculer B
    $B_k(f,m) = Y(f,m) - \hat{Y}(f,m) + \hat{Y}_k(f,m)$.
  fin pour
fin tant que
Sorties : $\hat{Y}, \hat{Y}_k, \psi, \lambda, \phi$.

6.5 Résultats expérimentaux

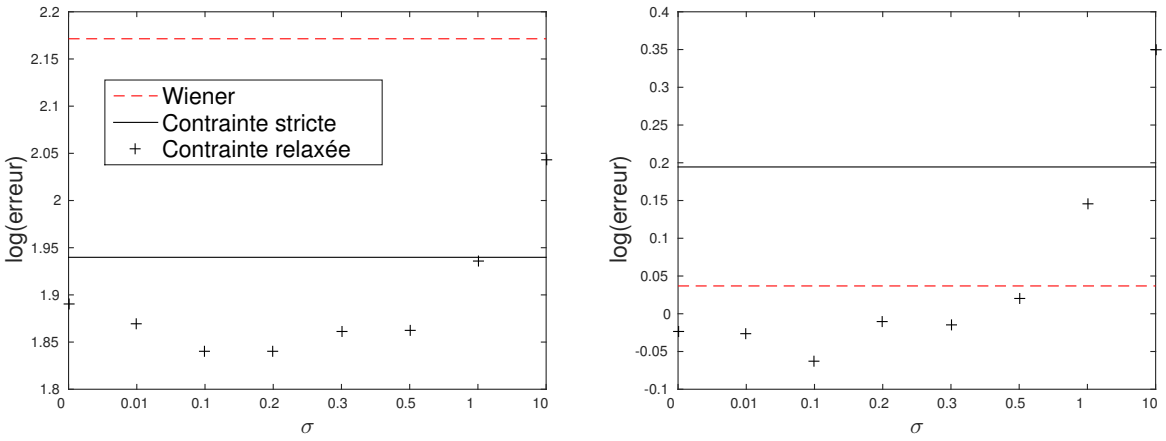
Dans cette partie, nous présentons des expériences menées pour évaluer le potentiel de notre méthode. Les signaux sont échantillonnés à $F_s = 11025$ Hz. La TFCT est calculée avec une fenêtre de Hann de longueur 512 échantillons et 75 % de recouvrement. Les phases des sources sont calculées par 100 itérations de nos algorithmes (strict et relaxé), la performance n'étant pas améliorée au-delà.

La boîte à outils MATLAB Tempogram Toolbox est utilisée pour estimer les trames d'attaque. Nous extrayons ensuite la matrice d'attaque Y à partir de la matrice complète X. Nous utilisons les SDR, SIR et SAR comme mesure de la qualité de la séparation de sources.

6.5.1 Influence du paramètre σ

Dans cette expérience, on cherche à estimer les paramètres de phases d'attaque sur des mélanges de deux sources ($K = 2$) dans lesquels on observe successivement chaque source seule, puis la superposition des deux. Nous testons les algorithmes tout d'abord sur des données synthétiques qui sont fabriquées en suivant le modèle (6.9), puis sur des données plus réalistes, des mélanges de notes de piano.

On teste les algorithmes avec contrainte stricte et avec contrainte relaxée pour différentes valeurs du paramètre de relaxation σ. Les valeurs des amplitudes des sources V_k sont supposées



(a) Données construites avec le modèle (b) Notes de piano

FIGURE 6.3 – Erreur (6.36) entre données et modèle estimé au niveau des attaques.

connues afin de s'intéresser spécifiquement à la reconstruction de phase. Ces algorithmes sont comparés à l'approche traditionnelle du filtrage de Wiener. Pour toutes ces méthodes, nous calculons l'erreur d'estimation moyennée sur les sources :

$$\frac{1}{K}\sum_{k=1}^{K}||Y_k - \hat{Y}_k||_2, \tag{6.36}$$

où $\hat{Y}_k$ est l'estimée de la k-ième source Y_k dans les trames d'attaque. Les résultats sont moyennés sur 30 signaux pour chaque jeu de données, et présentés sur la figure 6.3.

Lorsque les données suivent exactement notre modèle, l'algorithme avec contrainte stricte conduit à un meilleur résultat que le filtrage de Wiener. Pour des données réelles, qui ne suivent donc plus rigoureusement le modèle, l'algorithme d'estimation strict ne donne pas de meilleurs résultats que le filtrage de Wiener, certainement en raison d'une contrainte trop forte qui ne correspond plus à la réalité des données. Néanmoins, lorsque l'on applique l'algorithme relaxé, on voit que pour certaines valeurs de σ, les résultats sont meilleurs qu'avec le filtrage de Wiener. Cela traduit le potentiel d'un tel modèle de phase, à condition qu'il soit utilisé de façon relaxée pour prendre en compte les écarts au modèle.

6.5.2 Performance du modèle de phase

Nous proposons d'appliquer cette technique de reconstruction de phase à la séparation de sources. Nous considérons plusieurs jeux de données :

A : 30 mélanges de deux sources, composées de sinusoïdes amorties synthétiques. Les sources ne se recouvrent pas dans le domaine TF ;

B : 30 mélanges de deux sources, composées de sinusoïdes amorties synthétiques. Les sources se recouvrent dans le domaine TF ;

C : 30 mélanges de deux notes de piano tirées de la base MAPS ;

D : Un extrait MIDI de 1.57 secondes. Il est composé de plusieurs occurrences de trois notes de basse, de trois notes de piano et d'un accord de guitare.

Pour les mélanges dans les jeux de données A, B et C, chaque source est successivement observée seule, puis les deux sources sont activées simultanément. On applique l'algorithme

Données	SDR	SIR	SAR
A	0.16	0.3	0.14
B	0.002	0.003	0.001
C	0.05	0.1	0.02
D	−0.09	0.15	−0.3

TABLEAU 6.4 – Variation de score (en dB) entre méthodes de reconstruction de phase par modèle d'attaque et par filtrage de Wiener.

Données	Méthode	SDR	SIR	SAR
A	Wiener	**34.8**	**44.9**	**35.3**
	RePU	6.9	28.9	6.9
B	Wiener	**11.4**	15.7	**13.6**
	RePU	8.7	**28.4**	8.8
C	Wiener	**12.6**	**17.5**	**14.6**
	RePU	2.0	15.5	2.6
D	Wiener	**15.9**	**18.6**	**19.7**
	RePU	5.5	15.2	6.4

TABLEAU 6.5 – Performance de la séparation de source (SDR, SIR et SAR en dB) pour le filtrage de Wiener et la méthode RePU.

relaxé avec $\sigma = 0.1$ pour restaurer les phases des attaques. Alternativement, on peut utiliser le filtrage de Wiener pour la restauration des phases d'attaque. Par suite, les phases des partiels des composantes sont estimées par filtrage de Wiener, ce qui permet de comparer les deux méthodes utilisées pour restaurer les phases d'attaque en employant la même technique pour reconstruire les phases des autres trames.

Nous présentons dans le tableau 6.4 la valeur $\Delta = SR(\text{Méthode proposée}) - SR(\text{Wiener})$, ou SR représente le score considéré (SDR,SIR ou SAR). Ainsi, une valeur positive de cet écart traduit une amélioration des performances par notre méthode par rapport au filtrage de Wiener pour la restauration des phases d'attaque.

On constate que le modèle de phase améliore légèrement les résultats par rapport au filtrage de Wiener sur ces jeux de données. Cette amélioration est faible en valeur absolue et en erreur relative puisque les niveaux de performance sont déjà très élevés (entre 20 et 30 dB). Les hauts niveaux de performance sont expliqués par la simplicité des données, et la relative modération de l'augmentation due au fait que les signaux reconstruits ne diffèrent qu'au niveau des attaques. Ainsi, nous allons par la suite utiliser deux approches différentes pour la reconstruction des phases de partiels.

6.5.3 Modèle de phase d'attaque et déroulé linéaire

L'expérience précédente montre l'intérêt d'utiliser notre modèle de phase d'attaque pour l'estimation des composantes complexes dans le mélange. Néanmoins, les phases des partiels étaient estimées par filtrage de Wiener. À présent, on estime les phases des partiels en appliquant l'algorithme de déroulé linéaire de phases introduit dans le chapitre 4. Cette méthode est désignée par l'acronyme **RePU** (pour ***Re**peating **P**hase with **U**nwrapping*). Nous la comparons au filtrage de Wiener intégral (attaques + partiels) et présentons les résultats dans le tableau 6.5.

Ces résultats montrent que le filtrage de Wiener fournit globalement de meilleurs résultats que la combinaison du modèle de phases d'attaque et du déroulé linéaire.

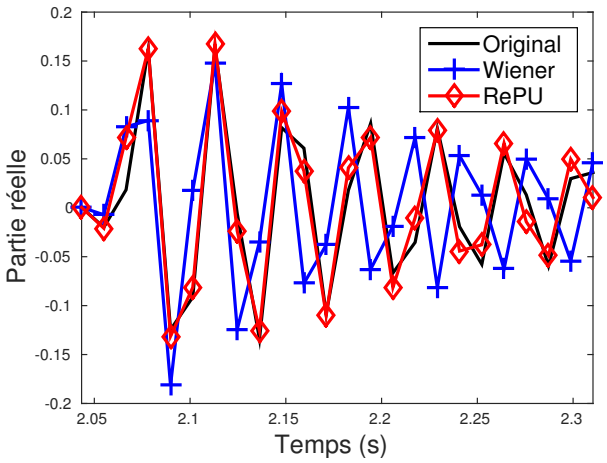


FIGURE 6.4 – Partie réelle de la source correspondant à la note G4 dans la bande de fréquences autour de 796 Hz, sur une fenêtre de temps ou deux sources se recouvrent.

Une évaluation perceptive informelle de notre part sur les données audio obtenues suggère cependant que les indicateurs utilisée dans ces tests (les SDR, SIR et SAR) ne sont pas aptes à rendre compte de certaines propriétés des signaux séparés. Par exemple, le phénomène de battements observé dans le jeu de données B (ainsi que sur d'autres tests effectués sur notes de piano) est complètement supprimé lors de la séparation de sources avec notre méthode, alors que le filtrage de Wiener ne permet pas de s'en affranchir. On illustre ce phénomène par la figure 6.4 : considérant un mélange de deux notes de piano (C4 et G4), on représente sur cette figure la partie réelle de la deuxième source (note G4) sur la fenêtre de temps et dans une bande de fréquences où l'on observe un recouvrement. La méthode RePU approche mieux le signal original que le filtrage de Wiener, qui conduit à des interférences entre sources.

Bien qu'une perte de définition des transitoires d'attaque soit également à déplorer avec notre technique, ces critères perceptifs ne sont donc pas nécessairement capturés par ces indicateurs.

On constate donc que similairement aux expériences menées dans le chapitre 5, lorsque le déroulé est appliqué isolément sur chaque source (c'est-à-dire sans tenir compte de la phase du mélange), les performances sont moindres qu'en utilisant le filtrage de Wiener. La baisse de performance que nous avons constatée dans cette expérience provient donc probablement de l'utilisation isolée du déroulé linéaire, qui n'exploite pas la phase du mélange au niveau des partiels.

6.5.4 Prise en compte de la phase du mélange pour le déroulé

Dans cette expérience, nous cherchons à comparer uniquement l'impact de la méthode choisie pour la reconstruction des phases d'attaque. Les phases des partiels sont reconstruites en utilisant l'algorithme 3 introduit au chapitre 5. Trois choix sont possibles pour estimer les composantes dans les trames d'attaque :

— Utiliser la phase du **Mélange** ;

— Le **Modèle** de phase introduit ici ;

— Une approche **Oracle** : les phases d'attaque sont supposées connues.

Les tests sont conduits sur des mélanges de sinusoïdes synthétiques avec recouvrement TF et des mélanges de notes de piano, et les résultats sont présentés dans le tableau 6.6.

Données	Méthode	SDR	SIR	SAR
Données synthétiques	**Mélange**	46.0	55.2	46.6
	Modèle	46.7	56.1	47.3
	Oracle	49.5	58.5	50.1
Notes de piano	**Mélange**	20.4	27.1	21.5
	Modèle	21.0	27.9	22.1
	Oracle	22.6	29.8	23.6

TABLEAU 6.6 – Performance de la séparation de sources (SDR, SIR et SAR en dB) pour diverses méthodes de reconstruction des phases d'attaque.

Les interprétations que l'on peut faire à partir de ce tableau sont les mêmes pour les deux jeux de données. Une comparaison entre **Mélange** et **Oracle** montre qu'en utilisant la technique de déroulé linéaire, il existe une certaine marge de progression possible entre une donner la phase du mélange aux sources dans les trames d'attaque et une approche Oracle (phases d'attaque connues). Cette marge de progression est d'environ 3 dB selon l'indicateur et le jeu de données considérés. Notre méthode conduit à un gain moyen de 0.7 dB par rapport à **Mélange**.

La méthode proposée ici combine modèle de phases d'attaque et déroulé linéaire pour les partiels. L'impact du déroulé linéaire est plus significatif que celui du modèle de phase dans les trames d'attaque, mais ceci est expliqué par le fait que les trames concernées par cette méthode (qui ne sont pas des trames d'attaque) représentent une proportion des données très importante : sur un mélange de notes de piano tel qu'utilisé ici, il y a 3 trames d'attaque pour un total de 262 trames. Le déroulé linéaire concerne donc presque 99 % des trames, ce qui explique que son impact sur le résultat final soit plus important que celui de la technique utilisée pour la reconstruction des phases d'attaque. Comme le déroulé linéaire dépend cependant de son initialisation (effectuée au niveau des trames d'attaque), la méthode employée pour reconstruire les phases d'attaque influe à la fois sur la qualité des attaques reconstruites mais également sur celle des partiels via le déroulé.

6.6 Conclusion

Nous avons introduit un modèle de phase basé sur la répétition d'évènements audio qui exploite cette propriété au sein des trames d'attaque. Ce modèle a montré expérimentalement sa pertinence sur plusieurs types de données. Combiné au déroulé linéaire dans le cas aveugle, on constate que cette méthode permet une meilleure reconstruction du signal que l'utilisation d'une phase aléatoire. Nous l'avons également intégré à un modèle de mélange de sources dans les trames d'attaque pour la séparation de sources. De futures expériences pourraient être conduites sur des signaux de musique plus réalistes et complexes. Néanmoins celles-ci sont compliquées à réaliser car il faudrait pour cela disposer des amplitudes de chaque atome spectral (i.e. de chaque note), ce qui peut être compliqué à obtenir pour un scénario Oracle.

Au demeurant assez simple, ce modèle pourrait être affiné, en prenant en compte le fait qu'une partie du signal est tronquée d'une trame d'attaque à l'autre, et que la forme de la fenêtre d'analyse et son positionnement pourraient également être utilisés. Les premiers calculs que nous avons effectués sur ce sujet n'ont cependant pas abouti à la formulation d'un modèle analytiquement simple, aussi nous avons préféré inclure ce premier modèle dans le cadre de la séparation de sources afin d'évaluer son potentiel. Les résultats de séparation étant mitigés du fait de la dépendance vis-à-vis du paramètre σ, il pourrait être utile de calculer automatiquement un paramètre de relaxation optimal. Par ailleurs, outre la modélisation des

répétitions de phase, il pourrait être intéressant de modéliser les phases au sein des trames d'attaque afin d'exploiter les corrélations entre canaux fréquentiels Kirchhoff et al. (2014); Badeau et Plumbley (2014) et donc de structurer la phase de référence ψ.

Les amplitudes des sources étaient supposées connues, ce qui n'est pas le cas en pratique. Le chapitre 7 introduit cette contrainte de répétition dans un modèle de NMF complexe pour l'estimation conjointe des amplitudes et des phases des différentes sources.

Chapitre 7

NMF complexe à phase contrainte

Sommaire

Les modèles de phase étudiés jusqu'à présent ont été utilisés dans un cadre de séparation de sources, mais les amplitudes étaient supposées connues (ou estimées préalablement à la reconstruction de phase). Nous nous intéressons dans ce chapitre à un modèle inspiré de la NMF, de type NMF complexe (CNMF), qui factorise les spectrogrammes d'amplitude tout en modélisant les phases. Comme nous l'avons vu dans le chapitre 3, les représentations CNMF sont prometteuses pour la séparation de sources, mais requièrent que la phase soit contrainte pour conduire à une séparation de qualité. Nous proposons donc des contraintes de phase issues des modèles étudiés dans ce manuscrit : le déroulé linéaire introduit au chapitre 4, et un modèle de phase dans les trames d'attaque (*cf.* chapitre 6).

Les résultats de ce travail ont fait l'objet d'une publication à la conférence ICASSP 2016 Magron et al. (2016b).

La section 7.1 introduit le modèle de CNMF à phase contrainte. La section 7.2 présente les algorithmes par lesquels le modèle est estimé. La section 7.3 consiste en une validation expérimentale de cette technique appliquée à la séparation de sources, et nous livrons nos principales conclusions dans la section 7.4.

7.1 Modèle de NMF complexe à phase contrainte

7.1.1 Approche intuitive

Le modèle de NMF complexe que nous avons présenté dans le chapitre 2 consiste à approcher une matrice de données complexes X (une TFCT en général) par le modèle $\hat{X}$ tel que pour tout canal fréquentiel f et trame temporelle t :

$$\hat{X}(f,t) = \sum_{k=1}^{K} W(f,k)H(k,t)e^{i\phi_k(f,t)}, \tag{7.1}$$

avec $W \in \mathbb{R}_+^{F\times K}$ et $H \in \mathbb{R}_+^{K\times T}$. L'estimation de ce modèle Kameoka et al. (2009) conduit à minimiser une fonction de coût qui s'exprime comme la somme de deux termes : la distance euclidienne entre X et $\hat{X}$, et un terme de parcimonie $\mathcal{C}_s(H) = 2\sum_{k,t} H(k,t)^p$ $(p \in]0,2[)$ multiplié par un poids σ_s.

Intuitivement, on peut considérer le problème d'estimation d'un modèle de NMF complexe à phase contrainte comme un problème de minimisation d'une fonction de coût $\mathcal{C}$ qui serait la somme de divers termes, qui traduisent les contraintes intégrées au modèle :

- Un terme de distance euclidienne D entre le modèle et les données, qui évalue la précision de la reconstruction (terme d'attache aux données) ;
- Un terme de parcimonie $\mathcal{C}_s$ qui force cette propriété ;
- Un terme $\mathcal{C}_u$ qui tienne compte du déroulé linéaire de phases ;
- Un terme $\mathcal{C}_r$ qui tienne compte d'un modèle de phases d'attaque.

On reprend, pour D et $\mathcal{C}_s$, les expressions obtenues dans le modèle originel Kameoka et al. (2009), et on intègre les nouvelles contraintes décrites ci-après.

Déroulé linéaire Le déroulé linéaire de phase introduit dans le chapitre 4 modélise la phase ϕ_k d'une source de la façon suivante :

$$\phi_k(f,t) = \phi_k(f,t-1) + 2\pi S\nu_k(f), \tag{7.2}$$

où S est le décalage temporel entre deux trames successives (en échantillons) et $\nu_k(f)$ est la fréquence réduite de la source k dans le canal f.

Afin d'injecter cette propriété de déroulé linéaire sous forme de contrainte dans notre problème d'estimation du modèle (7.1), on considère la fonction de coût suivante :

$$\mathcal{C}_u(\phi) = \sum_{f,k} \sum_{t \notin \Omega_k} |X(f,t)|^2 |e^{i\phi_k(f,t)} - e^{i\phi_k(f,t-1)} e^{2i\pi S \nu_k(f)}|^2, \tag{7.3}$$

où Ω_k désigne l'ensemble des trames d'attaque de la source k. En effet, le modèle de déroulé linéaire n'est valable qu'en dehors des trames d'attaque. Ce type d'approche a déjà été utilisé dans la littérature, notamment dans Bronson et Depalle (2014) et plus récemment dans Rodriguez-Serrano et al. (2016). Contrairement à ces approches, la notre est adaptée à la séparation aveugle car elle ne suppose pas connues les nombres d'harmoniques et les fréquences fondamentales, et ne requiert pas d'information externe (partition).

En posant $\forall (k,f,t)$, $u_k(f,t) = |e^{i\phi_k(f,t)} - e^{i\phi_k(f,t-1)} e^{2i\pi S \nu_k(f)}|^2$, (7.3) se réécrit :

$$\mathcal{C}_u(\phi) = \sum_{f,k} \sum_{t \notin \Omega_k} |X(f,t)|^2 u_k(f,t). \tag{7.4}$$

Nous proposons d'estimer les fréquences réduites par QIFFT sur les colonnes de W, et reprenons l'expression de la région d'influence utilisée dans le chapitre 4 pour associer à chaque canal fréquentiel f la fréquence réduite correspondante $\nu_k(f)$.

Remarque : Dans nos précédents travaux, la fréquence instantanée dépendait de la trame temporelle considérée t, afin de prendre en compte les variations de celle-ci, et donc les signaux non-stationnaires. Ici, le modèle NMF utilisé suppose que les fréquences sont fixes au cours du temps puisque le dictionnaire W d'atomes spectraux ne dépend pas du temps (les variations de fréquence fondamentale ne sont pas prises en compte dans ce modèle). Ainsi, nous estimerons les fréquences ν_k à partir de la matrice W, ces valeurs étant ensuite utilisées pour la totalité du déroulé. Notons cependant qu'un modèle plus fin prenant en compte les variations de fréquence pourra être ultérieurement envisagé, basé sur les travaux de Hennequin et al. (2011a).

Phases dans les trames d'attaque Afin de tenir compte du caractère répétitif des atomes temporels, nous proposons d'utiliser le modèle introduit dans le chapitre 6. La phase d'une source ϕ_k dans une trame d'attaque $t \in \Omega_k$ est modélisée par une phase de référence ψ_k à laquelle est ajoutée un délai linéaire en fréquence de pente λ_k :

$$\phi_k(f,t) = \psi_k(f) + \lambda_k(t) f. \tag{7.5}$$

La notation $t \in \Omega_k$ est préférée à la notation t_m du chapitre 6 pour alléger les notations. On a donc le critère suivant :

$$\mathcal{C}_r(\phi,\psi,\lambda) = \sum_{f,k} \sum_{t \in \Omega_k} |X(f,t)|^2 |e^{i\phi_k(f,t)} - e^{i\psi_k(f)} e^{i\lambda_k(t) f}|^2. \tag{7.6}$$

soit, en posant $r_k(f,t) = |e^{i\phi_k(f,t)} - e^{i\psi_k(f)} e^{i\lambda_k(t) f}|^2$,

$$\mathcal{C}_r(\phi,\psi,\lambda) = \sum_{f,k} \sum_{t \in \Omega_k} |X(f,t)|^2 r_k(f,t). \tag{7.7}$$

Fonction de coût globale En ajoutant les termes (7.4) et (7.7) à la distance euclidienne entre les données et le modèle ainsi qu'à la contrainte de parcimonie, on obtient alors la fonction de coût globale suivante :

$$\mathcal{C}(\theta) = D(X, \hat{X}) + \sigma_u \mathcal{C}_u(\phi), + \sigma_r \mathcal{C}_r(\phi,\psi,\lambda) + \sigma_s \mathcal{C}_s(H). \tag{7.8}$$

Dans l'article Magron et al. (2016b), nous avons formulé la fonction de coût de cette façon, mais on peut également utiliser un cadre probabiliste pour y parvenir. C'est ce que nous détaillons dans la section suivante.

7.1.2 Modélisation probabiliste

Estimateurs ML et MAP

On considère que les données X sont égales au modèle $\hat{X}$ défini par (7.1) auquel est ajouté un terme d'erreur, modélisé par un bruit blanc gaussien : $\forall(f,t)$,

$$X(f,t) = \hat{X}(f,t) + \epsilon(f,t), \tag{7.9}$$

où les $\epsilon(f,t)$ sont indépendants et de même loi $\mathcal{N}(0,\tilde{\sigma}^2)$. On a alors $X(f,t) \sim \mathcal{N}(\hat{X}(f,t),\tilde{\sigma}^2)$. En notant $\theta = \{W,H,\phi\}$ l'ensemble des paramètres du modèle, la log-vraisemblance est donc donnée par :

$$\begin{aligned}
L(\theta) &= \log(p_{X|\theta}(X)) = \sum_{f,t} \log(p_{X(f,t)|\theta}(X(f,t))) \\
&= \sum_{f,t} -\log(\pi\tilde{\sigma}^2) - \frac{|X(f,t) - \sum_k \hat{X}_k(f,t)|^2}{\tilde{\sigma}^2} \\
&\overset{c}{=} -\sum_{f,t} \frac{|X(f,t) - \sum_k W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2}{\tilde{\sigma}^2} \\
&\overset{c}{=} -\frac{1}{\tilde{\sigma}^2}\sum_{f,t} |X(f,t) - \sum_k W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2 \\
&\overset{c}{=} -\frac{1}{\tilde{\sigma}^2} D(X,\hat{X}).
\end{aligned}$$

On constate donc que maximiser la log-vraisemblance des données (approche ML) revient à minimiser D. Néanmoins, on ne s'intéresse pas ici à la méthode ML puisque celle-ci ne permet pas d'injecter des à priori sur les paramètres. Pour ce faire, on adopte plutôt une approche MAP. Comme nous l'avons déjà rappelé dans le chapitre 2, section 2.2.4, l'approche MAP consiste à maximiser la loi à postériori, ce qui revient à maximiser :

$$\mathcal{C}_{MAP}(\theta) = L(\theta) + \log(p_\theta), \tag{7.10}$$

où p_θ désigne la loi à priori sur les variables θ. Nous allons donc incorporer la parcimonie via un à priori sur H et les contraintes de phase via un à priori sur ϕ.

Parcimonie

Pour introduire une contrainte de parcimonie, on modélise chaque $H(k,t)$ comme une variable aléatoire suivant une loi normale généralisée Kameoka et al. (2009) :

$$p_{H(k,t)}(H(k,t)) = \frac{1}{2\Gamma(1+\frac{1}{p})b} e^{-\frac{|H(k,t)|^p}{b^p}}, \tag{7.11}$$

où p et b sont deux paramètres qui déterminent la forme de la distribution, et Γ désigne la fonction Gamma d'Euleur Artin (2015). Ainsi, en supposant toutes les sources et trames

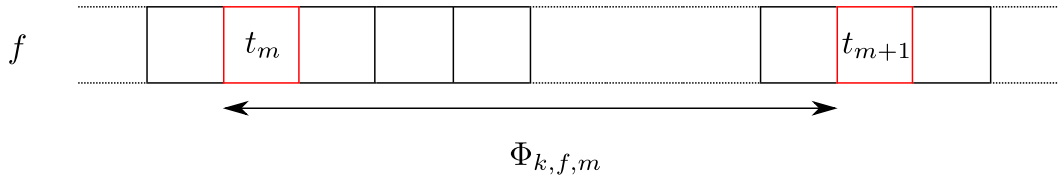


FIGURE 7.1 – Découpage de l'ensemble des trames selon les trames d'attaque.

temporelles indépendantes, on a :

$$\begin{aligned}\log(p_H) &= \sum_{k,t} \log(p_{H(k,t)}(H(k,t))) \\ &\stackrel{c}{=} -\frac{1}{b^p} \sum_{k,t} |H(k,t)|^p.\end{aligned}$$

En posant $\sigma_s = \dfrac{\tilde{\sigma}^2}{2b^p}$ et $\mathcal{C}_s(H) = 2\sum_{k,t} |H(k,t)|^p$, on a :

$$\log(p_H) = -\frac{\sigma_s}{\tilde{\sigma}^2}\mathcal{C}_s(H), \tag{7.12}$$

Remarque : Sans à priori sur la phase, maximiser la distribution à postériori revient à minimiser la fonction de coût $D(X, \hat{X}) + \sigma_s \mathcal{C}_s(H)$: on retrouve exactement le modèle de CNMF complexe originel KAMEOKA et al. (2009).

Contraintes de phase

Nous enrichissons le modèle précédent avec un terme d'à priori sur la phase ϕ. On commence par décomposer ces termes, dans chaque bande de fréquences et pour chaque source, sous hypothèse d'indépendance des sources et canaux fréquentiels :

$$p_\Phi = \prod_{f,k} p_{\Phi_{k,f}}. \tag{7.13}$$

On introduit les trames d'attaque t_m (nous revenons temporairement à cette notation, similaire à celle du chapitre 6, par souci de clarté) pour la source k, avec $m \in [\![0, M-1]\!]$ où M désigne le nombre de trames d'attaque pour la source k [1]. On découpe les composantes $\Phi_{k,f}$ selon les différentes attaques, comme c'est illustré sur la figure 7.1 :

$$p_{\Phi_{k,f}} = \prod_{m=0}^{M-1} p_{\Phi_{k,f,m}}. \tag{7.14}$$

On utilise une structure d'à priori en chaînes de Markov pour tenir compte de la dépendance des phases entre trames successives, comme cela est proposé dans BERTIN et al. (2010) pour modéliser la continuité temporelle des activations dans un modèle NMF :

$$p_{\Phi_{k,f,m}} = p(\phi_k(f,t_m)) \prod_{t=t_m+1}^{t_{m+1}-1} p(\phi_k(f,t)|\phi_k(f,t-1)). \tag{7.15}$$

1. En toute rigueur, M et t_m devraient dépendre de k, mais on retire ces indices pour plus de lisibilité. Nous revenons temporairement à cette notation car elle rend le découpage des trames plus simple.

On considère que $\phi_k(f,t)|\phi_k(f,t-1)$ suit une loi de Von Mises Mardia et Zemroch (1975) de mode $\phi_k(f,t-1)+2\pi S\nu_k(f)$ et de paramètre de concentration $\kappa_k(f,t)$. Nous détaillons cette loi et ses paramètres dans le chapitre 8, section 8.1.1. Cette distribution modélise des variables 2π-périodiques, ce qui est adapté à la phase. Sa densité de probabilité est :

$$p(\phi_k(f,t)|\phi_k(f,t-1)) = \frac{e^{\kappa_k(f,t)\cos(\phi_k(f,t)-\phi_k(f,t-1)-2\pi S\nu_k(f))}}{2\pi I_0(\kappa_k(f,t))}, \tag{7.16}$$

où I_0 désigne la fonction de Bessel modifiée de première espèce d'ordre 0. Ainsi :

$$\log(p_\Phi) \stackrel{c}{=} \sum_{k,f}\sum_{m}\log(p(\phi_k(f,t_m)))+\sum_{k,f}\sum_{m}\sum_{t=t_m+1}^{t_{m+1}-1}\kappa_k(f,t)\cos(\phi_k(f,t)-\phi_k(f,t-1)-2\pi S\nu_k(f)), \tag{7.17}$$

ce que l'on peut réécrire en utilisant la notation Ω_k :

$$\log(p_\Phi) \stackrel{c}{=} \sum_{k,f}\sum_{t\in\Omega_k}\log(p(\phi_k(f,t))) + \sum_{k,f}\sum_{t\notin\Omega_k}\kappa_k(f,t)\cos(\phi_k(f,t)-\phi_k(f,t-1)-2\pi S\nu_k(f)). \tag{7.18}$$

Par ailleurs, on considère que les phases dans les trames d'attaque sont distribuées selon une loi de Von Mises de localisation $\psi_k(f)+\lambda_k(t)f$, afin d'introduire le modèle de phase dans les trames d'attaque :

$$\forall t\in\Omega_k,\ p(\phi_k(f,t)) = \frac{e^{\kappa_k(f,t)\cos(\phi_k(f,t)-\psi_k(f)-\lambda_k(t)f)}}{2\pi I_0(\kappa_k(f,t))}. \tag{7.19}$$

Ainsi :

$$\begin{aligned}\log(p_\Phi) \stackrel{c}{=} &\sum_{k,f}\sum_{t\in\Omega_k}\kappa_k(f,t)\cos(\phi_k(f,t)-\psi_k(f)-\lambda_k(t)f)\\ &+\sum_{k,f}\sum_{t\notin\Omega_k}\kappa_k(f,t)\cos(\phi_k(f,t)-\phi_k(f,t-1)-2\pi S\nu_k(f)).\end{aligned} \tag{7.20}$$

Remarque : On aurait pu aboutir au même résultat en structurant les phases dans les trames d'attaque avec un modèle de chaîne de Markov. Les termes ψ et λ auraient alors eu une signification différente : ψ désignerait la phase dans la première trame d'attaque, et λ le délai entre deux trames d'attaque consécutives. Dans le modèle tel que présenté ici, les λ représentent le délai par rapport à une phase de référence ψ. Ces deux modèles sont équivalents à condition de fixer $\lambda=0$ pour la première trame d'attaque.

Afin de réécrire cette expression, on utilise le résultat suivant : $\forall(a,b)\in\mathbb{R}^2$,

$$|e^{ia}-e^{ib}|^2 = |e^{ia}|^2+|e^{ib}|^2-2\Re(e^{ia}e^{-ib}) = 2-2\cos(a-b). \tag{7.21}$$

Ainsi :

$$\cos(\phi_k(f,t)-\psi_k(f)-\lambda_k(t)f) \stackrel{c}{=} -\frac{1}{2}|e^{i\phi_k(f,t)}-e^{i\psi_k(f)}e^{i\lambda_k(t)f}|^2, \tag{7.22}$$

et :

$$\cos(\phi_k(f,t)-\phi_k(f,t-1)-2\pi S\nu_k(f)) \stackrel{c}{=} -\frac{1}{2}|e^{i\phi_k(f,t)}-e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}|^2. \tag{7.23}$$

On pose $\kappa_k(f,t) = 2\frac{\sigma_r|X(f,t)|^2}{\tilde{\sigma}^2}$ dans les trames d'attaque et $\kappa_k(f,t) = 2\frac{\sigma_u|X(f,t)|^2}{\tilde{\sigma}^2}$ dans les autres trames[2], avec $\sigma_u > 0$ et $\sigma_r > 0$. Ainsi,

$$\log(p_\Phi) \stackrel{c}{=} \sum_{k,f}\sum_{t\in\Omega_k} \frac{\sigma_r}{\tilde{\sigma}^2}|X(f,t)|^2|e^{i\phi_k(f,t)} - e^{i\psi_k(f)}e^{i\lambda_k(t)f}|^2 + \sum_{k,f}\sum_{t\notin\Omega_k} \frac{\sigma_u}{\tilde{\sigma}^2}|X(f,t)|^2|e^{i\phi_k(f,t)} - e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}|^2, \quad (7.24)$$

soit, en utilisant les expressions (7.3) et (7.6),

$$\log(p_\Phi) \stackrel{c}{=} -\frac{1}{\tilde{\sigma}^2}\left(\sigma_r\mathcal{C}_r(\phi,\psi,\lambda) + \sigma_u\mathcal{C}_u(\phi)\right). \quad (7.25)$$

Modèle complet

La maximisation de la distribution à postériori revient donc à maximiser :

$$\begin{aligned}\mathcal{C}_{MAP}(\theta) &= L(\theta) + \log(p_\Phi) + \log(p_H)\\ &\stackrel{c}{=} -\frac{1}{\tilde{\sigma}^2}\left(D(X,\hat{X}) + \sigma_u\mathcal{C}_u(\phi), +\sigma_r\mathcal{C}_r(\phi,\psi,\lambda) + \sigma_s\mathcal{C}_s(H)\right),\end{aligned}$$

ce qui revient à minimiser la fonction de coût complète suivante :

$$\mathcal{C}(\theta) = D(X,\hat{X}) + \sigma_u\mathcal{C}_u(\phi), +\sigma_r\mathcal{C}_r(\phi,\psi,\lambda) + \sigma_s\mathcal{C}_s(H). \quad (7.26)$$

On retrouve donc la fonction de coût introduite de façon intuitive dans la section 7.1.1.

7.2 Estimation du modèle

Nous présentons dans cette section les algorithmes d'estimation du modèle de NMF complexe contrainte. Nous introduisons certaines notations afin de simplifier l'écriture des règles de mise à jour. On commence par définir les matrices suivantes, $\forall k \in [\![1,K]\!]$:

$$\begin{aligned}&\mu_k \in \mathbb{C}^{F\times 1},\ \mu_k(f) = e^{2i\pi S\nu_k(f)},\\ &\Lambda_k \in \mathbb{C}^{F\times T},\ \Lambda_k(f,t) = \mathbb{1}_k(t)e^{if\lambda_k(t)},\\ &\Psi_k \in \mathbb{C}^{F\times 1},\ \Psi_k(f) = e^{i\psi_k(f)},\\ &\Phi_k \in \mathbb{C}^{F\times T},\ \Phi_k(f,t) = e^{i\phi_k(f,t)}.\end{aligned}$$

On désigne par H_k la k-ième ligne de H et par W_k la k-ième colonne de W. On utilise également les notations suivantes :

- $M_\downarrow$ (respectivement $M_\uparrow$) est la matrice obtenue en retirant la dernière ligne (respectivement la première ligne) de M.
- $M_\rightarrow$ (respectivement $M_\leftarrow$) désigne la matrice obtenue en retirant la dernière (respectivement la première) colonne de M et en insérant une colonne de 0 en tant que première (respectivement en temps que dernière) colonne.

2. Les hyper-paramètres dépendent donc des observations $|X(f,t)|^2$, ce qui peut sembler étrange, mais donne expérimentalement de bons résultats. Ce type d'approches est connu sous le nom d'approches *Bayésiennes empiriques* Efron (2012).

— $\texttt{diag}_\texttt{v}(v)$ désigne la matrice diagonale dont les éléments diagonaux sont les termes du vecteur v, et $\texttt{diag}_\texttt{m}(M)$ désigne le vecteur colonne obtenu par extraction des éléments diagonaux de la matrice M.

— $\texttt{vand}(v)$ désigne la matrice de Vandermonde obtenue à partir du vecteur v : si $v \in \mathbb{C}^{1\times T}$, alors $M = \texttt{vand}(v) \in \mathbb{C}^{F\times T}$, avec $\forall(f,t)$, $M(f,t) = v(t)^f$.

La minimisation de la fonction de coût (7.8) peut être effectuée par minimisation successive par rapport à chacune des variables (annulation des dérivées partielles). Cela permet d'aboutir à une procédure itérative. On n'a pas de garantie de convergence dans le cas général. Nous donnons le détail mathématique de cette technique d'optimisation appliquée à la fonction de coût (7.8) dans l'Annexe C (section C.1) de ce manuscrit afin de ne pas surcharger ce chapitre. La procédure complète est décrite dans l'Algorithme 8. Notons que l'ordre dans lequel sont effectuées les mises à jour des paramètres est arbitraire, et il pourrait être intéressant de considérer un ordre différent pour en évaluer l'impact sur les résultats.

Alternativement, on peut appliquer la méthode de la fonction auxiliaire, qui fournit un cadre théorique pour obtenir des règles de mise à jour. Les détails de calcul sont donnés dans l'Annexe C (section C.2). Essentiellement, la différence avec la première méthode se trouve dans la présence d'un terme G_k qui est un gain (ici choisi comme étant celui de Wiener) qui permet de "redistribuer" l'erreur d'approximation sur les diverses composantes. Ces gains proviennent des inégalités de convexité utilisées pour obtenir une fonction auxiliaire, et pourraient tout à fait prendre des valeurs différentes que celles proposées ici. Les mises à jour obtenues sont synthétisées dans l'Algorithme 9.

Pour l'initialisation, nous proposons d'appliquer une première NMF à $|X|$ afin d'obtenir une première approximation de W et de H. Alternativement, celles-ci peuvent être gardées aléatoires et l'algorithme de NMF Complexe peut être appliqué directement sur ces valeurs. Nous initialisons les phases en leur donnant celle du mélange : $\forall k \in [\![1, K]\!]$, $\Phi_k = \frac{X}{|X|}$, et les paramètres de phases d'attaque Ψ_k et Λ_k par des valeurs aléatoires de module 1. Il faut également déterminer $\mathbb{1}_k$, qui est l'indicatrice de Ω_k. Il est possible d'estimer directement celle-ci à partir de H_k en utilisant par exemple Paulus et Virtanen (2005). Nous souhaitons avoir une estimation précise des trames d'attaque (afin de se concentrer essentiellement sur l'influence de la reconstruction de phase), aussi nous avons utilisé la boîte à outils MATLAB Tempogram Toolbox Grosche et Müller (2011) appliquée sur chaque spectrogramme de source isolément.

Algorithme 8 Estimation du modèle CNMF à phase contrainte (méthode de relaxation)

Entrées :
X, K, σ_r, σ_u, σ_s.
Initialisation $\forall k \in [\![1, K]\!]$:
W_k, H_k, Φ_k,
Λ_k, Ψ_k, $\mathbb{1}_k$, $\overline{\mathbb{1}}_k$,
$\hat{X}_k = (W_k H_k) \odot \Phi_k$,
$B_k = X - \sum_{l \neq k} \hat{X}_l$.
tant que critère non atteint **faire**
pour $k = 1$ à K **faire**

Calculer μ_k

QIFFT sur W_k et découpage en régions d'influence.

Calculer Ψ_k

$$\Psi_k = \frac{\mathtt{diag_m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)}{|\mathtt{diag_m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)|}.$$

Calculer Λ_k

$$\Lambda_k = \mathtt{vand}\left(\frac{(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H (|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})}{|(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H (|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})|}\right) \mathtt{diag_v}(\mathbb{1}_k).$$

Calculer ρ_k

$$\rho_k = \sigma_r(\Psi_k \mathbb{1}_k) \odot \Lambda_k \odot |X|^{\odot 2} + \sigma_u(\mu_k \overline{\mathbb{1}}_k) \odot \Phi_{k,\rightarrow} \odot |X|^{\odot 2} + \sigma_u(\overline{\mu}_k \overline{\mathbb{1}}_{k,\leftarrow}) \odot \Phi_{k,\leftarrow} \odot |X_\leftarrow|^{\odot 2}.$$

Calculer Φ_k

$$\Phi_k = \frac{B_k \odot (W_k H_k) + \rho_k}{|B_k \odot (W_k H_k) + \rho_k|}.$$

Actualiser $\hat{X}_k$ **et** B_k

Calculer β_k

$$\beta_k = \Re(B_k \odot \overline{\Phi}_k).$$

Calculer W

$$W_k = \frac{\beta_k (H_k)^T}{\alpha((H_k)^{\odot 2})^T}.$$

Calculer H

$$H_k = \frac{(W_k)^T \beta_k}{p\sigma_s (H_k)^{\odot p-2} + ((W_k)^{\odot 2})^T \alpha}.$$

Projection de W **et** H **sur l'orthant positif**
Normaliser W **et** H
Actualiser $\hat{X}^k$ **et** B_k
fin pour
fin tant que
Sorties :
$\hat{X}_k$, W_k, H_k, Φ_k,
Λ_k, Ψ_k, μ_k.

Algorithme 9 Estimation du modèle CNMF à phase contrainte (méthode de la fonction auxiliaire)

Entrées :
X, K, σ_r, σ_u, σ_s.
Initialisation $\forall k \in [\![1, K]\!]$:
W_k, H_k, Φ_k,
Λ_k, Ψ_k, $\mathbb{1}_k$, $\overline{\mathbb{1}}_k$,
$\hat{X}_k = (W_k H_k) \odot \Phi_k$,
$G_k = \frac{W_k H_k}{\sum_l W_l H_l}$,
$B_k = \hat{X}_k + G_k \odot (X - \hat{X})$.
tant que critère non atteint **faire**
pour $k = 1$ à K **faire**

Calculer μ_k

QIFFT sur W_k et découpage en régions d'influence.

Calculer Ψ_k

$$\Psi_k = \frac{\texttt{diag}_\texttt{m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)}{|\texttt{diag}_\texttt{m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)|}.$$

Calculer Λ_k

$$\Lambda_k = \texttt{vand}\left(\frac{(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H(|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})}{|(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H(|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})|}\right)\texttt{diag}_\texttt{v}(\mathbb{1}_k).$$

Calculer ρ_k

$$\rho_k = \sigma_r(\Psi_k \mathbb{1}_k) \odot \Lambda_k \odot |X|^{\odot 2} + \sigma_u(\mu_k \overline{\mathbb{1}}_k) \odot \Phi_{k,\rightarrow} \odot |X|^{\odot 2} + \sigma_u(\overline{\mu}_k \overline{\mathbb{1}}_{k,\leftarrow}) \odot \Phi_{k,\leftarrow} \odot |X_\leftarrow|^{\odot 2}.$$

Calculer Φ_k

$$\Phi_k = \frac{B_k \odot (W_k H_k)/G_k + \rho_k}{|B_k \odot (W_k H_k)/G_k + \rho_k|}.$$

Actualiser $\hat{X}_k$, B_k **et** G_k

Calculer β_k

$$\beta_k = \Re(B_k \odot \overline{\Phi}_k).$$

Calculer W

$$W_k = \frac{\frac{\beta_k}{G_k}(H_k)^T}{\frac{1}{G_k}((H_k)^{\odot 2})^T}.$$

Calculer H

$$H_k = \frac{(W_k)^T \frac{\beta_k}{G_k}}{p\sigma_s(\bar{H}_k)^{\odot p-2} + ((W_k)^{\odot 2})^T \frac{1}{G_k}}.$$

Projection de W **et** H **sur l'orthant positif**
Normaliser W **et** H
Actualiser $\hat{X}_k$, B_k **et** G_k
fin pour
fin tant que
Sorties :
$\hat{X}_k$, W_k, H_k, Φ_k,
Λ_k, Ψ_k, μ_k.

7.3 Résultats expérimentaux

Nous avons conduit un ensemble d'expériences afin d'évaluer la performance de notre modèle de NMF complexe. Tout d'abord, nous avons cherché à étudier l'influence des paramètres σ_r et σ_u sur la performance de l'algorithme, mesurée grâce au SDR, SIR et SAR Vincent et al. (2006). Puis, nous avons comparé cette approche dans le cadre de la séparation de sources à d'autres méthodes : la NMF complexe non contrainte, et une NMF classique avec filtrage de Wiener pour reconstruire la phase.

7.3.1 Données et protocole

Les jeux de données sont :

A : 30 mélanges de deux sources, composées de sinusoïdes amorties synthétiques. Les sources se recouvrent dans le domaine TF.

B : 30 mélanges de deux notes de piano tirées de la base de données MAPS Emiya et al. (2010). Les sources se recouvrent également dans le domaine TF.

C : 100 mélanges de morceaux de musique polyphonique tirés de la base DSD100 Ono et al. (2015), que nous avons eu l'occasion de présenter dans le chapitre 4.

Pour les données A et B, chaque source est observée seule, puis les deux sont activées simultanément. Les signaux sont échantillonnés à 11025 Hz sur ces jeux de données, et à 44100 Hz sur la base C. La TFCT est calculée avec une fenêtre de Hann de longueur 46 ms (512 échantillons pour les données A et B et 2048 échantillons pour les données C). Le taux de recouvrement est de 75 %. Les paramètres de parcimonie sont quant à eux fixés à des valeurs régulièrement utilisées dans la litérature (*cf.* par exemple Kameoka et al. (2009)) : $p = 1$ et $\sigma_s = ||X||_2^2 K^{-(1-p/2)} 10^{-5}$.

7.3.2 Mélanges simples

Influence des paramètres

Dans cette expérience, nous évaluons, pour les jeux de données A et B, l'influence des paramètres σ_r et σ_u sur la qualité de la séparation effectuée avec 10 itérations de l'algorithme 9, initialisé avec 30 itérations de KLNMF.

Les résultats présentés dans l'article ICASSP Magron et al. (2016b) étaient obtenus par l'algorithme 8. Nous présentons ici les résultats obtenus avec la méthode de la fonction auxiliaire, qui sont globalement meilleurs. Néanmoins, en utilisant l'algorithme 9, nous constatons que le paramètre σ_r a une influence négligeable sur les résultats (ce qui n'est pas le cas avec l'algorithme 8). Pour les expériences sur les données A et B, on considère donc $\sigma_r = 0$. Les résultats de l'impact du paramètre σ_u sur la séparation de sources sont donnés sur la figure 7.2.

Globalement, des valeurs non-nulles de σ_u conduisent à améliorer les résultats par rapport à une valeur nulle, ce qui montre l'intérêt de notre méthode par rapport à une CNMF non contrainte. Un compromis entre SDR, SIR et SAR semble obtenu pour $\sigma_u = 1$ (données synthétiques) et $\sigma_u = 0.1$ (notes de piano). Des valeurs plus faibles sont insuffisantes pour contraindre efficacement le problème (on se rapproche alors du modèle de CNMF non contrainte), et des valeurs plus importantes sont au contraire trop contraignantes car les données ne respectent pas "suffisamment" le modèle (et l'influence du terme d'attache aux données dans la fonction de coût devient négligeable).

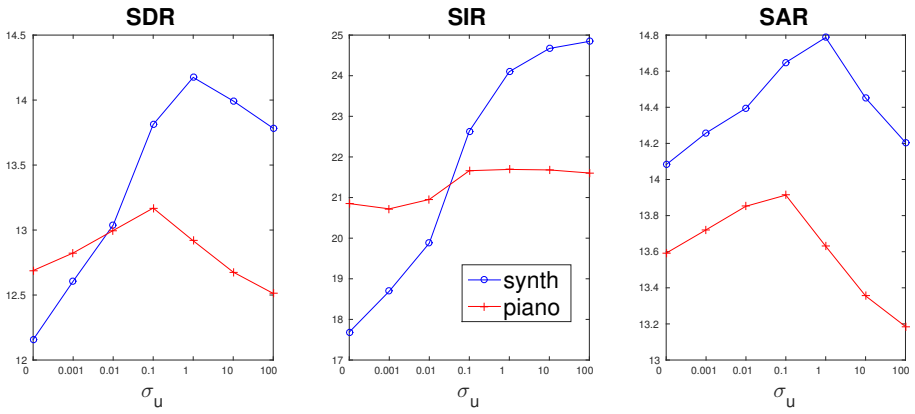


FIGURE 7.2 – Influence du paramètre σ_u sur la qualité de séparation de sources.

Données	Méthode	SDR	SIR	SAR
A	NMF-W	11.7	16.8	**13.5**
	CNMF	9.6	16.7	10.7
	CNMF-ϕ	**12.3**	**23.2**	12.8
B	NMF-W	**14.7**	18.5	**17.4**
	CNMF	13.3	19.6	14.9
	CNMF-ϕ	14.6	**22.1**	15.7

TABLEAU 7.1 – Performance de la séparation de source (SDR, SIR et SAR en dB) pour divers jeux de données et méthodes.

Séparation de sources

Nous effectuons une tâche de séparation de sources avec différentes méthodes :

— **NMF-W** : Une KLNMF [3] suivie d'un filtrage de Wiener pour la reconstruction de phase ;

— **CNMF** : La NMF complexe sans contraintes de phase ;

— **CNMF-ϕ** : Notre algorithme de NMF complexe à phase contrainte, avec $\sigma_u = 1$ pour les données A et $\sigma_u = 0.1$ pour les données B.

La KLNMF est effectuée avec 30 itérations de mises à jour multiplicatives. Celle-ci sert d'initialisation aux NMF complexes, qui utilisent ensuite 10 itérations de l'algorithme 9 (en effet, l'algorithme **CNMF** décrit dans KAMEOKA et al. (2009) revient à utiliser l'algorithme 9 avec $\sigma_r = \sigma_u = 0$). Par ailleurs, à partir de la KNLMF initiale, on effectue 10 itérations supplémentaires de KLNMF afin que les différentes méthodes utilisent le même nombre d'itérations en tout. Les résultats de la qualité de la séparation de sources (SDR, SIR et SAR) sont présentés dans le tableau 7.1.

On constate un résultat similaire à celui observé dans le chapitre précédent, à savoir que le gain en performance dû à cet algorithme est à nuancer selon les indicateurs. Pour des signaux synthétiques (données A), on note une augmentation modérée de SDR et un gain plus important de SIR, au détriment d'une légère dégradation de SAR par rapport au filtrage de Wiener. L'amélioration reste significative par rapport à une NMF complexe non contrainte.

Sur les signaux de piano, le filtrage de Wiener conduit aux meilleurs résultats (en SDR et SAR), bien que notre algorithme permette une nette augmentation de SIR. Il est possible que

3. Par souci d'équité, la KLNMF est également assortie d'une contrainte de parcimonie, pour ne pas avantager les autres méthodes.

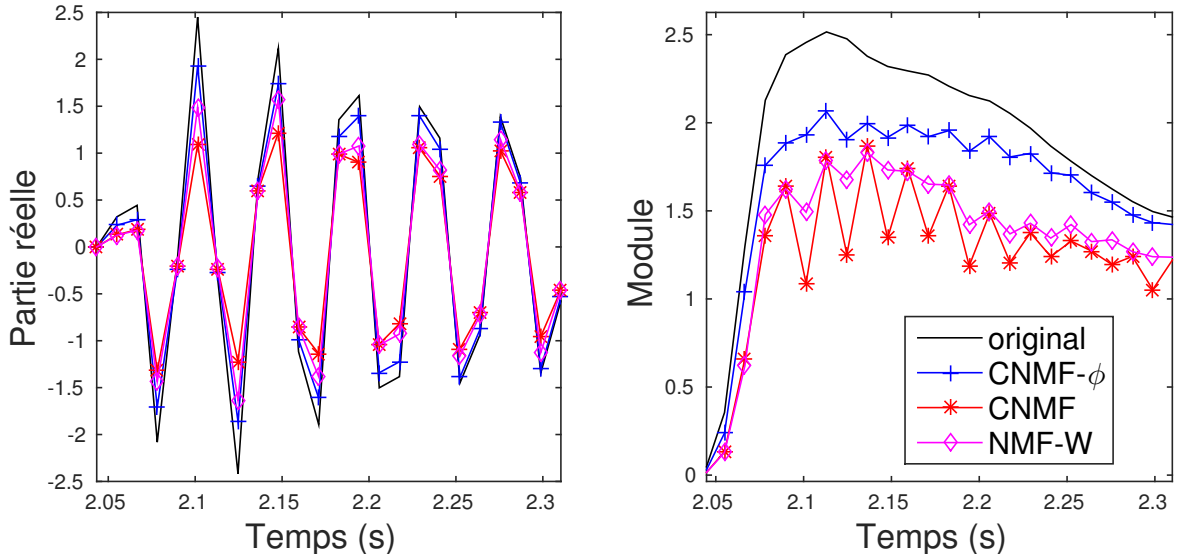


FIGURE 7.3 – Reconstruction d'un partiel de la note B2 à partir d'un mélange (E2 et B2), dans le canal fréquentiel à 495 Hz où les deux notes se recouvrent. Partie réelle (gauche) et amplitude (droite).

la NMF avec filtrage de Wiener initial soit déjà un résultat de bonne qualité, aussi poursuivre la séparation avec une CNMF ne semble pas forcément judicieux. Dans l'expérience suivante, nous verrons si tel est toujours le cas lorsque l'on traite des signaux plus compliqués. Pour nuancer ce résultat, notons que sur les notes de piano, la dégradation du SDR entre **NMF-W** et **CNMF-ϕ** est faible (0.1 dB) alors que le gain en SIR est plus important (3.6 dB). La diminution du SAR est également modérée.

Par ailleurs, on représente les signaux reconstruits (partie réelle et amplitude) dans une bande de fréquences et sur une durée où il y a recouvrement afin d'illustrer ces résultats. On choisit pour cela un mélange de deux notes de piano E2 et B2 et on représente le résultat sur la figure 7.3. On constate que les signaux reconstruits par les méthodes **NMF-W** et **CNMF** sont assez éloignés du signal original, dans le sens ou le phénomène de battement dû au recouvrement des sources impacte les signaux séparés. Le fait de contraindre la phase permet de s'en affranchir partiellement, ce qui se traduit également par une amélioration perceptive de la qualité des signaux.

Enfin, en analysant plus précisément les résultats, on constate que pour des notes de piano graves, la CNMF à phase contrainte donne de meilleurs résultats que **NMF-W**, alors que cette tendance s'inverse à mesure que l'on monte dans les aigus. Ce phénomène peut s'expliquer par le fait que dans les basses fréquences, les recouvrements et les battements qui en découlent sont plus marqués que dans les hautes fréquences, d'où l'intêret particulier d'une contrainte de phase pour cette gamme de sons.

7.3.3 Morceaux de musique polyphoniques

Nous reprenons les expériences précédentes, appliquées ici au jeu de données C. Plus précisément, cette base de données comprend 50 morceaux d'apprentissage et 50 morceaux de test. Nous étudions l'influence des paramètres sur les morceaux d'apprentissage, afin d'en déduire les valeurs optimales pour effectuer la séparation de sources sur la base de test. L'algorithme 9 est utilisé. Les sources sont reconstruites par clustering Oracle BARKER et al. (2013) à partir des atomes NMF, dont le principe est rappelé dans la section 2.2.6.

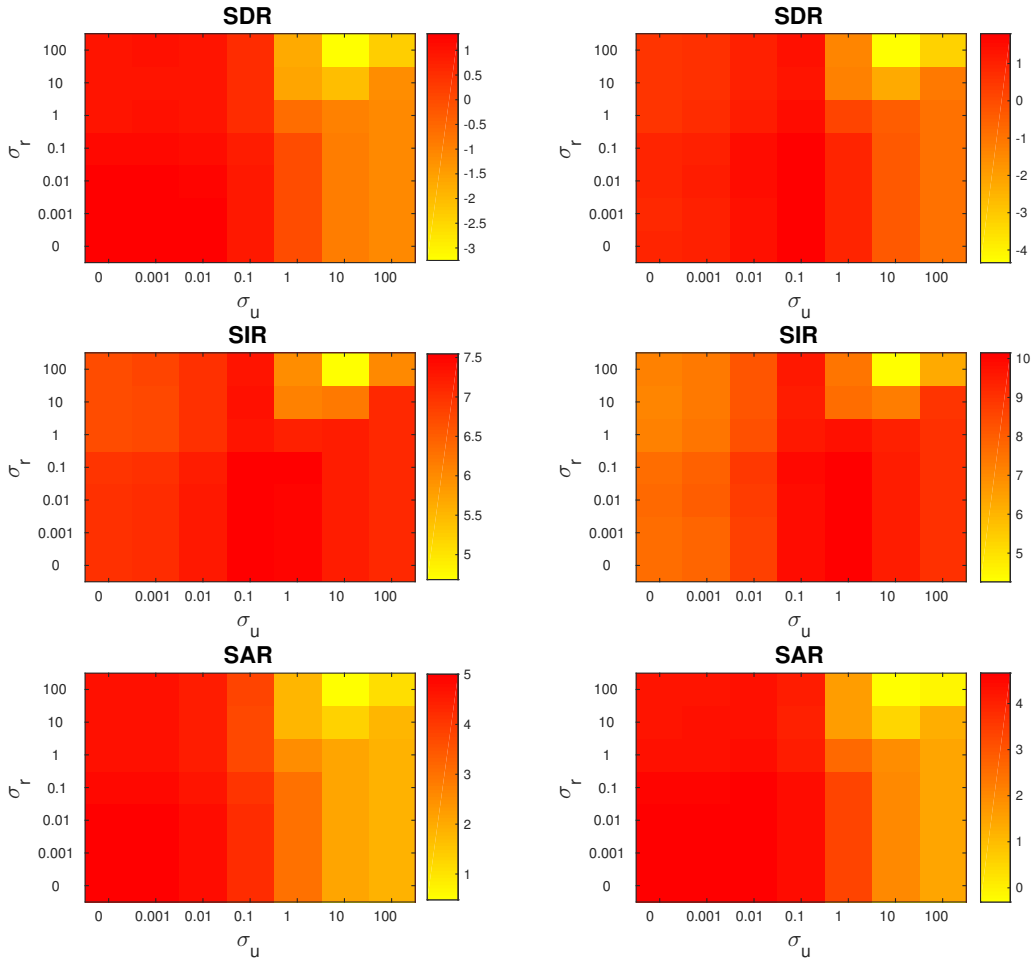


FIGURE 7.4 – Influence des paramètres σ_u et σ_r sur la qualité de séparation de sources (données DSD100), pour une initalisation par NMF (à gauche) et aléatoire (à droite).

Influence des paramètres

Nous testons deux initialisations :

— Une initialisation aléatoire, auquel cas l'algorithme de CNMF utilise 30 itérations ;

— Une initialisation par 30 itérations d'une KLNMF préalable, auquel cas l'algorithme de CNMF utilise 10 itérations.

Les résultats obtenus sont illustrés sur la figure 7.4. Nous constatons que le jeu de paramètres qui permet de maximiser la qualité de la séparation de sources (au niveau des trois indicateurs) semble être localisé autour du couple $(\sigma_r, \sigma_u) \approx (0.1, 0.1)$. Des valeurs plus faibles sont insuffisantes pour contraindre efficacement le problème (on se rapproche alors du modèle de CNMF non contrainte), et des valeurs plus importantes semblent au contraire trop contraignantes car les données ne respectent pas suffisamment le modèle. La valeur du paramètre σ_r est peu influente : les résultats sont principalement dépendants de la contrainte de déroulé linéaire, et non de la contrainte de répétition dans les trames d'attaque.

L'initialisation par KLNMF semble conduire à des résultats plus robustes (moins de variabilité notamment en SIR) mais en revanche, les valeurs de SIR en particulier sont moindres.

Méthode	SDR	SIR	SAR
NMF-W	**1.9**	10.2	**3.7**
CNMF	1.4	10.9	2.9
CNMF-ϕ	1.7	**12.2**	2.9

TABLEAU 7.2 – Performance de la séparation de source (SDR, SIR et SAR en dB) pour diverses méthodes sur la base DSD100.

En d'autres termes, le choix de l'initialisation n'impacte que peu les SDR et SAR, mais avec une initialisation aléatoire, on peut s'attendre à un gain d'environ 3 dB en SIR. Ce résultat, bien que suprenant, peut s'expliquer par le fait que la KLNMF initiale conduit à un minimum local de la fonction de coût, dont il est ensuite difficile de s'éloigner par application de la CNMF (similairement aux résultats de HRNMF dans le cas aveugle dans l'étude comparative du chapitre 3). En outre, cette approche est moins coûteuse en nombre d'itérations. Ce point est néanmoins à relativiser puisqu'on n'utilise alors aucune itération de KLNMF, mais davantage d'itérations de CNMF (qui sont un peu plus lourdes en temps de calcul).

Pour la dernière expérience, nous choisissons donc une initialisation aléatoire et on fixe les valeurs des paramètres comme suit : $(\sigma_r, \sigma_u) \approx (0.1, 0.1)$.

Séparation de sources

Nous testons la performance de notre méthode pour la séparation de sources sur la base de test DSD100. Chacune des méthodes utilise 30 itérations des algorithmes impliqués, et est initialisée avec des valeurs aléatoires pour W et H. Les résultats sont présentés dans le tableau 7.2.

La CNMF à phase contrainte donne de meilleurs résultats en SDR et SIR que la CNMF non contrainte, et une valeur similaire de SAR. Cela confirme l'intérêt de contraindre la NMF complexe pour en améliorer la performance. Par rapport au filtrage de Wiener, elle conduit à une baisse de 0.2 dB et 0.8 dB en SDR et SAR respectivement, et à un gain de 2 dB en SIR. Aussi, le choix d'une technique plutôt qu'une autre pourrait être motivée par la recherche d'un compromis entre ces différents indicateurs. On observe en particulier que l'utilisation d'une contrainte de déroulé linéaire de phase à tendance à réduire les interférences entre sources, conclusion déjà observée au chapitre 5.

7.4 Conclusion

Le modèle de NMF complexe à phase contrainte introduit dans ce chapitre est un outil prometteur pour la séparation de sources, notamment pour la réduction des interférences qui surviennent lorsqu'une méthode plus traditionnelle de reconstruction de phase est appliquée à des mélanges dans lesquels les sources se recouvrent dans le domaine TF. Ces contraintes de phase, basées sur les modèles de signaux que nous avons développés dans ce manuscrit, améliorent les résultats par rapport à une approche non contrainte. Nous avons observé expérimentalement que la contrainte au niveau des trames d'attaque n'avait pas un impact significatif sur les résultats. On pourrait donc envisager, à l'avenir, de travailler principalement à la reconstruction des phases des partiels dans le cadre d'une NMF complexe.

Nous avons supposé que le mélange était égal au modèle auquel est ajouté un bruit. Il pourrait être plus judicieux de directement modéliser les sources comme des variables latentes, ce qui permettrait de nouvelles techniques d'estimation. En outre, le modèle n'est pas conservatif, c'est-à-dire que la somme des sources n'est (en général) pas égale au mélange observé.

Le cadre probabiliste est donc plus adapté à ce type de problèmes, et c'est la direction que nous envisageons de suivre dans la troisième partie de ce manuscrit.

Troisième partie

Modèles probabilistes de sources à phase non-uniforme

Chapitre 8

Modèle gaussien anisotrope à phase informée

Sommaire

La technique de déroulé linéaire de phase introduite au chapitre 4 a été appliquée à la séparation de sources au chapitre 5. Cette méthode était basée sur la minimisation d'une fonction de coût pénalisant l'écart entre modèle et mélange observé, et l'information sur la phase provenant du modèle sinusoïdal était introduite dans l'algorithme via son initialisation. Ce type d'approche souffre néanmoins de deux écueils. Tout d'abord, l'ensemble des sources estimées ne constitue pas un modèle conservatif (la somme des estimées n'est pas égale au mélange observé). En outre, nous avons constaté dans le cas non-Oracle (où les spectrogrammes ne sont plus connus) une baisse des résultats, qui est due à une contrainte trop forte sur les amplitudes.

Les modèles probabilistes sont des outils adaptés à l'injection de connaissance préalable sur les paramètres que l'on cherche à estimer. Ils permettent donc d'éviter ces deux écueils, puisque l'on peut alors introduire une certaine incertitude sur les paramètres. La plupart des modèles probabilistes de sources, comme Févotte et al. (2009), sont basés sur une hypothèse d'uniformité de la phase. Cette hypothèse est vérifiée lorsqu'on considère que tous les points TF sont indépendants, et qu'on ne tient pas compte de la structure de la phase Parry et Essa (2007). Néanmoins, comme le suggère le modèle de phase introduit dans le chapitre 4, introduire des dépendances entre phases de points TF successifs améliore la qualité des signaux reconstruits. Dans ce chapitre, nous proposons un modèle probabiliste de mélange dans lequel les phases ne sont plus des variables aléatoires uniformes.

Les principales contributions de ce chapitre ont fait l'objet d'une publication à la conférence ICASSP 2017 Magron et al. (2017a).

Dans la section 8.1, nous introduisons un nouveau modèle de mélange basé sur la distribution de Von Mises pour représenter la phase. Nous approchons ensuite dans la section 8.2 celui-ci par un modèle gaussien équivalent, plus simple à manipuler. Dans la section 8.3, nous montrons expérimentalement le potentiel de ce modèle pour la séparation de sources musicales. Enfin, nous concluons dans la section 8.4.

8.1 Modèle de Von Mises

Nous proposons dans cette partie d'introduire un modèle de mélange basé sur la loi de Von Mises pour représenter les phases.

8.1.1 Modèle de sources

On raisonne dans un point TF (f, t) donné, tous les points étant supposés indépendants. Ainsi, on retire les indices f et t dans ce qui suit afin de clarifier les écritures. Le mélange X est égal à une somme de sources Z_k. Nous supposons les modules V_k des sources estimés au préalable (par une NMF par exemple) ou bien connus (ils sont déterministes dans ce modèle). Ainsi, sous forme polaire, on a, $\forall k$:

$$Z_k = V_k e^{i\phi_k}. \tag{8.1}$$

Notre idée consiste à considérer que les phases ϕ_k des sources sont des variables aléatoires qui sont relativement "proches" d'une première estimation (obtenue par déroulé linéaire) que nous notons μ_k. À la différence de l'approche probabiliste présentée dans le chapitre précédent, notre but ici n'est pas d'estimer au sens MAP la phase, mais d'obtenir un estimateur MMSE des variables Z_k. Une façon naturelle de modéliser ce comportement serait de considérer que ϕ_k est une gaussienne centrée en μ_k. Néanmoins, comme la phase est une variable périodique, il est nécessaire d'utiliser une statistique circulaire. La loi normale périodique Mardia et Jupp (2000) est une possibilité, et elle a été étudiée notamment dans Agiomyrgiannakis

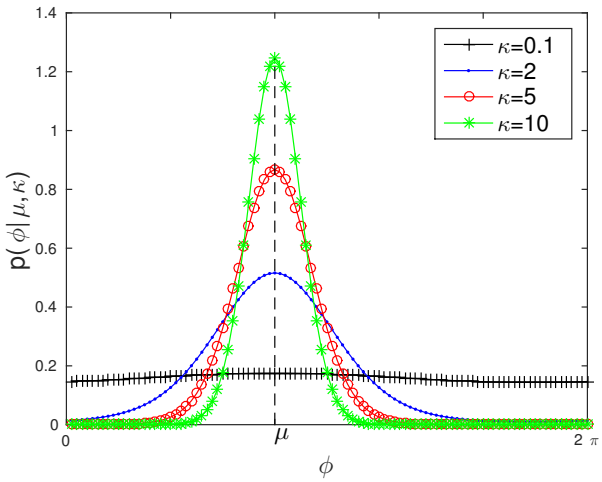


FIGURE 8.1 – Densité de probabilité d'une distribution de Von Mises pour un paramètre de localisation μ et plusieurs valeurs du paramètre de concentration κ.

et STYLIANOU (2009) pour modéliser la phase dans une application de rehaussement de la parole. Néanmoins, cette loi ne possède pas de densité analytiquement simple (elle s'écrit comme une somme de série infinie), aussi nous avons choisi d'utiliser son approximation : la loi de Von Mises MARDIA et ZEMROCH (1975). Celle-ci est très populaire dans le domaine des statistiques circulaires, et ce pour trois raisons :

— Sa densité de probabilité est simple à écrire analytiquement ;

— Elle constitue une bonne approximation de la loi normale périodique ;

— C'est la statistique circulaire à entropie maximale (à paramètres de localisation et de concentration donnés), ce qui garantit une certaine régularité et donc de bonnes propriétés mathématiques.

La loi de Von Mises, notée $\mathcal{VM}(\mu, \kappa)$ dépend de deux paramètres : un paramètre de localisation $\mu \in [0, 2\pi[$ (μ peut être réel, mais on retient en général sa valeur principale), qui joue le rôle de moyenne, et un paramètre de concentration $\kappa \in]0, +\infty[$ qui joue le rôle de l'inverse d'une variance. Sa densité, définie sur un intervalle de longueur 2π, est donnée par :

$$p(\phi|\mu, \kappa) = \frac{e^{\kappa \cos(\phi - \mu)}}{2\pi I_0(\kappa)}, \tag{8.2}$$

où I_n désigne la fonction de Bessel modifiée de première espèce d'ordre n WATSON (1995). En particulier, on remarquera que :

— Lorsque $\kappa \to 0$, la concentration autour du mode μ devient nulle : la loi de Von Mises devient alors équivalente à une loi uniforme.

— Lorsque $\kappa \to +\infty$, toute la masse est concentrée sur le mode μ de la distribution : la densité tend alors vers un dirac centré en μ, ce qui revient à dire que ϕ n'est plus une variable aléatoire et devient déterministe égale à μ (modulo 2π).

La figure 8.1 représente la densité de la loi de Von Mises pour plusieurs valeurs de κ.

Le modèle de mélange est donc :

$$X = \sum_k Z_k \text{ avec } Z_k = V_k e^{i\phi_k} \text{ et } \phi_k \sim \mathcal{VM}(\mu_k, \kappa_k). \tag{8.3}$$

8.1.2 Verrous du modèle

Connaissant les V_k, et supposant les κ_k fixés (par exemple estimés au préalable sur une base d'apprentissage) et les μ_k estimés par déroulé linéaire, la tâche de séparation de sources consiste à obtenir un estimateur des composantes complexes Z_k. Les estimateurs les plus naturels sont donnés par l'estimateur du maximum de vraisemblance, l'estimateur du maximum à postériori, et l'espérance à postériori de ces variables sachant les observations.

Le problème est qu'en pratique, le calcul de ces estimateurs requiert nécessairement le calcul de la vraisemblance (la densité à postériori dépendant de la vraisemblance). Cela impose donc d'être capable d'estimer la loi de X, ainsi que la loi des Z_k. Or, malgré nos efforts calculatoires, nous ne sommes pas parvenus à les obtenir sous une forme analytique simple. En effet, bien que la densité de ϕ_k soit aisée à écrire, la densité de X s'écrit sous la forme d'une intégrale d'ordre $K-1$ à l'intérieur de laquelle apparaissent des sommes de séries de Bessel.

L'estimation de l'espérance à postériori des variables Z_k pourrait être effectuée par des techniques de simulation telles que les méthodes de Monte Carlo par chaînes de Markov (MCMC) Andrieu et al. (2003); Robert et Casella (2013). L'idée principale est d'approcher l'espérance par la moyenne empirique calculée à partir d'échantillons suivant la loi dont on cherche l'espérance (méthodes de Monte Carlo). Comme on n'a pas forcément accès à cette loi (ici la loi à postériori des sources), on génère des échantillons selon un mécanisme de chaîne de Markov qui est construite de sorte à "ressembler" à cette loi (c'est le principe de l'algorithme de Metropolis-Hasting). Néanmoins, là encore, les algorithmes MCMC nécessitent certaines informations comme la loi à priori sur les variables. Or, celle-ci n'est pas exprimable analytiquement sous forme simple.

Ainsi, le modèle (8.3), malgré son apparente simplicité, semble trop complexe pour que les méthodes traditionnelles permettent d'obtenir un estimateur des sources. Par ailleurs, bien que des développements sur les méthodes MCMC pourraient permettre d'obtenir une solution, celles-ci sont basées sur la simulation et donc sur une procédure itérative, nécessitant un grand nombre d'itérations.

Nous proposons donc de l'approcher par un modèle simplifié, qui permette d'effectuer un certain nombre de calculs, tout en conservant la contrainte de phase qui motive cette approche.

8.1.3 Relations avec d'autres modèles

De tels modèles de phase pour tenir compte d'un modèle sinusoïdal ont été utilisés dans un cadre bayésien pour l'estimation de parole en milieu bruité, notamment dans Gerkmann (2014); Kulmer et Mowlaee (2015); Mowlaee et Kulmer (2015). La loi de Von Mises gagne donc depuis peu une certaine popularité en audio. Néanmoins, dans Mowlaee et Kulmer (2015), la phase est estimée au sens du Maximum à Postériori, alors que nous nous intéressons ici à l'obtention d'un estimateur MMSE des sources. Cette approche est conduite dans Gerkmann (2014), mais ce modèle est limité à un mélange parole plus bruit : aussi, les calculs deviennent plus aisés et il est possible d'obtenir une écriture de l'estimateur MMSE du signal de parole sous la forme d'une intégrale relativement simple à approcher numériquement. Dans notre cas (mélange de K sources), ce type d'approche ne fonctionne plus, d'où notre idée d'une approximation par modèle simplifié.

8.2 Modèle gaussien anisotrope

8.2.1 Loi normale complexe

L'idée que nous avons retenue consiste à approcher le modèle (8.3) par un modèle qui soit plus aisé à manipuler. Pour cela, nous avons considéré un modèle gaussien complexe non isotrope. En effet, les distributions gaussiennes sont omniprésentes en probabilités, avec de très nombreux développements et résultats théoriques rendant leur utilisation agréable. Elles apparaissent naturellement, comme c'est formalisé dans le théorème central limite notamment, comme une famille de référence pouvant approcher bon nombre de lois. L'idée qui consiste à approcher une somme de variables indépendantes à phases aléatoires par une somme de gaussiennes équivalentes n'est d'ailleurs pas neuve, comme en témoigne Beckmann (1962). On propose donc le modèle suivant :

$$X = \sum_k X_k \text{ avec } X_k \sim \mathcal{N}(m_k, \gamma_k, c_k), \tag{8.4}$$

où les X_k sont indépendants, et $m_k \in \mathbb{C}$, $\gamma_k \in \mathbb{R}_+$ et $c_k \in \mathbb{C}$ désignent respectivement la moyenne, la variance et le terme de *relation* de X_k, définis comme suit :

$$\begin{aligned} m_k &= \mathbb{E}(X_k), \\ \gamma_k &= \mathbb{E}(|X_k - m_k|^2), \\ c_k &= \mathbb{E}((X_k - m_k)^2). \end{aligned}$$

Nous définissons également la matrice de covariance Γ_k du vecteur $\underline{X}_k = \begin{pmatrix} X_k \\ \bar{X}_k \end{pmatrix}$:

$$\Gamma_k = \begin{pmatrix} \gamma_k & c_k \\ \bar{c}_k & \gamma_k \end{pmatrix}, \tag{8.5}$$

où $\bar{.}$ désigne le conjugué complexe. Nous rappelons également l'expression de la densité de probabilité d'une telle loi :

$$p(X_k | m_k, \gamma_k, c_k) = \frac{1}{\pi\sqrt{\det(\Gamma_k)}} e^{-\frac{1}{2}(\underline{X}_k - \underline{m}_k)^H \Gamma_k^{-1} (\underline{X}_k - \underline{m}_k)}, \tag{8.6}$$

où det est le déterminant et $.^H$ désigne la transposée Hermitienne. En particulier, si $m_k = 0$ et $c_k = 0$, on parle de loi normale complexe circulaire-symétrique. Si en outre la variance γ_k est égale à 1, il s'agit de la loi normale complexe standard (l'équivalent complexe de la loi normale centrée réduite).

La propriété d'additivité de la famille des lois normales implique alors :

$$X \sim \mathcal{N}(m_X, \gamma_X, c_X) \text{ avec } m_X = \sum_k m_k,\ \gamma_X = \sum_k \gamma_k \text{ et } c_X = \sum_k c_k, \tag{8.7}$$

la matrice de covariance de X étant égale à $\Gamma_X = \sum_k \Gamma_k$.

8.2.2 Paramètres du modèle

Afin de garantir la cohérence entre notre modèle initial et son approximation gaussienne, il convient que les moments des gaussiennes X_k soient identiques à ceux des sources Z_k dans

le modèle précédent de Von Mises :

$$\begin{aligned} m_k &= \mathbb{E}(X_k) = \mathbb{E}(Z_k), \\ \gamma_k &= \mathbb{E}(|X_k - m_k|^2) = \mathbb{E}(|Z_k - m_k|^2), \\ c_k &= \mathbb{E}((X_k - m_k)^2) = \mathbb{E}((Z_k - m_k)^2). \end{aligned}$$

La conservation de ces moments est donc ce qui permet de conserver, dans ce nouveau modèle, une propriété de non-uniformité des phases.

Moyenne Commençons par déterminer la moyenne m_k de chaque composante :

$$m_k = \mathbb{E}(X_k) = \mathbb{E}(Z_k) = V_k \mathbb{E}(e^{i\phi_k}). \tag{8.8}$$

Lorsque une variable est circulaire, on nomme *moment circulaire d'ordre n* la quantité $\mathbb{E}(e^{in\phi_k})$. Dans le cas de la loi de Von Mises, on ne sait pas déterminer la loi de $e^{i\phi_k}$ mais on sait en revanche calculer ses moments circulaires Mardia et Zemroch (1975) :

$$\forall n \in \mathbb{Z},\ \mathbb{E}(e^{in\phi_k}) = \frac{I_{|n|}(\kappa_k)}{I_0(\kappa_k)} e^{in\mu_k}. \tag{8.9}$$

Ainsi, la moyenne de X_k est donnée par :

$$m_k = V_k \frac{I_1(\kappa_k)}{I_0(\kappa_k)} e^{i\mu_k} = \lambda_k \tilde{X}_k, \tag{8.10}$$

où l'on a posé $\lambda_k = \dfrac{I_1(\kappa_k)}{I_0(\kappa_k)}$, et $\tilde{X}_k = V_k e^{i\mu_k}$ est la source estimée par déroulé linéaire appliqué isolément sur les sources.

Variance On cherche à présent à déterminer les termes de la matrice de covariance Γ_k de X_k, ce qui implique de déterminer les paramètres de variance et de relation. On a :

$$\begin{aligned} \gamma_k &= \mathbb{E}(|Z_k - m_k|^2) \\ &= \mathbb{E}(|Z_k|^2) - |m_k|^2 \\ &= V_k^2 - V_k^2 \frac{I_1(\kappa_k)^2}{I_0(\kappa_k)^2}, \end{aligned}$$

d'après le théorème de König-Huygens. Avec la notation λ_k, on a donc :

$$\gamma_k = \left(1 - \lambda_k^2\right) V_k^2. \tag{8.11}$$

Terme de relation De façon similaire, on calcule le terme de relation :

$$\begin{aligned} c_k &= \mathbb{E}((Z_k - m_k)^2) \\ &= \mathbb{E}(Z_k^2) - m_k^2 \\ &= V_k^2 \mathbb{E}(e^{i2\phi_k}) - V_k^2 \frac{I_1(\kappa_k)^2}{I_0(\kappa_k)^2} e^{i2\mu_k} \\ &= V_k^2 \frac{I_2(\kappa_k)}{I_0(\kappa_k)} e^{i2\mu_k} - V_k^2 \frac{I_1(\kappa_k)^2}{I_0(\kappa_k)^2} e^{i2\mu_k} \\ &= \rho_k \tilde{X}_k^2. \end{aligned}$$

où l'on a introduit la notation suivante :

$$\rho_k = \frac{I_2(\kappa_k) I_0(\kappa_k) - I_1(\kappa_k)^2}{I_0(\kappa_k)^2}. \tag{8.12}$$

Mélange L'intérêt du modèle gaussien est que l'on peut sommer les distributions précédemment identifiées en vertu de l'équation (8.7). On a donc :

$$m_X = \sum_k \lambda_k \tilde{X}_k, \tag{8.13}$$

$$\gamma_X = \sum_k \left(1 - \lambda_k^2\right) V_k^2, \tag{8.14}$$

$$c_X = \sum_k \rho_k \tilde{X}_k^2. \tag{8.15}$$

8.2.3 Comparaison des deux modèles

Afin de comparer les deux modèles (Von Mises et gaussien anisotrope), nous générons 10000 échantillons de chaque variable aléatoire Z_k et X_k, et ce pour différentes valeurs du paramètre de concentration. Nous fixons le module à 1 et le paramètre de localisation de phase à $\pi/3$. Nous représentons les histogrammes 2D de ces distributions sur la figure 8.2. On observe tout d'abord que les échantillons générés dans le modèle de Von Mises sont tous situés sur le cercle de rayon 1 : c'est logique puisque dans ce modèle, l'amplitude ne varie pas. L'approximation par modèle gaussien introduit une incertitude sur cette amplitude, qui est liée aussi bien au module qu'à la phase, mais également au paramètre de concentration. Plus celui-ci est élevé, plus la variation d'amplitude devient faible. On remarque enfin que plus le paramètre de concentration est élevé, plus les deux distributions ont tendance à se ressembler.

8.2.4 Estimateur MMSE des sources

Obtention de l'estimateur

Maintenant que l'on a accès à tous les éléments du modèle, on peut fournir un estimateur MMSE des sources, donné par l'espérance à postériori de celles-ci. Dans le cadre d'un modèle gaussien, il existe une formule simple qui donne cette espérance (voir par exemple Bishop (2006)) :

$$\underline{\hat{X}}_k = \underline{m}_k + \Gamma_k \Gamma_X^{-1} (\underline{X} - \underline{m}_X). \tag{8.16}$$

L'expression ci-dessus donne l'estimateur $\hat{X}_k$ ainsi que son conjugué $\bar{\hat{X}}_k$. Ainsi, pour éviter cette redondance d'information, on ne considère que la première ligne de (8.16) :

$$\hat{X}_k = m_k + \begin{pmatrix}\gamma_k & c_k\end{pmatrix} \Gamma_X^{-1} \begin{pmatrix} X - m_X \\ \bar{X} - \bar{m}_X \end{pmatrix} \tag{8.17}$$

On constate d'ores et déjà que cet estimateur est conservatif. En effet, comme $\sum_k \Gamma_k = \Gamma_X$, on a :

$$\sum_k \underline{\hat{X}}_k = \sum_k \underline{m}_k + (\underline{X} - \underline{m}_X) = \underline{X}. \tag{8.18}$$

Cette propriété garantit que quelle que soit la valeur des κ_k, on est assuré que la somme des estimateurs restera égale au mélange, ce qui ne serait pas le cas si on avait considéré les estimateurs obtenus par déroulé linéaire isolé $\tilde{X}_k$.

Interprétation

L'estimateur (8.16) réalise en fait une interpolation entre le filtrage de Wiener (qui n'utilise que la phase du mélange), et l'estimateur construit en utilisant μ_k uniquement (et pas la phase du mélange). Considérons ici que $\forall k$, les κ_k sont constants et égaux à une valeur κ. On

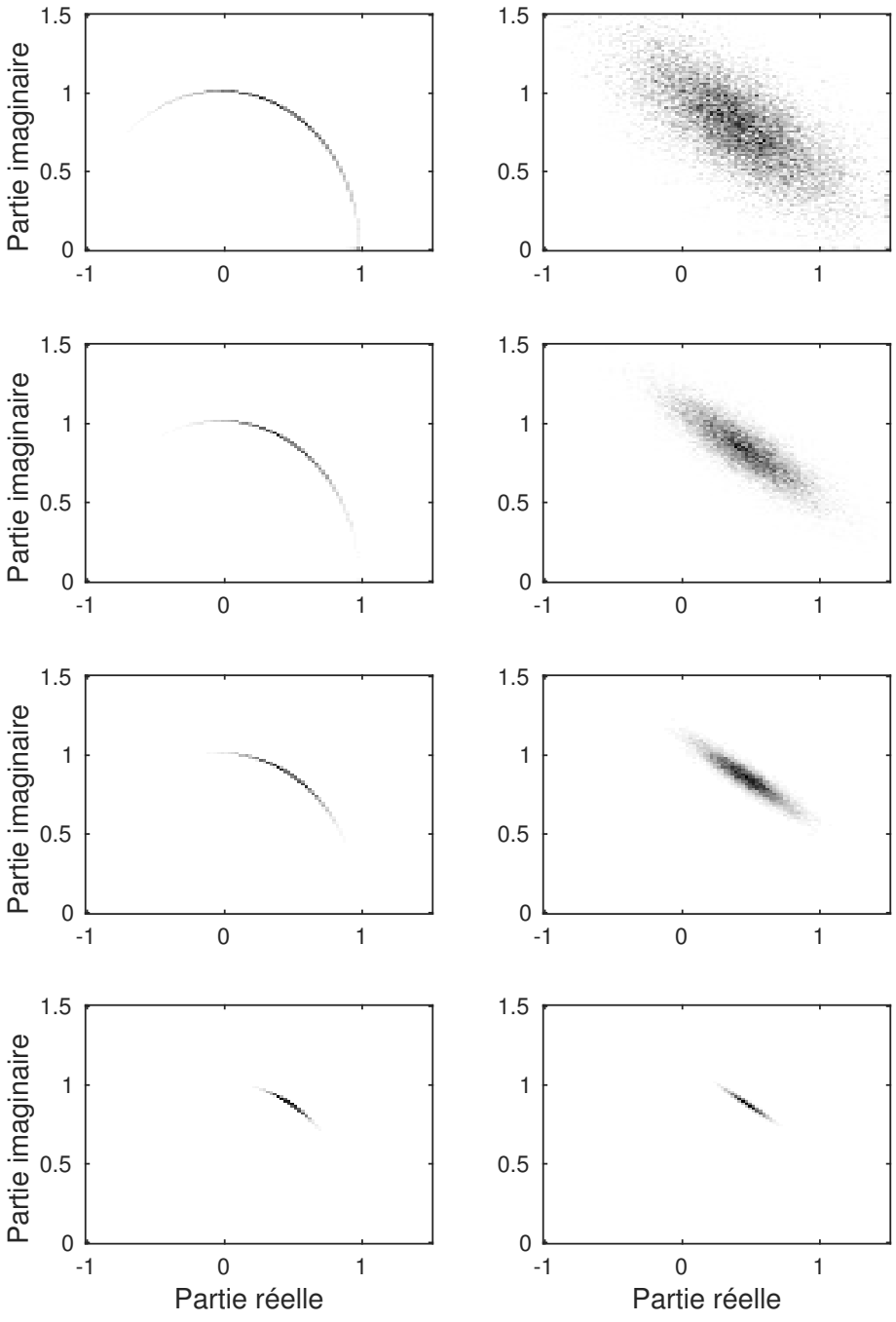


FIGURE 8.2 – Histogrammes en 2D d'échantillons générés par le modèle de Von Mises (gauche) et le modèle équivalent gaussien (droite), pour $V = 1$ et $\mu = \pi/3$. Le paramètre de concentration κ prend, de haut en bas, les valeurs 5, 10, 20 et 100.

désignera alors par λ et ρ les paramètres correspondants, et on illustre le comportement de ceux-ci en fonction de κ sur la figure 8.3.

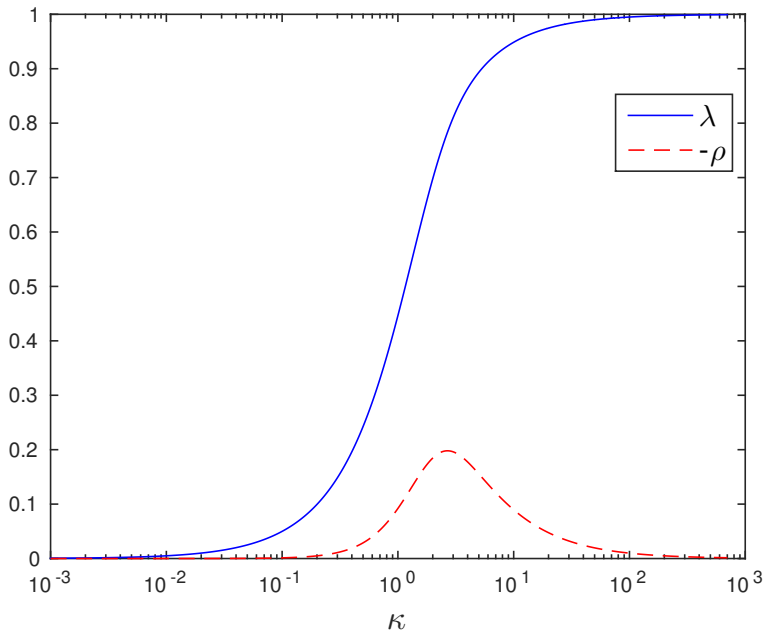


FIGURE 8.3 – Evolution des paramètres λ et ρ en fonction de κ.

Supposons que le paramètre de concentration κ tende vers 0. On montre alors, en utilisant les propriétés des fonctions de Bessel WATSON (1995) (et comme c'est suggéré sur la figure 8.3) que $\lambda \to 0$ et $\rho \to 0$. On a alors $m_k \to 0$, $c_k \to 0$ et $\gamma_k \to V_k^2$. Ainsi, on constate que l'estimateur (8.16) tend vers le filtrage de Wiener :

$$\hat{X}_k \underset{\kappa \to 0}{\longrightarrow} \begin{pmatrix} V_k^2 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sum_l V_l^2} & 0 \\ 0 & \frac{1}{\sum_l V_l^2} \end{pmatrix} \begin{pmatrix} X \\ \bar{X} \end{pmatrix} = \frac{V_k^2}{\sum_l V_l^2} X, \tag{8.19}$$

où l'on reconnaît le gain de Wiener traditionnel $G_k = \dfrac{V_k^2}{\sum_l V_l^2}$. Ce résultat est logique, puisqu'une valeur nulle du paramètre de concentration dans la distribution de Von Mises correspond, comme on l'a vu, à une distribution uniforme. Cela revient à dire que l'on a aucune information à priori sur la phase, et on obtient finalement un modèle gaussien isotrope classique FÉVOTTE et al. (2009) pour lequel on sait que l'estimateur MMSE des sources est donné par le filtrage de Wiener traditionnel.

Les deux estimateurs précédents (filtrage de Wiener et déroulé linéaire appliqué isolément sur les sources) ont montré leurs limites, comme nous avons eu plusieurs fois l'occasion de le rappeler. Le filtrage de Wiener introduit en effet des interférences lorsque les signaux sont recouvrants dans le domaine TF, et le déroulé linéaire seul, appliqué isolément sur les sources, a tendance à créer des artéfacts du fait des erreurs d'estimation sur les fréquences instantanées (voir chapitre 4). L'estimateur (8.17) que nous proposons est un compromis entre ces deux approches, dont la performance dépendra du réglage du paramètre κ.

De façon intéressante, ce type d'estimateur qui combine optimalement filtrage de Wiener et modèle de phase par déroulé linéaire a déjà été introduit dans KRAWCZYK et GERKMANN (2015) dans un but de débruitage de la parole, néanmoins sans utiliser un modèle de Von Mises.

Algorithme de séparation de sources

Comme on considère que tous les points TF sont indépendants, on pourrait appliquer l'estimateur 8.17 matriciellement et ainsi reconstruire directement toutes les sources. Néanmoins, similairement à ce que nous avons constaté dans le chapitre 5, il est préférable d'adopter une approche séquentielle sur les trames temporelles : en effet, mieux vaut estimer $\mu_k(f,t)$ par déroulé linéaire à partir de $\phi_k(f,t-1) = \angle\hat{X}_k(f,t-1)$, qui est un estimateur de la phase dans la trame précédente où l'erreur sur μ_k est "corrigée" par la phase du mélange, plutôt que directement à partir de $\mu_k(f,t-1)$. Autrement dit, on choisit de calculer séquentiellement les estimateurs afin d'éviter au maximum de propager les erreurs d'estimation qui proviennent du déroulé linéaire.

L'algorithme 10 résume cette procédure de séparation de sources dans un cas informé (les amplitudes sont préalablement estimées). On rappelle que Ω_k désigne l'ensemble des trames d'attaque pour la source k.

Un cas particulier : le modèle parole plus bruit

Comme nous l'avons rappelé plus haut, les modèles de mélanges de sources utilisant la loi de Von Mises ont déjà été étudiés en audio, notamment dans le cadre du débruitage de la parole Gerkmann (2014); Kulmer et Mowlaee (2015). On peut voir ce type de mélanges comme un cas particulier de notre modèle avec $K = 2$ sources, en supposant néanmoins que les paramètres de concentration ne sont pas les mêmes pour les deux sources. En effet, on a $X = \tilde{S} + B$, $\tilde{S}$ correspondant au signal de parole, que l'on peut écrire sous la forme $Ve^{i\phi}$, ϕ suivant une loi de Von Mises de paramètres κ et μ (phase obtenue par déroulé linéaire). B est généralement considéré comme étant un bruit blanc gaussien.

Ainsi, on peut utiliser notre approximation par mélange de gaussiennes anisotropes. $\tilde{S}$ est donc approchée par S, qui est une variable gaussienne dont les paramètres sont décrits par les équations données dans la section 8.2.2. Le bruit suit également ce modèle, à condition que l'on choisisse un paramètre de concentration nul (on a alors une gaussienne complexe circulaire-symétrique).

À la différence de Gerkmann (2014) qui obtient un estimateur MMSE directement à partir du modèle de Von Mises, le notre est obtenu à partir d'une approximation gaussienne. Néanmoins, il ne requiert pas le calcul numérique d'une intégrale, aussi on peut supposer qu'il soit plus efficace sur le plan du temps de calcul. L'application de ce modèle au débruitage de la parole pourra donc être envisagée à l'avenir.

Algorithme 10 Séparation de sources par application d'un estimateur MMSE combinant filtrage de Wiener et déroulé linéaire.

Entrées :

Mélange $X \in \mathbb{C}^{F\times T}$,

Spectrogrammes $V_k \in \mathbb{R}_+^{F\times T}$, $\forall k \in [\![1, K]\!]$,

Ensembles des trames d'attaque Ω_k, $\forall k \in [\![1, K]\!]$,

Phases d'attaque $\phi_k^o(f,t)$, $\forall t \in \Omega_k$,

Paramètre de concentration $\kappa_k \in \mathbb{R}_+^{F\times T}$.

pour $t = 1$ à $T-1$ **faire**

pour $k = 1$ à K **faire**

si $t \in \Omega_k$ **alors**

Phase d'attaque : $\forall f$, $\mu_k(f,t) = \phi_k^o(f,t)$.

sinon

$\mu_k(f,t)$ est obtenue par déroulé linéaire (*cf.* Algorithme 2).

fin si

fin pour

Pour toutes les sources k et tous les canaux fréquentiels f :

À prioris $\tilde{X}_k(f,t) = V_k(f,t)e^{i\mu_k(f,t)}$.

Paramètres $\lambda_k(f,t) = \dfrac{I_1(\kappa_k(f,t))}{I_0(\kappa_k(f,t))}$ et $\rho_k(f,t) = \dfrac{I_2(\kappa_k(f,t))I_0(\kappa_k(f,t)) - I_1(\kappa_k(f,t))^2}{I_0(\kappa_k(f,t))^2}$.

Moments

$m_k(f,t) = \lambda_k(f,t)\tilde{X}_k(f,t)$, $m_X(f,t) = \sum_k m_k(f,t)$,

$\gamma_k(f,t) = (1-\lambda_k(f,t)^2)V_k^2$, $\gamma_X(f,t) = \sum_k \gamma_k(f,t)$,

$c_k(f,t) = \rho_k(f,t)\tilde{X}_k(f,t)^2$, $c_X(f,t) = \sum_k c_k(f,t)$.

Estimateur MMSE

$$\hat{X}_k(f,t) = m_k(f,t) + \begin{pmatrix}\gamma_k(f,t) & c_k(f,t)\end{pmatrix} \begin{pmatrix}\gamma_X(f,t) & c_X(f,t)\\ \bar{c}_X(f,t) & \gamma_X(f,t)\end{pmatrix}^{-1} \begin{pmatrix}X(f,t) - m_X(f,t)\\ \bar{X}(f,t) - \bar{m}_X(f,t)\end{pmatrix}$$

fin pour

Sortie :

$\forall k \in [\![1, K]\!]$, $\hat{X}_k \in \mathbb{C}^{F\times T}$.

8.3 Validation expérimentale

Dans cette partie, nous conduisons une série d'expériences pour montrer le potentiel de l'estimateur (8.17). Les données utilisées proviennent de la base DSD100 Ono et al. (2015), constituée de 100 morceaux de musique polyphonique composés de $K = 4$ sources. Celle-ci est divisée en deux moitiés (une pour l'apprentissage du paramètre de concentration optimal et une autre pour l'expérience de séparation de sources). Les signaux sont échantillonnés à $F_s = 44100$ Hz et la TFCT est calculée avec une fenêtre de Hann de longueur 4096 échantillons et 75 % de recouvrement. Les amplitudes V_k sont soit supposées connues (cas Oracle) soit estimées par NMF sur les sources séparées (cas semi-Oracle). On initialise les phases dans les trames d'attaque en leur donnant celle du mélange. La qualité de la séparation est mesurée par le SDR, le SIR et le SAR calculés avec la boîte à outils BSS Eval Vincent et al. (2006).

8.3.1 Influence du paramètre de concentration

Tout d'abord, nous considérons 50 morceaux issus de la base de données (base de développement) et appliquons l'algorithme 10 pour estimer les sources séparées à partir des mélanges en faisant varier le paramètre de concentration κ. Nous examinons deux cas de figure : soit le paramètre de concentration κ est le même pour tous les points TF, soit il est variable, auquel cas on utilise un paramètre de concentration $\tilde{\kappa}_k(f,t)$ tel que :

$$\tilde{\kappa}_k(f,t) = \kappa(1 - G_k(f,t)), \tag{8.20}$$

où $G_k(f,t)$ est le gain de Wiener. En effet, ce choix est motivé par le fait que plus le gain de Wiener est important, plus une source est prédominante dans un point TF considéré. Dans ce cas, on peut s'attendre à ce que la phase du mélange soit très proche de la phase de la source dominante, auquel cas la phase de $\tilde{X}_k$ dans l'estimateur des sources devient moins importante que celle du mélange X.

Les résultats dans le cas Oracle sont présentés sur la figure 8.4. On constate tout d'abord que lorsque κ devient proche de 0, les performances de l'estimateur deviennent similaires à celles obtenues par filtrage de Wiener. Cela confirme notre interprétation de la section 8.2.4 et montre que l'estimateur (8.17) est une généralisation du filtrage de Wiener. Nous observons ce qui semble être un maximum de SDR et de SAR autour de $\kappa = 1.6$, qui permet d'augmenter les performances par rapport à Wiener d'environ 2 dB pour ces indicateurs, et d'environ 3 dB pour le SIR. Le SIR ne semble pas maximisé pour cette valeur : en effet, augmenter le paramètre de concentration continue d'augmenter le SIR. C'est logique puisque plus ce paramètre est élevé, moins la phase du mélange devient importante, ce qui implique une réduction des interférences entre les sources. Néanmoins, on constate que les SDR et SAR décroissent après ce pic, ce qui là encore est expliqué par le fait qu'une concentration trop forte autour de μ_k par rapport à la donnée de phase du mélange peut avoir tendance à créer des artéfacts (dûs aux erreurs d'estimation et aux écarts au modèle lorsqu'on applique le déroulé linéaire).

On constate par ailleurs que les paramètres de concentration variables (8.20) ne donnent pas globalement de meilleurs résultats qu'un paramètre constant. On pourrait cependant chercher une nouvelle définition de ces paramètres qui potentiellement mènerait à de meilleurs résultats, car le choix (8.20) était arbitraire.

Les résultats dans le scénario semi-Oracle, présentés sur la figure 8.5, sont similaires, et montrent qu'il est préférable, lorsque les amplitudes ne sont plus exactement égales à la vérité terrain, d'utiliser un paramètre de concentration constant $\kappa = 1$ plutôt qu'un paramètre variable.

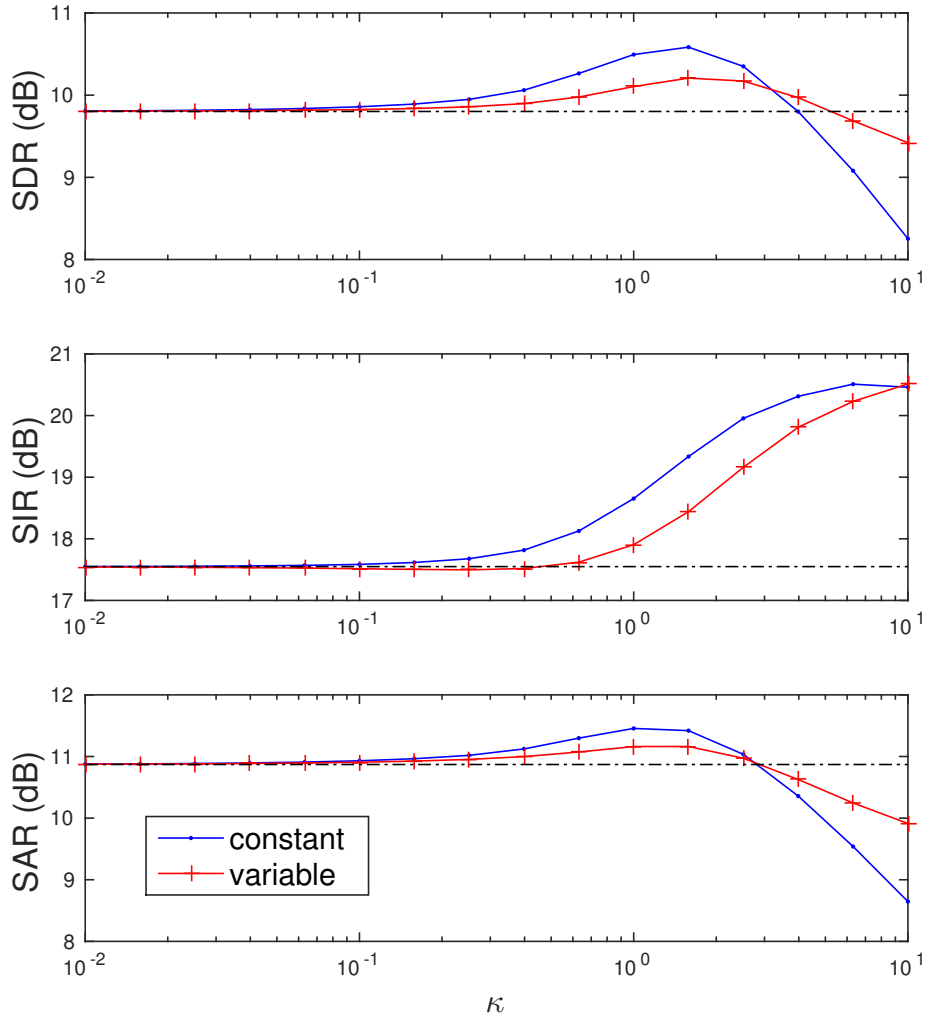


FIGURE 8.4 – Influence du paramètre de concentration κ sur la qualité de la séparation de sources dans l'algorithme 10 (courbes en traits pleins) et comparaison au filtrage de Wiener (courbe en pointillés) dans le cas Oracle. Le paramètre de concentration peut être constant pour toutes les sources et les points TF ou bien variable, selon l'équation (8.20).

Cette première expérience montre l'intérêt de notre estimateur. Pour la suite de nos expériences, nous utiliserons dans le cas Oracle la valeur $\kappa = 1.6$, qui semble être un bon compromis entre les différents indicateurs. Dans le cas semi-Oracle, on prendra $\kappa = 1$.

8.3.2 Séparation de sources musicales

Nous considérons à présent la deuxième moitié de la base de données (50 morceaux constituant la base de tests) et les valeurs du paramètre de concentration apprises précédemment. Nous appliquons le même protocole que dans l'expérience précédente, et comparons les résultats de notre approche (notée **MMSE**) avec plusieurs autres méthodes :

— le filtrage de Wiener FÉVOTTE et al. (2009) qui sera noté **Wiener** ;

— le filtrage de Wiener consistant LE ROUX et VINCENT (2013), noté **W-Cons**. Confor-

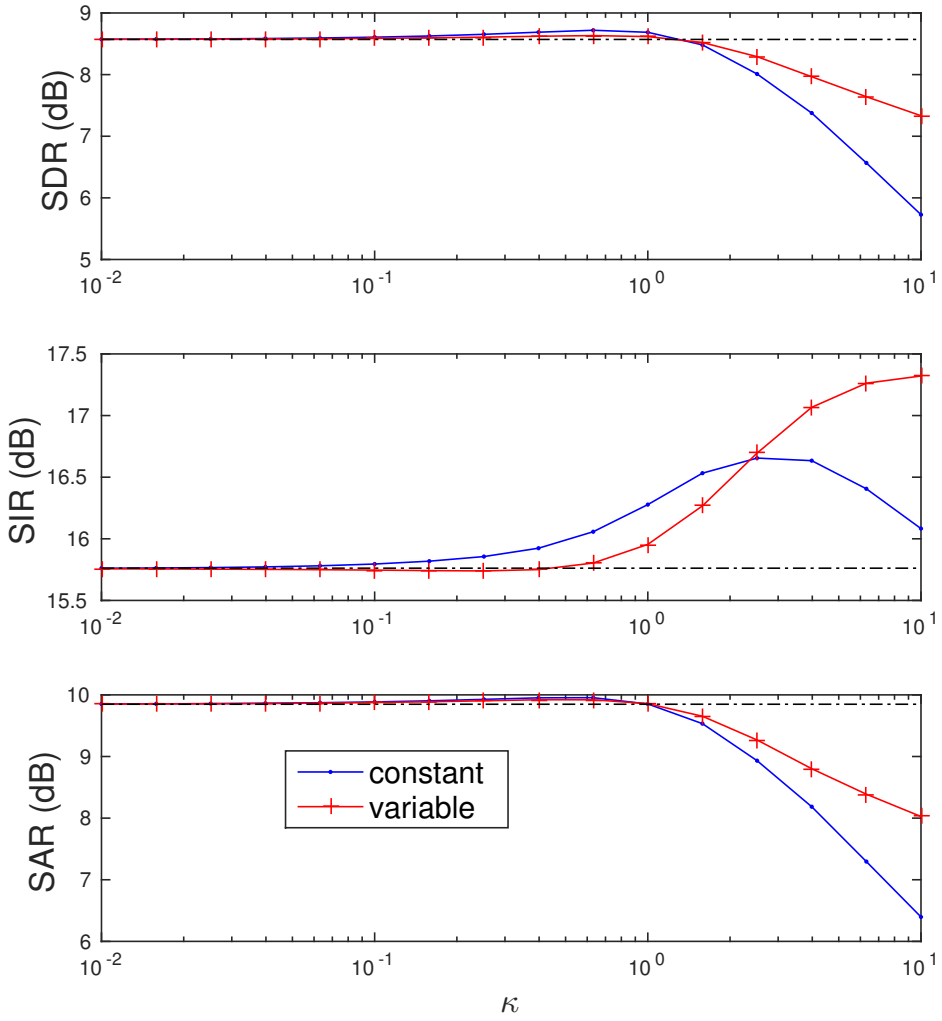


FIGURE 8.5 – Influence du paramètre de concentration κ sur la qualité de la séparation de sources dans l'algorithme 10 (courbes en traits pleins) et comparaison au filtrage de Wiener (courbe en pointillés) dans le cas semi-Oracle. Le paramètre de concentration peut être constant pour toutes les sources et les points TF ou bien variable, selon l'équation (8.20).

mément aux résultats obtenus au chapitre 5 dans la section 5.3.1, nous choisissons un paramètre de consistance égal à 4 pour cet algorithme.

— l'estimateur n'utilisant que le déroulé de phase $\tilde{X}_k$, c'est-à-dire sans tenir compte de la phase du mélange. Il sera noté **Unwrap**.

Les résultats sont présentés sur la figure 8.6. Nous constatons qu'aussi bien dans le cas Oracle que semi-Oracle, l'estimateur que nous proposons fournit des résultats légèrement meilleurs que le filtrage de Wiener, et légèrement moins bons que le filtrage de Wiener consistant en SDR, SIR et SAR. La méthode **Unwrap** ne conduit pas à des résultats satisfaisants, notamment au niveau du SDR et du SAR. On peut interpréter ceci comme au chapitre 5 : l'application du déroulé sans tenir compte de la phase du mélange à tendance à faire se propager l'erreur de déroulé linéaire sans le corriger au fur et à mesure, ce qui créé des artéfacts qui sont perceptibles. Le SIR, en revanche, n'est pas trop impacté, puisque cet estimateur

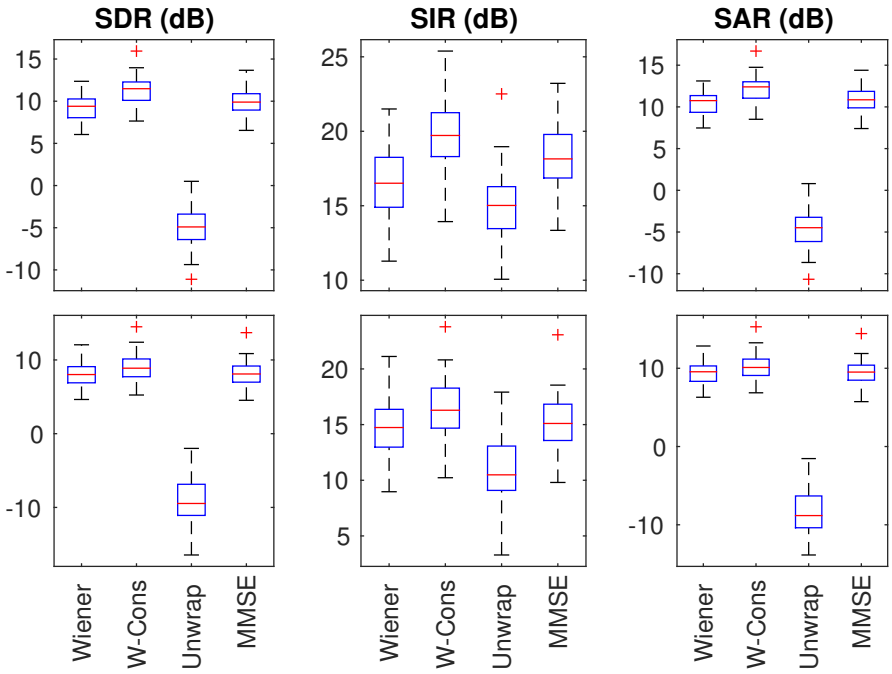


FIGURE 8.6 – Performance de la séparation de sources (SDR, SIR et SAR en dB) pour diverses méthodes sur la base DSD100. Résultats Oracle (en haut) et sur spectrogrammes approchés (en bas).

	Scénario Oracle			Scénario semi-Oracle		
	SDR	SIR	SAR	SDR	SIR	SAR
Iter	**10.0**	**20.5**	**10.4**	5.3	**15.0**	5.9
MMSE	9.0	16.7	9.9	**7.4**	13.9	**8.6**

TABLEAU 8.1 – Performance de reconstruction (SDR en dB) pour divers jeux de données.

agissant sur les sources séparées, il créé moins d'interférences entre sources que les autres méthodes, qui utilisent la phase du mélange. Du point de vue perceptif, nous n'avons pas entendu de différence significative entre les pistes `vocals`, `other` et `drum` estimées par filtrage de Wiener consistant et par notre approche, néanmoins nous avons noté une nette diminution des artéfacts dans la piste `bass` par notre méthode. Enfin, soulignons que notre approche est nettement moins gourmande que **W-Cons** en temps de calcul (d'un facteur 7).

Nous effectuons enfin une comparaison entre la méthode présentée au chapitre 5 (méthode **Iter**) et celle introduite ici (**MMSE**). Les résultats moyennés sur les 50 morceaux de la base de données sont présentés dans le tableau 8.1. Nous constatons que dans le cas Oracle, l'approche itérative donne les meilleurs résultats pour tous les indicateurs. Néanmoins, lorsque les amplitudes ne sont plus égales à la vérité terrain, l'estimateur **MMSE** conduit à un SDR et un SAR plus élevés, bien que **Iter** donne toujours un meilleur SIR. Cela s'explique par le fait que l'estimateur **MMSE** tienne compte d'une incertitude sur les amplitudes, alors que dans l'approche itérative, celles-ci sont fixées à chaque itération à la valeur objectif, qui n'est plus exacte dans le cas semi-Oracle. Les deux approches utilisent un temps de calcul similaire.

En fin de compte, on peut résumer ces conclusions comme suit :

— Si on est soucieux d'implémenter une méthode très rapide, quitte à perdre en qualité,

on utilisera le filtrage de Wiener ;

— Si les spectrogrammes qui ont été préalablement estimés sont considérés comme étant d'excellente qualité, on pourra appliquer la procédure **Iter** ;

— Si on s'éloigne de cette valeur Oracle, il est plus prudent d'appliquer l'estimateur **MMSE** introduit dans ce chapitre.

Enfin, le filtrage de Wiener consistant donne de bons résultats, mais reste limitant du point de vue du temps de calcul.

8.4 Conclusion

Le modèle introduit dans ce chapitre se base sur une hypothèse de phase non-uniforme, contrairement à ce qui est fréquemment supposé dans la littérature. Du modèle gaussien découle un estimateur MMSE des sources qui allie bonne qualité de séparation et rapidité de calcul. Nous avons expérimentalement démontré son potentiel pour une application de séparation de sources. Nous avons proposé que le paramètre de localisation de la phase soit estimé par déroulé linéaire, mais n'importe quel μ_k pourrait en réalité être utilisé dans le modèle.

Bien que l'on puisse considérer que l'approche itérative introduite au chapitre 5 mène à des résultats légèrement meilleurs dans le cas Oracle, l'avantage de ce modèle est qu'il pose les bases d'une méthode complète de séparation de sources. En effet, les amplitudes étaient ici supposées estimées par avance, et nous n'avons pas directement modélisé l'incertitude sur ces estimées (elle a été introduite par le passage au modèle gaussien anisotrope). On pourrait alors envisager de modéliser les V_k comme des variables aléatoires suivant une loi à support non-négatif (loi de Poisson, de Rayleigh, Gamma...), et structurer les paramètres de dispersion de celle-ci par un modèle de type NMF. Des applications en séparation aveugle de sources pourront alors être envisagées.

Chapitre 9

Modélisation robuste de données non-négatives

Sommaire

Dans la deuxième partie de ce manuscrit, nous nous sommes intéressés à des modèles de signaux qui permettent d'obtenir des contraintes de phase dans un but de reconstruction de celle-ci. Au chapitre précédent, nous avons proposé un modèle probabiliste qui prend en compte les à prioris que l'on peut obtenir sur la phase grâce à de tels modèles. Néanmoins, nous avons jusqu'à présent supposé que les amplitudes des sources étaient connues (cas Oracle) ou bien estimées par avance. En tout cas, elles étaient déterministes. Dans le but de mettre au point un modèle de sources complet pour une application de séparation de sources réaliste, nous allons chercher à modéliser les amplitudes des sources. En outre, nous avons constaté expérimentalement (aux chapitres 5 et 8 en particulier) que la performance des techniques de reconstruction de phase dépend de la qualité des amplitudes estimées préalablement.

C'est pourquoi, dans ce chapitre, nous nous intéressons à la modélisation de données non-négatives. En particulier, nous mettons en évidence le fait que les lois usuelles (telles que la loi de Poisson Virtanen et al. (2008); Cemgil (2009)) employées pour modéliser des données non-négatives, notamment en traitement du signal audio, ne sont pas à *queue lourde* : cela implique que de telles distributions représentent mal des valeurs éloignées de leurs modes et qui sont considérées comme "abbérantes", alors que de telles valeurs sont pourtant fréquentes en audio. Les lois α-stables Nolan (2015), en revanche, présentent cette propriété de robustesse aux valeurs abbérantes tout en conservant la propriété d'additivité des lois de Poisson. Ces distributions ont été utilisées avec succès dans des applications de traitement du signal robuste, notamment en audio Samoradnitsky et Taqqu (1994); Godsill et Kuruoglu (1999); Bassiou et al. (2013). Nous introduisons un sous-ensemble des lois stables : les distributions Positives α-stables (PαS), qui sont à support non-négatif, et peuvent donc modéliser des données non-négatives telles que des amplitudes de TFCT de signaux audio.

Après avoir introduit cette famille de lois, nous en étudions un cas particulier : la distribution de Lévy. Il s'agit du seul cas de loi PαS pour lequel nous pouvons écrire analytiquement la densité de probabilité. Nous proposons un modèle de mélange de données non-négatives appelé *Lévy NMF*, qui est appliqué à la séparation de sources non-négatives. Ceci ne porte que sur un sous-problème particulier de notre objectif final, qui consiste à séparer des sources complexes (en utilisant notamment un modèle de données non-négatives), mais nous avons trouvé intéressant d'étudier plus en détail ce modèle, car il s'agit d'une thématique de recherche active. La séparation de sources non-négatives est en effet un problème fondamental dans de nombreux domaines tels que l'imagerie par résonance magnétique Sajda et al. (2004), la reconnaissance de visage Guillamet et Vitria (2002) ou la fouille de données textuelles Pauca et al. (2004). Nous proposons des applications en débruitage de spectrogrammes audio et en séparation de sources en spectroscopie de fluorescence Liu et al. (2013).

Un des résultats forts de ce chapitre, la généralisation du filtrage de Wiener aux variables PαS, a fait l'objet d'une publication dans un rapport technique déposé dans la base de données de Télécom ParisTech Magron et al. (2016c). Le modèle de Lévy NMF ainsi que la partie expérimentale ont fait l'objet d'un article soumis dans la revue *IEEE Signal Processing Letters* Magron et al. (2017c). Ce travail ayant été mené conjointement avec Antoine Liutkus de l'INRIA Nancy, nous profitons de ce préambule pour le remercier.

Dans la section 9.1, nous introduisons les distributions PαS. Dans le section 9.2, nous présentons un modèle de mélange de données non-négatives, dont nous estimons les paramètres par différentes méthodes. Dans la section 9.3, nous fournissons une justification théorique à l'utilisation d'un filtrage de Wiener généralisé pour estimer les sources de mélanges PαS. Une validation expérimentale est proposée dans la section 9.4 avec des exemples d'application en audio et en spectroscopie de fluorescence. Enfin, nous concluons dans la section 9.5.

9.1 Distributions Positives α-stables

9.1.1 Lois stables

Les distributions α-stables Nolan (2015), notées $\mathcal{S}(\alpha, \sigma, \mu, \beta)$, sont définies comme l'ensemble des distributions sur la variable aléatoire X à valeurs dans $\mathbb{R}$ dont la fonction caractéristique est :

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \begin{cases} e^{it\mu - \sigma^\alpha |t|^\alpha (1 - i\beta\Phi sg(t))} \text{ si } \alpha \neq 1, \\ e^{it\mu - \sigma|t|(1 + i\beta\frac{2}{\pi}\log(|t|)sg(t))} \text{ si } \alpha = 1, \end{cases} \tag{9.1}$$

où $sg(t)$ désigne le signe de $t \in \mathbb{R}$, et $\Phi = \tan(\frac{\pi\alpha}{2})$. Elles dépendent de 4 paramètres :

- Un exposant caractéristique ou paramètre de forme $\alpha \in]0, 2]$ qui détermine la forme de la queue de la distribution (pour des valeurs faibles de α, la distribution est dite à queue lourde) ;
- Un paramètre d'échelle $\sigma \in]0, +\infty[$ qui mesure la dispersion de la distribution autour de son mode ;
- Un paramètre de localisation $\mu \in \mathbb{R}$;
- Un paramètre d'asymétrie $\beta \in [-1, 1]$.

La propriété de stabilité se traduit par le fait que si K variables X_k suivent une loi α-stable et sont indépendantes, alors la somme de ces variables suit également une loi α-stable. En particulier, si $X_k \sim \mathcal{S}(\alpha, \sigma_k, \mu_k, \beta_k)$, alors :

$$X = \sum_k X_k \sim \mathcal{S}(\alpha, \sigma, \mu, \beta) \text{ avec } \begin{cases} \sigma^\alpha = \sum_k \sigma_k^\alpha, \\ \mu = \sum_k \mu_k, \\ \beta = \dfrac{\sum_k \beta_k \sigma_k^\alpha}{\sum_k \sigma_k^\alpha}. \end{cases} \tag{9.2}$$

Cette propriété est utilisée pour manipuler les grandeurs X et obtenir un certain nombre de résultats. Elle donne des informations sur la densité du mélange, ce qui permet l'application de méthodes de type ML ou MAP pour l'estimation des paramètres.

Les distributions dont le paramètre d'asymétrie β est nul sont dites Symétriques α-stables (SαS) Liutkus et Badeau (2015), notée $\mathcal{S}\alpha\mathcal{S}(\alpha, \sigma, \mu) = \mathcal{S}(\alpha, \sigma, \mu, 0)$. Ces distributions sont à support dans $\mathbb{R}$ et peuvent être étendues à $\mathbb{C}$ en raisonnant sur les parties réelles et imaginaires des variables complexes considérées (lois SαS isotropes).

9.1.2 Variables à support non-négatif

Les distributions SαS ne sont pas adaptées à la modélisation de données non-négatives puisqu'elles sont à support dans $\mathbb{R}$ (ou $\mathbb{C}$). Nous proposons d'introduire un autre cas particulier de distributions α-stables qui soient à support non-négatif : les distributions Positives α-stables (PαS).

Lorsque $\beta = 1$ et $\alpha < 1$, on peut montrer Nolan (2015) que la distribution a pour support $[\mu, +\infty[$. Pour $\mu = 0$, on obtient un ensemble de distributions qui sont à support non-négatif. Une distribution PαS est donc telle que $\mathcal{P}\alpha\mathcal{S}(\sigma) = \mathcal{S}(\alpha, \sigma, 0, 1)$ avec $\alpha < 1$.

Soit $X_k \sim \mathcal{P}\alpha\mathcal{S}(\sigma_k)$ pour $k \in [\![1, K]\!]$, les X_k étant indépendants. On a, par stabilité :

$$X = \sum_k X_k \sim \mathcal{P}\alpha\mathcal{S}(\sigma) \text{ avec } \sigma^\alpha = \sum_k \sigma_k^\alpha. \tag{9.3}$$

En règle générale, nous ne pouvons pas écrire analytiquement la densité de probabilité des distributions α-stables. Ainsi, pour estimer les paramètres de modèles de mélanges comme (9.3), il faut mettre en oeuvre des méthodes de type MCMC. C'est ce qui a été effectué dans Simsekli et al. (2015) pour les distributions SαS, et pourrait par la suite être étendu au cas de distributions PαS.

9.1.3 Distribution de Lévy

Il existe trois cas de figure pour lesquels on a une expression analytique de la densité pour des distributions α-stables :

— $\alpha = 2$ et $\beta = 0$: C'est la loi normale, très populaire notamment en audio Févotte et al. (2009) ;

— $\alpha = 1$ et $\beta = 0$: C'est la loi de Cauchy, utilisée dans Liutkus et al. (2015) ;

— $\alpha = 1/2$ et $\beta = 1$: C'est la loi de Lévy.

Les deux premiers cas correspondent à des distributions SαS et ne nous intéressent donc pas ici puisqu'on cherche à modéliser des données non-négatives. Par contre, la loi de Lévy présente l'avantage d'être une distribution PαS puisque $\beta = 1$ et $\alpha < 1$. Son paramètre de localisation peut être fixé arbitrairement, et comme on cherche à modéliser des données non-négatives, on choisit $\mu = 0$. Par abus de langage, on désignera par loi de Lévy cette loi dans le cas particulier où le paramètre de localisation est nul. Sa densité de probabilité, illustrée sur la figure 9.1 est donnée par :

$$p(x|\sigma) = \begin{cases} \sqrt{\frac{\sigma}{2\pi}} \frac{1}{x^{3/2}} e^{-\frac{\sigma}{2x}} & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases} \tag{9.4}$$

Outre le fait d'avoir une densité de probabilité que l'on peut écrire analytiquement, la distribution de Lévy possède une queue lourde, comme illustré sur la figure 9.2. Cela signifie que des valeurs éloignées du mode de la distribution restent tout de même probables. Cette propriété se traduit par la robustesse d'un modèle de Lévy par rapport aux valeurs aberrantes : en effet, celles-ci modifient moins l'estimation des paramètres de la loi que pour des distributions à queue non lourde.

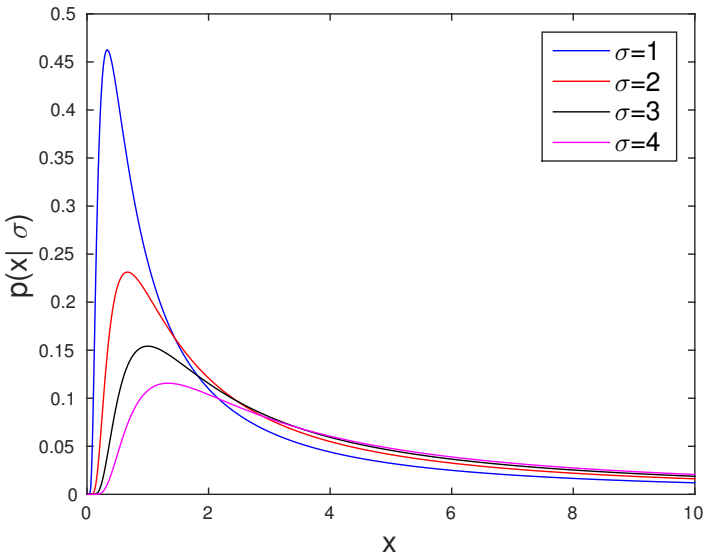


FIGURE 9.1 – Densité de probabilité d'une loi de Lévy de paramètre de localisation nul.

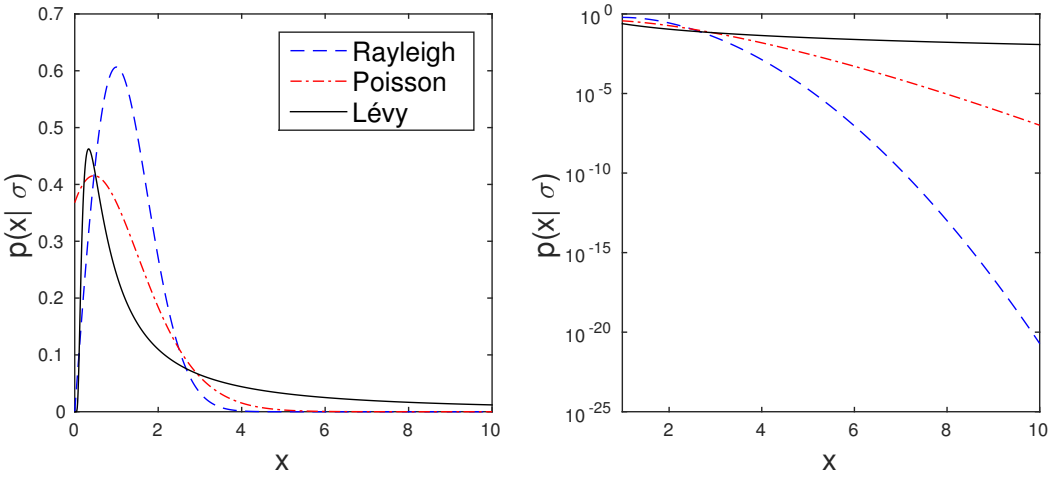


FIGURE 9.2 – Densité de probabilité pour plusieurs lois avec le même paramètre $\sigma = 1$. On observe que la distribution de Lévy possède la queue la plus lourde.

9.2 Modèle de Lévy NMF

Dans cette partie, nous proposons un modèle de mélange de sources non-négatives basé sur la distribution de Lévy.

9.2.1 Modèle de mélange

Considérons une matrice de données non-négatives $X \in \mathbb{R}_+^{F\times T}$ qui est le mélange de K sources non-négatives $X_k \in \mathbb{R}_+^{F\times T}$. On modélise les sources X_k comme des variables aléatoires matricielles suivant une loi de Lévy de paramètre $\sigma_k \in \mathbb{R}_{+*}^{F\times T}$, notée $\mathscr{L}(\sigma_k)$, $\forall k \in [\![1, K]\!]$. On étend en effet les définitions aux variables matricielles, étant donné que l'on traite tous les points TF indépendemment. Ainsi, on a, $\forall (f,t)$, $X_k \sim \mathscr{L}(\sigma_k(f,t))$.

Comme $X = \sum_k X_k$, on a, par additivité des distributions α-stables :

$$X \sim \mathscr{L}(\sigma) \text{ , avec } \sigma^{\odot 1/2} = \sum_k \sigma_k^{\odot 1/2}. \tag{9.5}$$

Comme les observations et paramètres ci-dessus sont matriciels, on propose de structurer les paramètres de dispersion par un modèle NMF :

$$\sqrt{\sigma_k(f,t)} = W(f,k)H(k,t). \tag{9.6}$$

Ainsi, la dispersion du mélange est :

$$\sigma^{\odot 1/2} = WH. \tag{9.7}$$

Ce modèle de mélange est appelé *Lévy NMF*. Un tel modèle permet donc, similairement au modèle de ISNMF, de conserver une structuration des données (qui peuvent être par exemple des spectres de notes de musique), tout en autorisant en outre aux observations de s'éloigner localement du modèle. Cela suggère donc que l'estimation des paramètres du modèle sera moins affectée par des valeurs aberrantes que dans un modèle moins robuste (comme ISNMF, qui n'est pas basé sur une distribution à queue lourde).

9.2.2 Estimation du modèle

Afin d'estimer les paramètres W et H du modèle précédemment établi, nous maximisons la vraisemblance des données (ou de façon équivalente mais plus pratique, sa log-vraisemblance). La log-vraisemblance est :

$$\begin{aligned} L(W,H) &= \sum_f \sum_t \log(p(X(f,t)|\sigma(f,t))) \\ &= \sum_f \sum_t \frac{1}{2}\log(\sigma(f,t)) - \frac{1}{2}\log(2\pi) - \frac{3}{2}\log(X(f,t)) - \frac{\sigma(f,t)}{2X(f,t)} \\ &\overset{c}{=} \frac{1}{2}\sum_f \sum_t \log([WH](f,t)^2) - \frac{[WH](f,t)^2}{X(f,t)}. \end{aligned}$$

On remarque alors que :

$$L(W,H) \overset{c}{=} -\frac{1}{2} D_{IS}([WH]^{\odot 2}, X), \tag{9.8}$$

où D_{IS} désigne la divergence d'Itakura-Saito. Ainsi, maximiser la vraisemblance des données dans le modèle Lévy NMF revient à minimiser la divergence IS entre $[WH]^{\odot 2}$ et X. En retirant les termes constants, cela revient à minimiser la fonction de coût suivante :

$$\mathcal{C}(W,H) = \sum_{f,t} \frac{[WH](f,t)^2}{X(f,t)} - 2\log([WH](f,t)). \tag{9.9}$$

Approche naïve

L'optimisation de la fonction $\mathcal{C}$ donnée par (9.9) peut-être effectuée par une approche heuristique similaire à Lee et Seung (1999) et rappelée au chapitre 2, section 2.2.4. La dérivée partielle de cette fonction par rapport à une variable θ est écrite sous forme d'une différence entre deux termes positifs :

$$\frac{\partial \mathcal{C}}{\partial \theta} = \nabla_\theta^+ - \nabla_\theta^-. \tag{9.10}$$

On considère alors la règle de mise à jour multiplicative suivante :

$$\theta \leftarrow \theta \times \frac{\nabla_\theta^-}{\nabla_\theta^+}. \tag{9.11}$$

Une telle mise à jour permet de s'assurer que le signe de θ ne change pas au cours des itérations (ce qui garantit la positivité des variables tant que l'initialisation est faite avec des valeurs positives), et que θ varie dans le sens de la décroissance (locale) de $\mathcal{C}$ (ce qui, rappelons-le, n'est pas une garantie de décroissance de la fonction de coût).

On calcule donc la dérivée partielle de $\mathcal{C}$ par rapport à $W(f,k)$:

$$\frac{\partial \mathcal{C}}{\partial W(f,k)} = \underbrace{\sum_t 2H(k,t)\frac{[WH](f,t)}{X(f,t)}}_{\nabla^+_{W(f,k)}} - \underbrace{\sum_t 2\frac{H(k,t)}{[WH](f,t)}}_{\nabla^-_{W(f,k)}}. \tag{9.12}$$

La mise à jour de $W(f,k)$ est donnée par :

$$W(f,k) \leftarrow W(f,k)\frac{\sum_t \frac{H(k,t)}{[WH](f,t)}}{\sum_t H(k,t)\frac{[WH](f,t)}{X(f,t)}}. \tag{9.13}$$

De façon tout à fait similaire, on peut calculer la mise à jour de $H(k,t)$:

$$H(k,t) \leftarrow H(k,t)\frac{\sum_f \frac{W(f,k)}{[WH](f,t)}}{\sum_f W(f,k)\frac{[WH](f,t)}{X(f,t)}}. \tag{9.14}$$

Sous forme matricielle, on obtient donc :

$$W \leftarrow W \odot \frac{[WH]^{\odot -1}H^T}{([WH]\odot X^{\odot -1})H^T}, \tag{9.15}$$

et

$$H \leftarrow H \odot \frac{W^T[WH]^{\odot -1}}{W^T([WH] \odot X^{\odot -1})}. \tag{9.16}$$

Les équations (9.15) et (9.16) fournissent des règles de mise à jour pour les variables W et H qui assurent que les variables sont actualisées dans le sens de la décroissance locale de la fonction de coût. Nous avons cependant observé expérimentalement que ces règles ne faisaient pas décroître la fonction de coût (9.9), et que celle-ci ne convergeait pas, même après un grand nombre d'itérations. D'après Badeau et al. (2010), il peut être souhaitable d'utiliser les règles de mise à jour suivantes :

$$W \leftarrow W \odot \left(\frac{[WH]^{\odot -1} H^T}{([WH] \odot X^{\odot -1}) H^T} \right)^{\odot \eta}, \tag{9.17}$$

et

$$H \leftarrow H \odot \left(\frac{W^T[WH]^{\odot -1}}{W^T([WH] \odot X^{\odot -1})} \right)^{\odot \eta}, \tag{9.18}$$

où η est un exposant qui, choisi judicieusement, augmente significativement la vitesse de convergence. Nous proposons par la suite d'utiliser une nouvelle technique d'optimisation, qui permet de se ramener à ce type de règles.

Approche Majoration-Minimisation

L'algorithme Majoration-Minimisation (MM) Hunter et Lange (2004) fournit un cadre théorique qui permet de minimiser la fonction de coût $\mathcal{C}(\theta)$. Le principe de l'algorithme MM (*cf.* chapitre 2, section 2.2.4) est de majorer la fonction de coût $\mathcal{C}(\theta)$ par une fonction auxiliaire $G(\theta, \bar{\theta})$, telle que $G(\bar{\theta}, \bar{\theta}) = \mathcal{C}(\bar{\theta})$. À $\bar{\theta}$ fixé, la minimisation de G par rapport à θ produit une mise à jour de ce paramètre qui conduit, par construction de G, à la décroissance de $\mathcal{C}$.

Pour construire la fonction auxiliaire G, on s'inspire de la démarche conduite dans Févotte et Idier (2011); Févotte (2011), qui consiste à écrire $\mathcal{C}$ comme somme de fonctions convexes et concaves. Les termes convexes sont majorés via l'inégalité de Jensen, et les termes concaves sont majorés par leurs tangentes.

Nous détaillons ici la construction de la fonction auxiliaire en un point $\overline{W}(f,k)$ (cette notation ne désigne pas ici la conjugaison complexe, puisqu'on travaille sur des données non-négatives). Un calcul semblable permet d'obtenir la fonction auxiliaire en un point $\overline{H}(k,t)$ pour obtenir la règle de mise à jour sur H.

On s'intéresse tout d'abord au premier terme dans l'expression de $\mathcal{C}$ donnée par l'équation (9.9) :

$$\mathcal{C}_1(W,H) = \sum_{f,t} \frac{[WH](f,t)^2}{X(f,t)}. \tag{9.19}$$

On note $\overline{V}(f,t) = [\overline{W}H](f,t)$ et $\rho_k(f,t) = \dfrac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)}$. On peut alors introduire ces termes dans la somme ci-dessous :

$$\begin{aligned}
[WH](f,t)^2 &= \left(\sum_k W(f,k)H(k,t) \right)^2 \\
&= \left(\sum_k \rho_k(f,t) \frac{W(f,k)H(k,t)}{\rho_k(f,t)} \right)^2.
\end{aligned}$$

Les termes ρ_k jouent le rôle de poids et ont pour propriété de sommer à l'unité (par définition de $\overline{V}$). Comme la fonction carré est convexe et que les poids somment à l'unité, on peut appliquer l'inégalité de Jensen :

$$\begin{aligned}[WH](f,t)^2 &\leq \sum_k \rho_k(f,t)\left(\frac{W(f,k)H(k,t)}{\rho_k(f,t)}\right)^2 \\ &\leq \sum_k \frac{W(f,k)^2H(k,t)^2}{\overline{W}(f,k)H(k,t)}\overline{V}(f,t) \\ &\leq \overline{V}(f,t)\sum_k \frac{W(f,k)^2H(k,t)}{\overline{W}(f,k)}.\end{aligned}$$

Ainsi, on obtient finalement la majoration suivante du premier terme dans la fonction de coût :

$$\mathcal{C}_1(W,H) \leq \sum_{f,t}\frac{\overline{V}(f,t)}{X(f,t)}\sum_k \frac{W(f,k)^2H(k,t)}{\overline{W}(f,k)}. \tag{9.20}$$

On procède à présent au même type de majoration sur le deuxième terme de la fonction de coût :

$$-2\log([WH](f,t)) = -2\log\left(\sum_k \rho_k(f,t)\frac{W(f,k)H(k,t)}{\rho_k(f,t)}\right). \tag{9.21}$$

La fonction $-2\log$ étant convexe (le logarithme étant concave), on applique de nouveau l'inégalité de convexité et on obtient :

$$\begin{aligned}-2\log([WH](f,t)) &\leq \sum_k \rho_k(f,t)\left(-2\log\frac{W(f,k)H(k,t)}{\rho_k(f,t)}\right) \\ &\leq -2\sum_k \frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)}\log\left(W(f,k)\frac{\overline{V}(f,t)}{\overline{W}(f,k)}\right).\end{aligned}$$

Cela nous mène à une majoration du deuxième terme de la fonction de coût :

$$\mathcal{C}_2(W,H) \leq -2\sum_{f,t}\sum_k \frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)}\log\left(W(f,k)\frac{\overline{V}(f,t)}{\overline{W}(f,k)}\right). \tag{9.22}$$

En combinant (9.20) et (9.22), on obtient finalement une majorante de la fonction de coût : $\mathcal{C}(W,H) \leq G(W,H,\overline{W})$ avec :

$$G(W,H,\overline{W}) = \sum_{f,t,k}\frac{\overline{V}(f,t)H(k,t)}{X(f,t)\overline{W}(f,k)}W(f,k)^2 - 2\frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)}\log\left(W(f,k)\frac{\overline{V}(f,t)}{\overline{W}(f,k)}\right). \tag{9.23}$$

Il est par ailleurs aisé de constater que $\mathcal{C}(W,H) = G(W,H,W)$ (c'est obtenu par construction). G est donc une fonction auxiliaire à $\mathcal{C}$.

On applique alors la méthode MM : on minimise la fonction auxiliaire G par rapport à la variable $W(f,k)$ à H et $\overline{W}$ fixés. La dérivée partielle est :

$$\frac{\partial G}{\partial W(f,k)}(W,H,\overline{W}) = \sum_t 2\frac{\overline{V}(f,t)H(k,t)}{X(f,t)\overline{W}(f,k)}W(f,k) - 2\frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)}\frac{1}{W(f,k)}, \tag{9.24}$$

et l'annulation de cette dérivée conduit à :

$$\frac{W(f,k)}{\overline{W}(f,k)}\sum_t \frac{\overline{V}(f,t)H(k,t)}{X(f,t)} = \frac{\overline{W}(f,k)}{W(f,k)}\sum_t \frac{H(k,t)}{\overline{V}(f,t)}, \tag{9.25}$$

soit, en remplaçant $\overline{V}$ par sa définition $[\overline{W}H]$:

$$W(f,k)^2 = \overline{W}(f,k)^2 \frac{\sum_t \frac{H(k,t)}{[\overline{W}H](f,t)}}{\sum_t H(k,t)\frac{[\overline{W}H](f,t)}{X(f,t)}}, \tag{9.26}$$

et donc (en ne gardant que la solution positive, la seule qui nous intéresse) :

$$W(f,k) = \overline{W}(f,k)\left(\frac{\sum_t \frac{H(k,t)}{[\overline{W}H](f,t)}}{\sum_t H(k,t)\frac{[\overline{W}H](f,t)}{X(f,t)}}\right)^{1/2}. \tag{9.27}$$

On retrouve une mise à jour similaire à celle de l'approche naïve, exceptée une puissance 1/2 supplémentaire. Comme on le faisait remarquer dans la section précédente, nous avons trouvé un exposant η qui permette d'assurer la décroissance de la fonction de coût. En tenant compte du fait que la variable $\overline{W}$ est ensuite actualisée en prenant la nouvelle valeur de W, on peut synthétiser la mise à jour sur W sous forme matricielle :

$$W \leftarrow W \odot \left(\frac{[WH]^{\odot -1}H^T}{([WH]\odot X^{\odot -1})H^T}\right)^{\odot 1/2}. \tag{9.28}$$

Par un calcul similaire (on introduit la variable auxiliaire $\overline{H}$ et on utilise les mêmes inégalités de convexité), on construit une fonction auxiliaire à $\mathcal{C}$ en un point $\overline{H}$, et on peut obtenir la mise à jour suivante :

$$H \leftarrow H \odot \left(\frac{W^T[WH]^{\odot -1}}{W^T([WH]\odot X^{\odot -1})}\right)^{\odot 1/2}. \tag{9.29}$$

Les équations (9.28) et (9.29) fournissent donc les règles de mise à jour pour l'estimation des paramètres W et H du modèle de Lévy NMF par l'approche MM. Par construction de ces règles, nous avons la garantie de la monotonie de la fonction de coût, contrairement au cas de l'approche naïve.

Approche Majoration-Égalisation

L'approche MM présentée ci-dessus est un moyen d'obtenir des règles de mise à jour pour lesquelles on a une garantie de décroissance de la fonction de coût. Néanmoins, il y a d'autres façons d'obtenir de telles mises à jour qu'en minimisant la fonction majorante : en effet, toute mise à jour vérifiant $\mathcal{C}(\theta^{(it+1)}) \leq \mathcal{C}(\theta^{(it)})$ convient. L'approche Majoration-Égalisation (ME) consiste à choisir $\theta^{(it+1)}$ tel que $G(\theta^{(it+1)}, \theta^{(it)}) = G(\theta^{(it)}, \theta^{(it)})$. Cette approche est illustrée sur la figure 9.3.

Par construction de la fonction auxiliaire :

$$\mathcal{C}(\theta^{(it+1)}) \leq G(\theta^{(it+1)}, \theta^{(it)}), \tag{9.30}$$

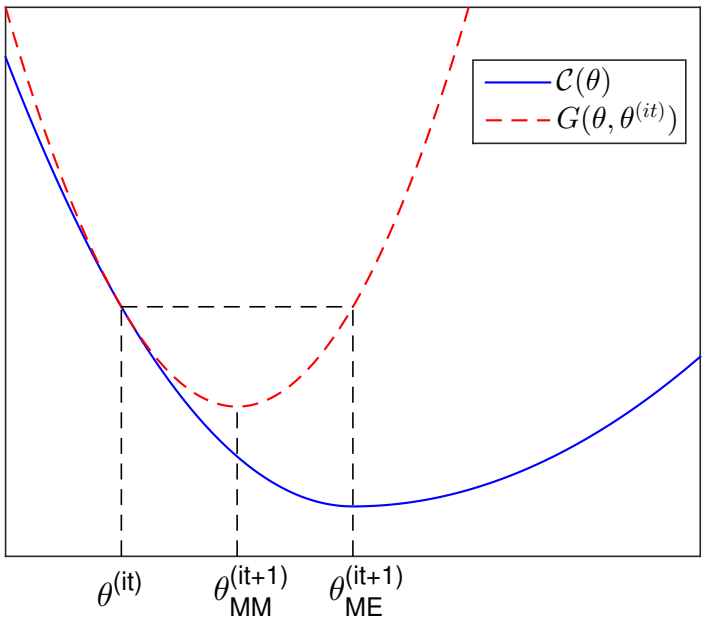


FIGURE 9.3 – Illustration des approches MM et ME : la fonction de coût $\mathcal{C}(\theta)$ est majorée à l'itération (it) par la fonction auxiliaire $G(\theta, \theta^{(it)})$. À partir de celle-ci, on peut obtenir les nouveaux estimateurs de θ, par minimisation (MM) ou par égalisation (ME).

et donc, en appliquant l'approche ME, on a :

$$\mathcal{C}(\theta^{(it+1)}) \leq G(\theta^{(it+1)}, \theta^{(it)}) = G(\theta^{(it)}, \theta^{(it)}) = \mathcal{C}(\theta^{(it)}), \tag{9.31}$$

ce qui prouve que la fonction objectif $\mathcal{C}$ est décroissante sous cette mise à jour. Nous proposons donc d'appliquer cette méthode pour obtenir une nouvelle procédure d'estimation des paramètres. En effet, cette approche peut être intéressante car il est montré dans FÉVOTTE et IDIER (2011) que les mises à jour ME convergent plus rapidement que les mises à jour MM dans le cas de la NMF avec β-divergence. Il est possible que pour notre cas d'application également, de telles mises à jour accélèrent la convergence de l'algorithme. Pour l'estimation de W (encore une fois, celle de H est équivalente, nous ne la détaillons donc pas), on est amené à résoudre l'équation suivante :

$$G(W, H, \overline{W}) = G(\overline{W}, H, \overline{W}). \tag{9.32}$$

En utilisant la séparabilité des variables dans l'expression de G, c'est équivalent à résoudre, pour tout (f, k) :

$$\begin{aligned}\sum_t \frac{\overline{V}(f,t)H(k,t)}{X(f,t)\overline{W}(f,k)} W(f,k)^2 - 2\frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)} \log\left(W(f,k)\frac{\overline{V}(f,t)}{\overline{W}(f,k)}\right) \\ = \sum_t \frac{\overline{V}(f,t)H(k,t)}{X(f,t)\overline{W}(f,k)} \overline{W}(f,k)^2 - 2\frac{\overline{W}(f,k)H(k,t)}{\overline{V}(f,t)} \log\left(\overline{W}(f,k)\frac{\overline{V}(f,t)}{\overline{W}(f,k)}\right),\end{aligned} \tag{9.33}$$

que l'on peut simplifier et réécrire :

$$\left(\frac{W(f,k)^2}{\overline{W}(f,k)} - \overline{W}(f,k)\right) \sum_t \frac{\overline{V}(f,t)H(k,t)}{X(f,t)} = 2\overline{W}(f,k) \log\left(\frac{W(f,k)}{\overline{W}(f,k)}\right) \sum_t \frac{H(k,t)}{\overline{V}(f,t)}. \tag{9.34}$$

En notant :

$$a_W(f,k) = \frac{\sum_t \frac{H(k,t)}{\overline{V}(f,t)}}{\sum_t \frac{\overline{V}(f,t)H(k,t)}{X(f,t)}}, \tag{9.35}$$

on a alors :

$$\frac{W(f,k)^2}{\overline{W}(f,k)^2} - 1 = 2a_W(f,k)\log\left(\frac{W(f,k)}{\overline{W}(f,k)}\right). \tag{9.36}$$

Dans le but d'obtenir des règles de mise à jour multiplicatives, on cherche, sans perte de généralité, les solutions sous la forme $W = \overline{W}a_W^{\odot\eta_W}$ où $\eta_W \geq 0$ est à déterminer. On peut alors réécrire l'équation ci-dessus :

$$a_W(f,k)^{2\eta_W(f,k)} = 1 + 2\eta_W(f,k)a_W(f,k)\log a_W(f,k). \tag{9.37}$$

On obtient exactement la même équation en raisonnant sur H plutôt que sur W : on utilise alors le paramètre

$$a_H(k,t) = \frac{\sum_f \frac{W(f,k)}{[W\overline{H}](f,t)}}{\sum_f W(f,k)\frac{[W\overline{H}](f,t)}{X(f,t)}}, \tag{9.38}$$

et un exposant η_H. On peut donc écrire, sous forme matricielle, l'équation définissant l'exposant η_θ (avec $\theta = W$ ou H) qui mène à l'estimation ME :

$$a_\theta^{2\eta_\theta} = 1 + 2\eta_\theta \odot a_\theta \odot \log a_\theta. \tag{9.39}$$

Comme pour les deux matrices W et H, et pour toutes les entrées de ces matrices, on est amené à résoudre l'équation ci-dessus, on choisit, par souci de clarté, de considérer une variable (W ou H) et une entrée de cette variable, et on s'affranchit des notations d'indices et de matrices. On cherche donc à résoudre sur $\mathbb{R}_+$ l'équation suivante :

$$a^{2\eta} = 1 + 2\eta a\log a, \tag{9.40}$$

ce qui autrement dit, revient à trouver un (ou plusieurs) zéro de la fonction définie sur $\mathbb{R}_+$ par :

$$d(\eta) = a^{2\eta} - 1 - 2\eta a\log a. \tag{9.41}$$

La fonction d est illustrée sur la figure 9.4.

On considère le cas non-trivial où $a \neq 1$. Il est aisé de constater que $d(0) = 0$ et que $\lim_{\eta\to+\infty} d(\eta) = +\infty$. Un calcul élémentaire de dérivée montre que cette fonction possède un minimum atteint en $\eta = 1/2$, que cette valeur minimale est négative, et que d est strictement décroissante sur $[0, 1/2]$ et croissante sur $[1/2, +\infty[$. Ainsi, lorsque $a \neq 1$, le théorème des valeurs intermédiaires garantit qu'il existe exactement 2 solutions à l'équation (9.40). Une de ces solutions est triviale et correspond à $\eta = 0$, ce qui signifie que l'on a alors $\theta = \bar{\theta}$: la mise à jour ne modifie pas la valeur des paramètres, ce qui ne nous intéresse pas ici. Nous cherchons donc à obtenir l'autre solution de l'équation. Les solutions de cette équation peuvent être écrites de façon exacte par le biais des fonctions de Lambert Corless et al. (1996), mais cela ne nous fournit pas d'expression analytique simple de la solution.

Nous proposons donc d'approcher cette solution en utilisant un développement limité. Au voisinage de $a = 1$, on a :

$$a^{2\eta} = (1 + (a-1))^{2\eta} \approx 1 + 2\eta(a-1) + \frac{2\eta(2\eta-1)}{2}(a-1)^2. \tag{9.42}$$

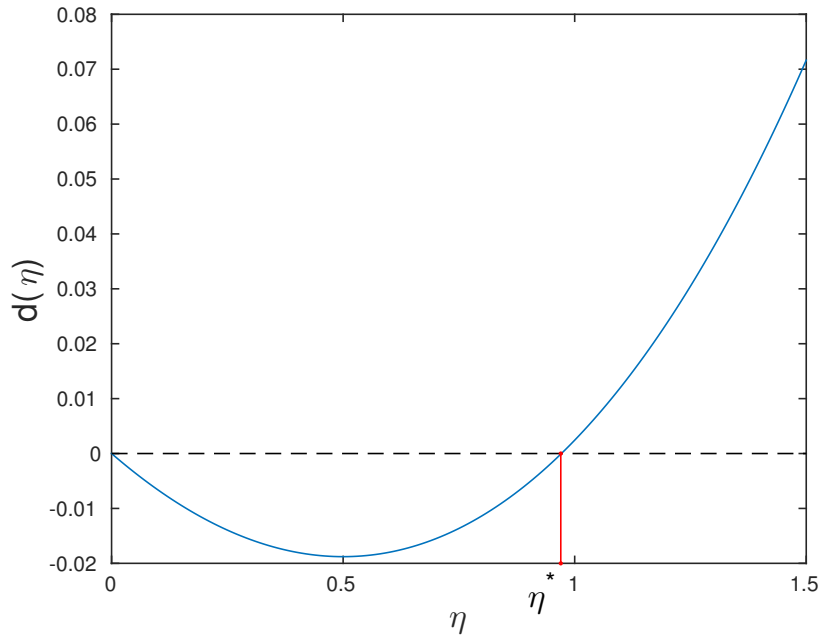


FIGURE 9.4 – Fonction d définie par (9.41) pour $a = 1.2$.

En réinjectant cette expression dans (9.40), on obtient :

$$1 + 2\eta(a-1) + \frac{2\eta(2\eta-1)}{2}(a-1)^2 = 1 + 2\eta a \log a, \tag{9.43}$$

et, en simplifiant par 2η (que l'on suppose non-nul pour s'affranchir de cette solution triviale) :

$$\begin{aligned}(a-1) + \frac{(2\eta-1)}{2}(a-1)^2 &= a\log a \\ (2\eta-1)(a-1)^2 &= 2(a\log a - (a-1)) \\ 2\eta - 1 &= 2\frac{a\log a - (a-1)}{(a-1)^2},\end{aligned}$$

soit finalement l'approximation de l'exposant optimal suivante :

$$\eta^* = \frac{1}{2} + \frac{a\log a - (a-1)}{(a-1)^2}. \tag{9.44}$$

Cette expression permet donc d'obtenir une approximation de l'exposant à utiliser dans les règles de mise à jour de Lévy NMF par l'approche ME.

Nous souhaitons examiner la validité de cette approximation. Nous calculons, pour différentes valeurs de a, la valeur de η^*. Nous calculons également une solution approchée $\tilde{\eta}$ de l'équation (9.40) par la fonction Matlab `fzero`. Cette valeur est considérée comme une référence (à la précision de Matlab près, elle annule l'équation (9.40)). Sur la figure 9.5, nous traçons les exposants approché et de référence, ainsi que l'erreur relative entre les deux.

Nous constatons tout d'abord que l'erreur d'approximation est relativement faible (de l'ordre de 3 %) pour des valeurs de a proches de 1. Cela signifie que plus on est proche de la convergence (ce qui correspond à $a = 1$), plus l'approximation que nous avons proposée est de bonne qualité. Par ailleurs, nous constatons que l'approximation η^* semble être inférieure à la valeur de référence $\tilde{\eta}$. Ce constat est intéressant car il traduit la propriété suivante : la mise

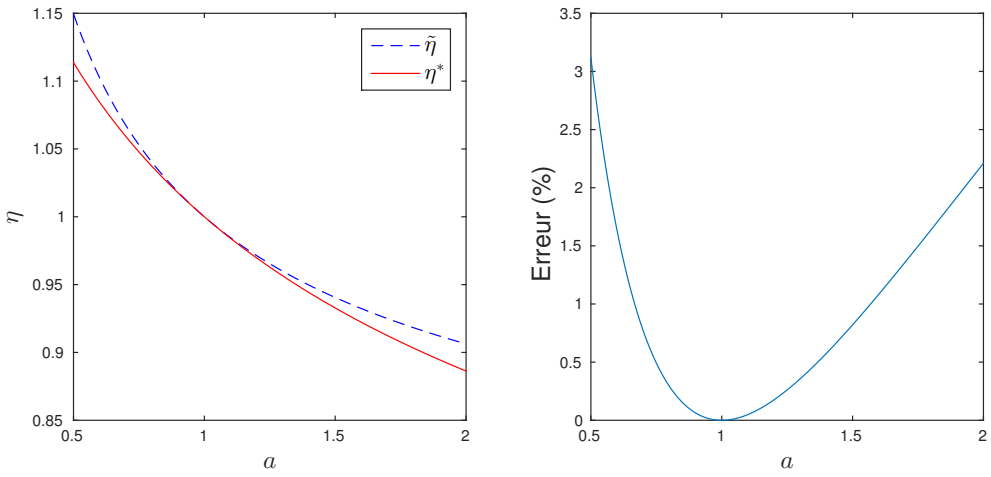


FIGURE 9.5 – Estimations de l'exposant optimal dans l'approche ME (gauche), et erreur relative (en %) entre référence et approximation (droite).

à jour obtenue avec l'exposant η^* est "moins" éloignée de la valeur courante du paramètre qu'avec l'exposant $\tilde{\eta}$. Autrement dit, la variation du paramètre θ est moins grande lors d'une mise à jour utilisant cette approximation, ce qui conduit à une garantie sur la décroissance de la fonction de coût. Considérons en effet le cas $a > 1$. On a alors :

$$\theta^{(it)} \leq \theta^{(it)} a^{\eta^*} \leq \theta^{(it)} a^{\tilde{\eta}}, \tag{9.45}$$

soit

$$\theta^{(it)} \leq \theta_{ME}^{(it+1)*} \leq \theta_{ME}^{(it+1)}, \tag{9.46}$$

et par ailleurs, $G(\theta^{(it)}, \theta^{(it)}) = G(\theta_{ME}^{(it+1)}, \theta^{(it)})$. Donc, en utilisant la convexité de la fonction G, on a :

$$G(\theta_{ME}^{(it+1)*}, \theta^{(it)}) \leq G(\theta^{(it)}, \theta^{(it)}) = \mathcal{C}(\theta^{(it)}). \tag{9.47}$$

Enfin, on a par construction de la fonction auxiliaire $\mathcal{C}(\theta_{ME}^{(it+1)*}) \leq G(\theta_{ME}^{(it+1)*}, \theta^{(it)})$, ce qui conduit à :

$$\mathcal{C}(\theta_{ME}^{(it+1)*}) \leq \mathcal{C}(\theta^{(it)}), \tag{9.48}$$

ce qui montre la décroissance de la fonction de coût avec cette mise à jour. La démonstration est similaire lorsque $a < 1$. En d'autres termes, tant que l'on utilise un exposant η^* qui est inférieur à l'exposant $\tilde{\eta}$, on a toujours une garantie de décroissance de la fonction de coût. Il est donc intéressant de noter, à partir de la figure 9.5, que l'approximation que nous avons proposée respecte cette condition. Néanmoins, il s'agit d'une observation empirique sur une échelle réduite, et nous n'avons aucune garantie que cette propriété reste vraie pour toute valeur de a.

On s'attend donc à constater la décroissance de la fonction de coût à l'approche de la convergence (lorsque a se rapproche de 1), mais celle-ci n'est pas globalement garantie, notamment pour les premières itérations. On pourrait alors imaginer un algorithme hybride qui utilise les règles de mises à jour MM pour les premières itérations, puis les règles ME ensuite.

Méthode	Naïve	MM	ME
η_θ	1	$\frac{1}{2}$	$\frac{1}{2} + \frac{a_\theta \odot \log a_\theta - a_\theta + 1}{(a_\theta - 1)^{\odot 2}}$

TABLEAU 9.1 – Exposant dans les règles de mise à jour pour l'estimation du modèle de Lévy NMF selon la méthode choisie.

Algorithme 11 Lévy NMF (Approche Majoration-Minimisation)

Entrées :

Données $X \in \mathbb{R}_+^{F \times T}$,

Matrices initiales $W \in \mathbb{R}_+^{F \times K}$ et $H \in \mathbb{R}_+^{K \times T}$,

Nombre d'itérations N_{it}.

pour $j = 1$ à N_{it} **faire**

$$W \leftarrow W \odot \left(\frac{[WH]^{\odot -1} H^T}{([WH] \odot X^{\odot -1}) H^T} \right)^{\odot 1/2}.$$

$$H \leftarrow H \odot \left(\frac{W^T [WH]^{\odot -1}}{W^T ([WH] \odot X^{\odot -1})} \right)^{\odot 1/2}.$$

Normalisation de W et H.

fin pour

Sortie :

$W \in \mathbb{R}_+^{F \times K}$ et $H \in \mathbb{R}_+^{K \times T}$.

9.2.3 Algorithmes de Lévy NMF

Dans les trois cas que nous avons étudiés, nous avons abouti à des règles de mise à jour des paramètres W et H qui sont multiplicatives, et qui possèdent une forme similaire :

$$\theta \leftarrow \theta \odot a_\theta^{\odot \eta_\theta}, \tag{9.49}$$

les paramètres a_θ étant donnés par :

$$a_W = \frac{[WH]^{\odot -1} H^T}{([WH] \odot X^{\odot -1}) H^T} \text{ et } a_H = \frac{W^T [WH]^{\odot -1}}{W^T ([WH] \odot X^{\odot -1})}. \tag{9.50}$$

L'exposant η_θ est donné, selon la méthode choisie, par le tableau 9.1. On présente dans l'algorithme 11 la procédure itérative détaillée d'estimation des paramètres du modèle de Lévy NMF par l'approche MM. La forme des autres algorithmes est donc la même, il suffit de modifier l'exposant dans les règles de mise à jour conformément aux valeurs données dans le tableau 9.1.

9.2.4 Interprétation des mises à jour

Nous proposons une interprétation des règles de mise à jour obtenues dans le cadre de l'approche MM. Plaçons-nous dans le cas $K = 1$ et supposons que $W(f) = 1$ pour tout f. La mise à jour sur H dans l'algorithme de Lévy NMF (par approche MM) s'écrit alors :

$$H(t) \leftarrow \sqrt{\frac{F}{\sum_f \frac{1}{X(f,t)}}}, \tag{9.51}$$

et, à titre de comparaison, nous donnons également la mise à jour dans le cas de l'algorithme de ISNMF Févotte et al. (2009) :

$$H(t) \leftarrow \frac{1}{F} \sum_f X(f,t). \tag{9.52}$$

On constate donc que dans l'algorithme de ISNMF, la mise à jour sur H revient à effectuer une moyenne arithmétique sur les lignes de X, alors que dans l'algorithme de Lévy NMF, on effectue une opération de moyenne harmonique (élevée à la puissance $1/2$).

Considérons que X est de dimensions 10×10 et ne contient que des 1, à l'exception d'une entrée (à la t-ième ligne) qui vaut 10^8. On cherche à factoriser X grâce à un modèle NMF de rang 1 (c'est-à-dire tel que tous les coefficients de W et H valent 1, à la normalisation près) : la valeur 10^8 est donc considérée comme aberrante dans un tel modèle. On souhaite mettre au point une technique d'estimation du modèle robuste à cette valeur. Supposons qu'à une itération donnée, on a $W(f) = 1$ pour tout f, et on cherche à estimer $H(t)$. La mise à jour dans le modèle de Lévy serait, d'après (9.51), $H(t) \leftarrow 1.05$, alors que dans le modèle de ISNMF, d'après (9.52), elle serait $H(t) \leftarrow 10^7$.

Cet exemple illustre la propriété de robustesse aux valeurs aberrantes de l'algorithme de Lévy NMF (construit sur une distribution à queue lourde) comparativement à l'algorithme de ISNMF (construit sur un modèle gaussien, donc à queue non-lourde).

9.3 Estimateur des sources

L'algorithme de Lévy NMF permet d'estimer les paramètres des sources. On cherche à présent à obtenir un estimateur des sources construit à partir de ces paramètres. Ce qui suit est valable pour toute loi positive α-stable, et pas uniquement la distribution de Lévy.

Afin de clarifier la démonstration qui suit, on considère dans un premier temps un mélange de 2 sources uniquement ($K = 2$) et une seule entrée matricielle. On note alors, pour un jeu d'indices (f,t), $x = X(f,t)$, $s_1 = X_1(f,t)$ et $s_2 = X_2(f,t)$, de sorte à ce que $x = s_1 + s_2$. Le résultat sera par la suite étendu à K sources et à des variables matricielles.

9.3.1 Somme de 2 variables

Soit $\alpha \in]0,1[$. Considérons deux variables s_1 et s_2 indépendantes et distribuées selon une loi PαS de paramètres de dispersion σ_1 et σ_2 respectivement. Nous allons nous inspirer de la démonstration qui est conduite dans Badeau et Liutkus (2014) (pour des variables aléatoires SαS) pour obtenir les estimateurs des sources donnés par l'espérance à postériori des sources sachant les observations, que nous notons $\hat{s}_1$ et $\hat{s}_2$. L'estimateur de s_1 est donc :

$$\hat{s}_1 = \mathbb{E}_{s_1|x}(s_1), \tag{9.53}$$

et par ailleurs, cette espérance est définie si et seulement si la fonction caractéristique de $s_1|x$, $\varphi_{s_1|x}(t_1) = \mathbb{E}_{s_1|x}(e^{it_1 s_1})$, est différentiable en $t_1 = 0$, auquel cas on a (par exemple d'après Badeau et Liutkus (2014)) :

$$\mathbb{E}_{s_1|x}(s_1) = \frac{1}{i} \frac{d\varphi_{s_1|x}}{dt_1}(0). \tag{9.54}$$

Ainsi, en déterminant la fonction caractéristique de $s_1|x$ puis la dérivée de celle-ci, on pourra en déduire l'estimateur de s_1. L'intérêt de cette méthode est d'utiliser les fonctions caractéristiques de lois stables, alors que l'on ne sait pas exprimer (dans le cas général) leurs densités de probabilités.

Étape 1 : Fonction caractéristique de $s_1|x$. Tout d'abord, par stabilité des processus PαS, x suit une loi PαS de paramètre σ tel que $\sigma^\alpha = \sigma_1^\alpha + \sigma_2^\alpha$. La fonction caractéristique d'une telle variable est donnée par :

$$\varphi_x(t_x) = e^{-\sigma^\alpha |t_x|^\alpha + i\Phi\sigma^\alpha |t_x|^\alpha sg(t_x)}, \tag{9.55}$$

où $\Phi > 0$ comme rappelé dans la section 9.1. La fonction caractéristique du vecteur (s_1, x) est :

$$\begin{aligned}\varphi_{s_1,x}(t_1, t_x) &= \mathbb{E}(e^{i(t_1 s_1 + t_x x)}) \\ &= \mathbb{E}(e^{i(t_1 s_1 + t_x (s_1 + s_2))}) \\ &= \mathbb{E}(e^{i((t_1 + t_x) s_1 + t_x s_2)}) \\ &= \varphi_{s_1,s_2}(t_1 + t_x, t_x).\end{aligned}$$

Or, les variables s_1 et s_2 étant indépendantes, $\varphi_{s_1,s_2}(t_1 + t_x, t_x) = \varphi_{s_1}(t_1 + t_x)\varphi_{s_2}(t_x)$. On obtient donc :

$$\varphi_{s_1,x}(t_1, t_x) = e^{-\sigma_1^\alpha |t_1 + t_x|^\alpha - \sigma_2^\alpha |t_x|^\alpha + i\Phi(\sigma_1^\alpha |t_1 + t_x|^\alpha sg(t_1 + t_x) + \sigma_2^\alpha |t_x|^\alpha sg(t_x))}. \tag{9.56}$$

Afin d'obtenir la fonction caractéristique de $s_1|x$ à partir de celle du vecteur (s_1, x), nous allons utiliser un résultat issu de Samoradnitsky et Taqqu (1994) (eq. (5.1.7) p. 226) :

$$\varphi_{s_1|x}(t_1) = \frac{\displaystyle\int_{\mathbb{R}} \varphi_{s_1,x}(t_1, t_x) e^{-it_x x} dt_x}{\displaystyle\int_{\mathbb{R}} \varphi_{s_1,x}(0, t_x) e^{-it_x x} dt_x}. \tag{9.57}$$

Étape 2 : Dérivation de la fonction caractéristique. La dérivée de la fonction caractéristique de $s_1|x$ est donnée par :

$$\frac{d\varphi_{s_1|x}}{dt_1}(t_1) = \frac{\displaystyle\int_{\mathbb{R}} \frac{\partial \varphi_{s_1,x}}{\partial t_1}(t_1, t_x) e^{-it_x x} dt_x}{\displaystyle\int_{\mathbb{R}} \varphi_{s_1,x}(0, t_x) e^{-it_x x} dt_x}, \tag{9.58}$$

que l'on applique en $t_1 = 0$:

$$\frac{d\varphi_{s_1|x}}{dt_1}(0) = \frac{\displaystyle\int_{\mathbb{R}} \frac{\partial \varphi_{s_1,x}}{\partial t_1}(0, t_x) e^{-it_x x} dt_x}{\displaystyle\int_{\mathbb{R}} \varphi_{s_1,x}(0, t_x) e^{-it_x x} dt_x}. \tag{9.59}$$

Notons que la relation (9.58) est valable à condition qu'il soit possible de dériver sous le signe $\int$, ce que nous montrons ci-après. On distingue deux cas à partir de l'équation (9.56) :

— Si $t_1 > -t_x$, on a :

$$\varphi_{s_1,x}(t_1, t_x) = e^{-\sigma_1^\alpha (t_1 + t_x)^\alpha - \sigma_2^\alpha |t_x|^\alpha + i\Phi(\sigma_1^\alpha (t_1 + t_x)^\alpha + \sigma_2^\alpha |t_x|^\alpha sg(t_x))}, \tag{9.60}$$

et donc :

$$\frac{\partial \varphi_{s_1,x}}{\partial t_1}(t_1, t_x) = [-\alpha\sigma_1^\alpha (t_1 + t_x)^{\alpha-1} + i\alpha\Phi\sigma_1^\alpha (t_1 + t_x)^{\alpha-1}]\varphi_{s_1,x}(t_1, t_x) \tag{9.61}$$

— Si $t_1 < -t_x$, on a :

$$\varphi_{s_1,x}(t_1,t_x) = e^{-\sigma_1^\alpha(-t_1-t_x)^\alpha-\sigma_2^\alpha|t_x|^\alpha+i\Phi(-\sigma_1^\alpha(-t_1-t_x)^\alpha+\sigma_2^\alpha|t_x|^\alpha sg(t_x))}, \tag{9.62}$$

et donc :

$$\frac{\partial\varphi_{s_1,x}}{\partial t_1}(t_1,t_x) = [\alpha\sigma_1^\alpha(-t_1-t_x)^{\alpha-1} + i\alpha\Phi\sigma_1^\alpha(-t_1-t_x)^{\alpha-1}]\varphi_{s_1,x}(t_1,t_x). \tag{9.63}$$

On obtient dans les deux cas la même expression de la dérivée :

$$\frac{\partial\varphi_{s_1,x}}{\partial t_1}(t_1,t_x) = \alpha\sigma_1^\alpha[-(t_1+t_x)|t_1+t_x|^{\alpha-2} + i\Phi|t_1+t_x|^{\alpha-1}]\varphi_{s_1,x}(t_1,t_x), \tag{9.64}$$

qui, appliquée en $t_1 = 0$, conduit à :

$$\frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x) = \alpha\sigma_1^\alpha[-t_x|t_x|^{\alpha-2} + i\Phi|t_x|^{\alpha-1}]\varphi_{s_1,x}(0,t_x), \tag{9.65}$$

avec :

$$\varphi_{s_1,x}(0,t_x) = e^{-\sigma_1^\alpha|t_x|^\alpha-\sigma_2^\alpha|t_x|^\alpha+i\Phi(\sigma_1^\alpha|t_x|^\alpha sg(t_x)+\sigma_2^\alpha|t_x|^\alpha sg(t_x))} \tag{9.66}$$

$$= e^{-(\sigma_1^\alpha+\sigma_2^\alpha)|t_x|^\alpha+i\Phi(\sigma_1^\alpha+\sigma_2^\alpha)|t_x|^\alpha sg(t_x)}. \tag{9.67}$$

Revenons à présent sur la dérivation sous l'intégrale (9.59). Pour montrer qu'il est possible d'écrire cette dérivée, nous allons prouver que :

$$\frac{\partial\int_{\mathbb{R}_+}\varphi_{s_1,x}(t_1,t_x)e^{-it_xx}dt_x}{\partial t_1}(t_1=0) = \int_{\mathbb{R}_+}\frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x)e^{-it_xx}dt_x, \tag{9.68}$$

$$\frac{\partial\int_{\mathbb{R}_-}\varphi_{s_1,x}(t_1,t_x)e^{-it_xx}dt_x}{\partial t_1}(t_1=0) = \int_{\mathbb{R}_-}\frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x)e^{-it_xx}dt_x. \tag{9.69}$$

Montrons l'équation (9.68) (la preuve est similaire pour (9.69)). Il nous faut trouver une borne supérieure de $\left|\frac{\partial\varphi_{s_1,x}}{\partial t_1}(t_1,t_x)e^{-it_xx}\right|$. Comme $\alpha \in]0,1[$, on montre avec les équations (9.56) et (9.64) que :

$$\forall t_1\in\mathbb{R}_+,\ \forall t_x\in\mathbb{R}_+\backslash\{0\},\ \left|\frac{\partial\varphi_{s_1,x}}{\partial t_1}(t_1,t_x)e^{-it_xx}\right| \le g(t_1+t_x)h(t_x) \le \|g\|_\infty h(t_x) \tag{9.70}$$

avec $g(t) = \alpha\sigma_1^\alpha\sqrt{1+\Phi^2}e^{-\sigma_1^\alpha|t|^\alpha} \in L^\infty(\mathbb{R}_+)$ et $h(t) = \frac{1}{|t|^{1-\alpha}}e^{-\sigma_2^\alpha|t|^\alpha} \in L^1(\mathbb{R}_+)$. Comme la borne (terme de droite dans (9.70)) est indépendente de t_1 et intégrable au sens de Lebesgue sur $\mathbb{R}_+\backslash\{0\}$,et comme $\{0\}$ est un ensemble négligeable, on peut conclure que :

$$\forall t_1\in\mathbb{R}_+,\ \frac{\partial\int_{\mathbb{R}_+}\varphi_{s_1,x}(t_1,t_x)e^{-it_xx}dt_x}{\partial t_1} = \int_{\mathbb{R}_+}\frac{\partial\varphi_{s_1,x}}{\partial t_1}(t_1,t_x)e^{-it_xx}dt_x, \tag{9.71}$$

ce qui prouve l'équation (9.68). La preuve de (9.69) étant similaire, on montre finalement (9.59).

Étape 3 : Intégration du numérateur dans (9.59). Calculons la dérivée du terme $\tilde{\varphi}(t_x) = \varphi_{s_1,x}(0,t_x)$ par rapport à t_x. On peut appliquer la même technique que précédemment (découper $\mathbb{R}$ en deux parties pour éliminer la valeur absolue) dans l'équation (9.67), et on a alors :

$$\frac{d\tilde{\varphi}}{dt_x}(t_x) = \alpha(\sigma_1^\alpha + \sigma_2^\alpha)[-t_x|t_x|^{\alpha-2} + i\Phi|t_x|^{\alpha-1}]\varphi_{s_1,x}(0,t_x). \tag{9.72}$$

Ainsi, en combinant les équations (9.65) et (9.72), on obtient la relation :

$$\frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x) = \frac{\sigma_1^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}\frac{d\tilde{\varphi}}{dt_x}(t_x). \tag{9.73}$$

L'équation (9.73) est utile car elle nous permet de calculer l'intégrale au numérateur dans l'équation (9.59). En effet :

$$\int_{\mathbb{R}} \frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x)e^{-it_x x}dt_x = \frac{\sigma_1^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}\int_{\mathbb{R}} \frac{d\tilde{\varphi}}{dt_x}(t_x)e^{-it_x x}dt_x, \tag{9.74}$$

et une intégration par parties permet d'écrire [1] :

$$\int_{\mathbb{R}} \frac{d\tilde{\varphi}}{dt_x}(t_x)e^{-it_x x}dt_x = -\int_{\mathbb{R}} \tilde{\varphi}(t_x)\frac{d(e^{-it_x x})}{dt_x}dt_x = ix\int_{\mathbb{R}} \tilde{\varphi}(t_x)e^{-it_x x}dt_x. \tag{9.75}$$

En combinant les équations (9.74) et (9.75), on obtient :

$$\int_{\mathbb{R}} \frac{\partial\varphi_{s_1,x}}{\partial t_1}(0,t_x)e^{-it_x x}dt_x = i\frac{\sigma_1^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}x\int_{\mathbb{R}} \varphi_{s_1,x}(0,t_x)e^{-it_x x}dt_x, \tag{9.76}$$

et on ré-injecte cette expression dans (9.59), ce qui conduit à :

$$\frac{d\varphi_{s_1|x}}{dt_1}(0) = i\frac{\sigma_1^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}x. \tag{9.77}$$

Étape 4 : Obtention de l'estimateur. Finalement, en utilisant (9.77) avec (9.53) et (9.54), on obtient l'estimateur de la source s_1 :

$$\hat{s}_1 = \frac{\sigma_1^\alpha}{\sigma_1^\alpha + \sigma_2^\alpha}x. \tag{9.78}$$

De façon complètement analogue, on a naturellement l'estimateur de la source s_2. Ainsi, pour $K = 2$, on a le résultat suivant :

$$\forall k \in [\![1,K]\!],\ \hat{s}_k = \frac{\sigma_k^\alpha}{\sum_{l=1}^{K}\sigma_l^\alpha}x. \tag{9.79}$$

Remarque : En fait, ce résultat reste valable pour n'importe-quelle valeur de β (nous avons ici supposé qu'il valait 1 pour se placer dans le cas des distributions PαS). En effet, si $\alpha \in]0,1[$, la fonction caractéristique (9.55) devient :

$$\forall t_x \in \mathbb{R},\ \varphi_x(t_x) = \mathbb{E}_x(e^{it_x x}) = e^{-\sigma^\alpha|t_x|^\alpha + i\beta\Phi\sigma^\alpha|t_x|^\alpha sg(t_x)}, \tag{9.80}$$

Ainsi, si on remplace Φ par $\beta\Phi$ dans notre démonstration, le résultat reste valable. La nouvelle constante $\beta\Phi$ peut cependant s'annuler (pour $\beta = 0$), ce qui perturbe la démonstration, mais la distribution est alors symétrique α-stable (SαS) : le résultat (9.79) est toujours valable, comme démontré dans Liutkus et Badeau (2015).

1. En toute rigueur, cette intégration par partie fait intervenir le produit des termes non dérivés évalué en $+\infty$ et $-\infty$. Or ce produit de termes a pour module $e^{-(\sigma_1^\alpha+\sigma_2^\alpha)|t_x|^\alpha}$, il est donc nul en $\pm\infty$.

9.3.2 Somme de K variables

Nous montrons à présent que (9.79) est valable quelque soit le nombre de sources K. Considérons une somme de $K \geq 2$ variables s_k qui suivent une loi PαS de paramètre σ_k. On définit alors, $\forall k \in [\![1, K]\!]$, $\tilde{s}_k = \sum_{l \neq k} s_l$. Ainsi :

$$s_k + \tilde{s}_k = x, \tag{9.81}$$

et par propriété de stabilité des lois PαS, $\tilde{s}_k$ est aussi une variable PαS de paramètre $\tilde{\sigma}_k$ telle que $\tilde{\sigma}_k^\alpha = \sum_{l \neq k} \sigma_l^\alpha$. On utilise alors (9.79) avec les deux sources s_k and $\tilde{s}_k$:

$$\hat{s}_k = \frac{\sigma_k^\alpha}{\sigma_k^\alpha + \tilde{\sigma}_k^\alpha} x = \frac{\sigma_k^\alpha}{\sigma_k^\alpha + \sum_{l \neq k} \sigma_l^\alpha} x = \frac{\sigma_k^\alpha}{\sum_l \sigma_l^\alpha} x. \tag{9.82}$$

Comme ce résultat est valable quel que soit $k \in [\![1, K]\!]$, l'équation (9.79) est donc valable pour tout K. Enfin, on peut écrire ce résultat matriciellement en considérant que tous les entrées de X sont indépendantes :

$$\hat{X}_k = \frac{\sigma_k^{\odot\alpha}}{\sum_l \sigma_l^{\odot\alpha}} \odot X. \tag{9.83}$$

En conclusion, nous avons démontré que le *filtrage de Wiener généralisé* (9.83) déjà justifié par Liutkus et Badeau (2015) dans le cas des distributions SαS, restait valable pour les distributions PαS. Nous avons à présent un outil pour estimer des sources à partir de mélange de données non-négatives.

Dans le cas du modèle de Lévy NMF, on a donc, $\forall k \in [\![1, K]\!]$,

$$\hat{X}_k = \frac{\sigma_k^{\odot 1/2}}{\sum_l \sigma_l^{\odot 1/2}} \odot X = \frac{W_k H_k}{\sum_l W_l H_l} \odot X, \tag{9.84}$$

où W_k (resp. H_k) désigne la k-ième colonne (resp. ligne) de la matrice W (resp. H).

Remarque : Cet estimateur des sources est généralement optimal au sens des moindres carrés (estimateur MMSE). Néanmoins, pour montrer que c'est le cas, il faut prouver que la variance de $s_k|x$ est bien définie (c'est le cas pour des distributions gaussiennes).

9.4 Expériences

Nous menons une série d'expériences pour montrer l'intérêt des distributions PαS et de l'algorithme de Lévy NMF pour diverses applications. Les données utilisées sont des mélanges de notes de piano issues de la base MAPS Emiya et al. (2010), des morceaux de guitare issus de la base IDMT-SMT-GUITAR Kehling et al. (2014), et des extraits de morceaux de musique polyphoniques issus de la base DSD100 Ono et al. (2015). Les signaux sont échantillonnés à 44100 Hz. La TFCT est calculée avec une fenêtre de Hann de longueur 93 ms et 75 % de recouvrement. Plusieurs méthodes sont comparées :

— ISNMF : NMF avec divergence d'Itakura-Saito, correspondant au modèle gaussien Févotte et al. (2009) ;

— KLNMF : NMF avec divergence de Kullback-Leibler, correspondant au modèle de Poisson Virtanen et al. (2008) ;

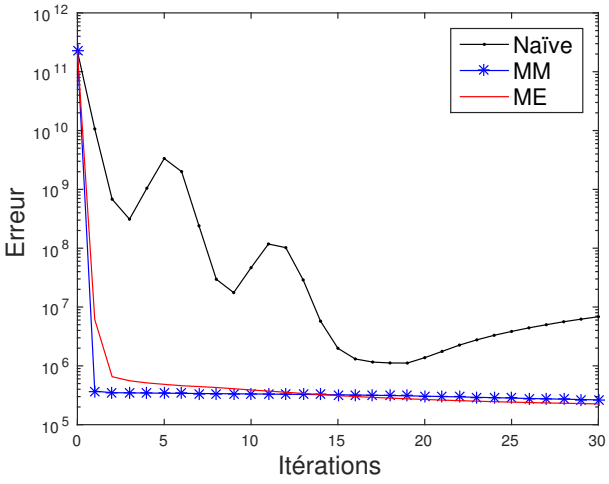


FIGURE 9.6 – Évolution de la fonction de coût au cours des itérations pour les différentes approches retenues pour l'algorithme de Lévy NMF.

— Cauchy NMF : modèle de Cauchy LIUTKUS et al. (2015) ;
— Lévy NMF : modèle de Lévy, présenté dans ce chapitre ;
— RPCA : Analyse en Composantes Principales Robuste CANDÈS et al. (2011). Nous avons utilisé l'implémentation HUANG et al. (2012). Il est en effet intéressant de comparer les approches NMF sus-citées à la PCA robuste, car celle-ci est précisément utilisée pour sa robustesse aux bruits, ce qui, on le verra, est l'un des atouts de la Lévy NMF.

Les algorithmes utilisent 200 itérations (sauf explicitement mentionné), valeur pour laquelle la convergence est observée et au-delà de laquelle les résultats ne sont pas améliorés.

9.4.1 Algorithmes de Lévy NMF

Ces premières expériences étudient certaines propriétés des algorithmes de Lévy NMF présentés dans la section 9.2.

Convergence des algorithmes

Nous testons les différents algorithmes sur un extrait de la base DSD100 ONO et al. (2015), qui est un morceau de musique polyphonique. Nous effectuons une Lévy NMF par les différentes approches avec un ordre de factorisation $K = 40$ et 30 itérations. Nous traçons sur la figure 9.6 la fonction de coût (qui est, rappelons-le, la divergence IS entre le modèle $[WH]^{\odot 2}$ et les données X, ici le spectrogramme d'amplitude du signal) pour les différentes approches : naïve, MM et ME.

Nous constatons tout d'abord que l'utilisation de l'approche naïve produit une fonction de coût non monotone. En outre, la convergence n'est pas observée (même en augmentant significativement le nombre d'itérations). Les approches MM et ME, quant à elles, conduisent à une fonction de coût décroissante. Conformément à notre résultat théorique, l'approche ME fait décroître la fonction de coût plus lentement que l'approche MM durant les premières itérations, mais à proximité du minimum local, c'est-à-dire pour les dernières itérations, cette approche permet une décroissance de la fonction de coût légèrement plus importante.

Nous utiliserons néanmoins l'approche MM dans la suite des expériences, préférant en effet avoir une garantie de décroissance de la fonction de coût. En outre, les résultats obtenus par les deux approches ne diffèrent pas significativement.

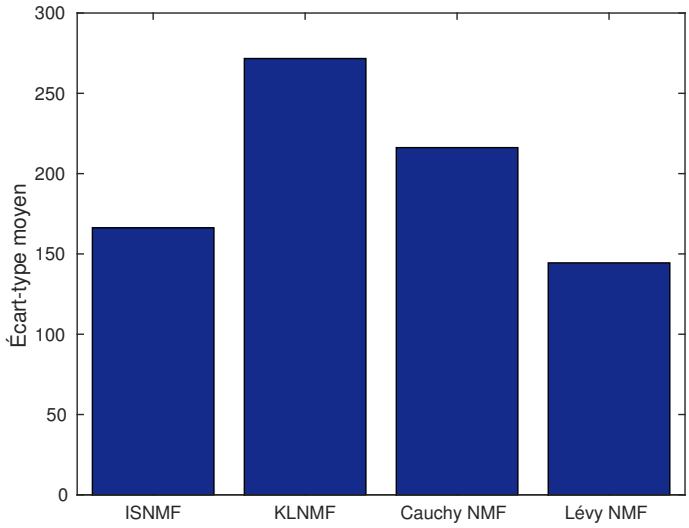


FIGURE 9.7 – Robustesse des algorithmes de NMF à l'initialisation (écart-type moyen de la distribution d'erreur pour plusieurs initialisation aléatoires).

Robustesse à l'initialisation

Nous proposons de tester la robustesse à l'initialisation, c'est-à-dire la capacité d'un algorithme à converger vers la même solution, avec des valeurs initiales différentes. Pour ce faire, nous considérons des signaux de mélanges de notes de piano issus de la base MAPS. Afin de mesurer la robustesse à l'initialisation, nous reconstruisons, par application du filtrage de Wiener généralisé (9.83), les spectrogrammes d'amplitude des différentes sources $\hat{X}_k$, ce qui permet de calculer l'erreur moyenne de reconstruction :

$$\frac{1}{K}\sum_k ||X_k - \hat{X}_k||_2. \tag{9.85}$$

Cette erreur est calculée pour 30 initialisations aléatoires et pour chaque méthode. On mesure la robustesse à l'initialisation par l'écart-type de la distribution de l'erreur (9.85) : plus celui-ci est faible, plus l'erreur est concentrée autour de sa moyenne, plus l'algorithme est alors dit robuste. Enfin, cet écart-type est moyenné sur 30 signaux différents, l'écart-type moyen étant représenté sur la figure 9.7. On constate que l'algorithme de Lévy NMF est le plus robuste parmi les 4 considérés. C'est un avantage pour des applications pratiques puisque une variation de l'initialisation conduit à faire moins varier le résultat de la factorisation qu'avec les autres approches.

9.4.2 Représentation de bruits impulsionnels

Nous testons à présent la capacité du modèle de Lévy NMF à représenter des bruits de nature impulsive. Pour ce faire, nous considérons des données synthétiques qui sont créées en générant $K = 5$ paires de composantes W et H, obtenues en prenant la puissance quatrième d'un bruit blanc gaussien (ce qui permet d'obtenir des composantes parcimonieuses). Le produit obtenu WH est alors de dimensions $F \times T = 50 \times 50$. On utilise celui-ci comme paramètre d'échelle de distributions PαS ($\sigma^{\odot\alpha} = WH$) pour diverses valeurs de α, notamment de faibles valeurs (comprises entre 0.01 et 0.5) : cela permet de représenter des signaux très impulsionnels. Pour générer ces signaux tests, on utilise la boîte à outils WERON (2010).

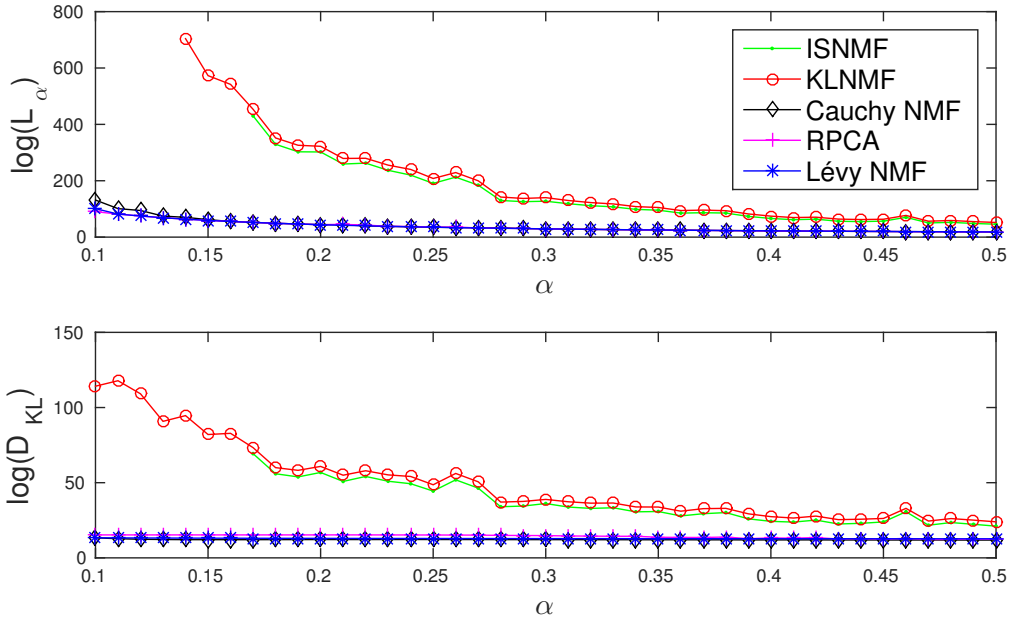


FIGURE 9.8 – Mesures de la qualité d'estimation des paramètres d'échelle.

On applique alors les différentes méthodes d'estimation des paramètres d'échelle $\hat{\sigma}$ sur ces données. Le rang de la factorisation est choisi à 5. Afin de quantifier la qualité d'estimation des paramètres d'échelle, on utilise deux mesures, similairement à ce qui est fait dans LIUTKUS et al. (2015) :

— La α-dispersion $L_\alpha = \sum_{f,t} |\sigma(f,t) - \hat{\sigma}(f,t)|^{1/\alpha}$;

— La divergence de Kullback-Leibler $D_{KL} = \sum_{f,t} \sigma(f,t) \log \frac{\sigma(f,t)}{\hat{\sigma}(f,t)} - \hat{\sigma}(f,t) + \sigma(f,t)$.

On calcule ces deux indicateurs (moyennés sur 100 signaux synthétiques) pour chaque méthode et diverses valeurs de α. Les résultats sont présentés sur la figure 9.8.

Le modèle de Lévy NMF fournit des résultats comparables à la RPCA et à la Cauchy NMF. En particulier, pour de petites valeurs de α (bruits très impulsionnels), la Lévy NMF fournit des résultats légèrement meilleurs que les autres. Cela montre le potentiel de ce modèle pour représenter des signaux de nature variable, qui peuvent aller jusqu'à des bruits très impulsionnels.

9.4.3 Applications

Nous proposons enfin de tester la Lévy NMF dans le cadre de plusieurs applications sur des données réalistes.

Restauration de signaux de musique corrompus synthétiquement

Nous reprenons l'idée de l'expérience précédente mais cette fois-ci en considérant des morceaux de musique réels, qui sont corrompus synthétiquement avec des bruits impulsionnels dans le domaine TF. Bien que ces bruits ne soient pas très réalistes, cela illustre le potentiel

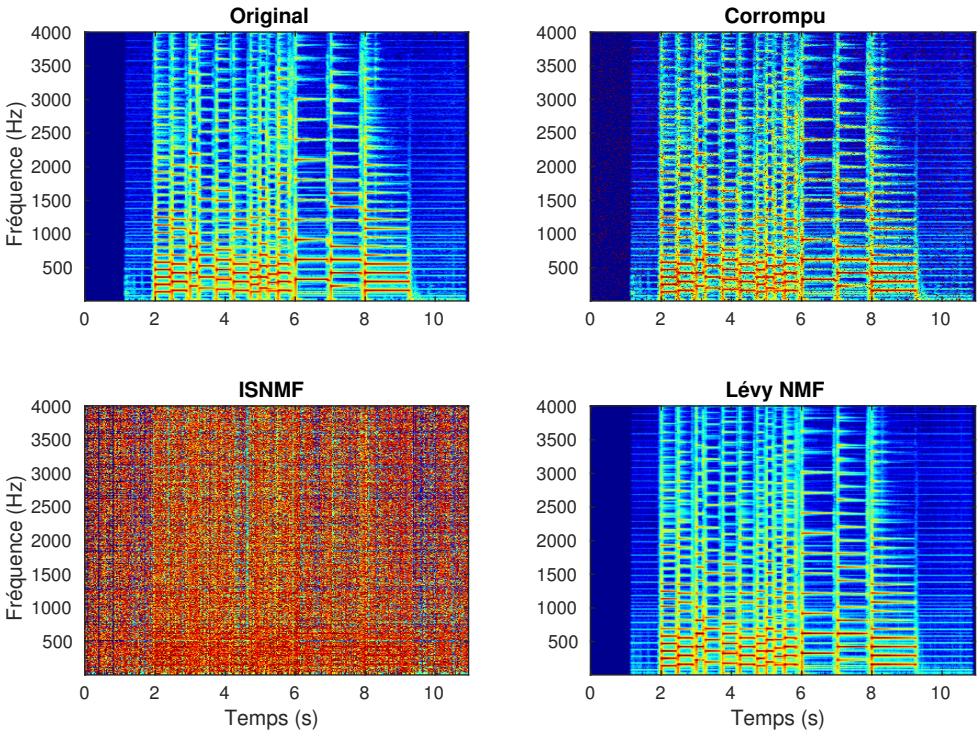


FIGURE 9.9 – Reconstruction de spectrogrammes corrompus grâce à diverses méthodes.

de notre technique en matière de robustesse par rapport à des bruits impulsionnels pour une tâche de restauration de signaux audio.

Les signaux considérés sont issus de la base IDMT-SMT-GUITAR KEHLING et al. (2014) et sont de courts extraits de morceaux de guitare. Nous traçons sur la figure 9.9 divers spectrogrammes que l'on obtient après débruitage par simple application des méthodes de NMF et de RPCA (KLNMF et ISNMF conduisent à des résultats similaires, de même que RPCA, Cauchy NMF et Lévy NMF). On constate que la Lévy NMF semble très robuste aux bruits et que l'estimation des paramètres d'échelle n'est pas perturbée par ceux-ci. Aucun modèle de bruit n'est injecté dans le modèle, ce qui traduit la robustesse naturelle du modèle de Lévy NMF aux bruits impulsionnels.

Ce constat est quantifiable par la divergence KL mesurée entre les spectrogrammes originaux et restaurés. On présente celle-ci sur la figure 9.10. On constate que la Lévy NMF conduit aux meilleurs résultats (c'est-à-dire à la plus basse divergence KL), ce qui confirme l'intuition préalablement faite à partir des spectrogrammes.

On reconstruit ensuite les signaux temporels en appliquant la phase du signal original non bruité (ce qui est un cas Oracle, mais n'a aucune justification théorique) aux spectrogrammes estimés. On mesure la qualité de reconstruction par le SDR VINCENT et al. (2006) et on présente les résultats sous forme de boîtes à moustaches sur la figure 9.10. On constate que la Lévy NMF donne les meilleurs résultats, confirmant le potentiel de celle-ci comme un outil d'estimation de paramètres très robuste aux bruits impulsionnels dans le domaine TF, notamment dans le cas de signaux audio.

Nous avons enfin mené une expérience complémentaire, qui consiste à informer le modèle ISNMF par la localisation des bruits : on applique alors une ISNMF pondérée LIMEM et al. (2013) par un masque qui vaut 0 pour les points TF corrompus, et 1 ailleurs. Les résultats s'en

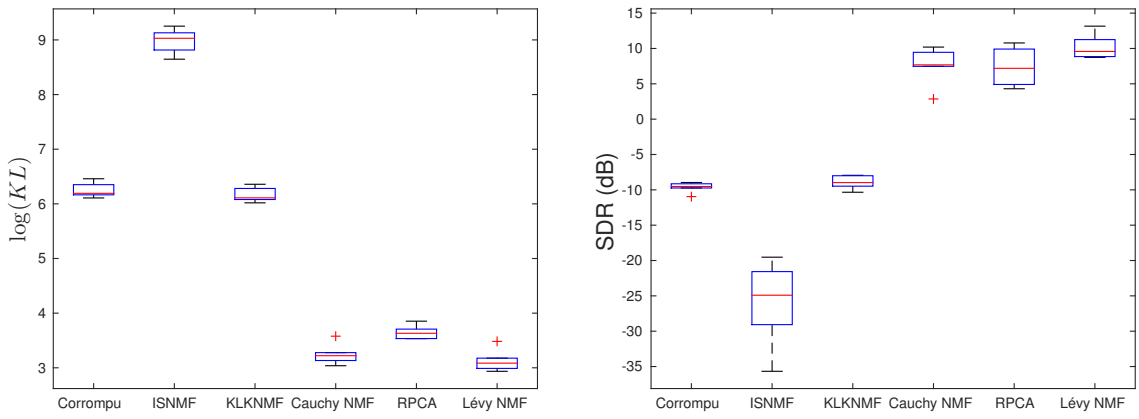


FIGURE 9.10 – Qualité de restauration des spectrogrammes mesurée par la divergence KL (gauche) et de reconstruction de signal mesurée par le SDR (droite).

voient améliorés (on passe en moyenne de 9 u. environ [2] pour la ISNMF non informée à 3.8 u. pour la ISNMF pondérée). Néanmoins, les résultats restent moins bons que ceux obtenus avec la Lévy NMF (environ 3.2 u.) sachant que celle-ci n'est pas informée. Cela traduit la supériorité de ce modèle, qui mène de façon aveugle à des performances meilleures que certains modèles informés (ISNMF pondérée) ou construits sur des propriétés de robustesse (RPCA).

Rehaussement de l'accompagnement musical

Nous considérons 30 signaux tirés de la base DSD100, composés d'une piste d'accompagnement et d'une piste comportant la voix chantée. En appliquant une Lévy NMF (sans aucun à priori ou modèle sur la voix), on s'attend à ce que la voix, traitée comme du bruit, soit partiellement supprimée au profit de l'accompagnement. On choisit un rang de factorisation égal à 30 et 200 itérations de NMF. L'accompagnement est alors estimé par filtrage de Wiener généralisé pour les différentes méthodes. Dans le cas de la Lévy NMF (et de la RPCA), on applique la phase du mélange aux amplitudes estimées (ce qui pourrait être amélioré par une technique plus avancée de reconstruction de phase). On calcule enfin le SDR entre l'accompagnement musical original et sa version estimée. Les résultats sont présentés sur la figure 9.11.

On constate que les méthodes Cauchy NMF, RPCA et Lévy NMF donnent des résultats similaires et supérieures aux techniques plus traditionnelles de ISNMF et KLNMF (avec un très léger avantage pour la Lévy NMF). Perceptivement, ce sont les seules méthodes sur lesquelles on entend effectivement un rehaussement de l'accompagnement musical et une quasi-suppression de la partie voix. Cette expérience montre donc le potentiel de Lévy NMF pour de telles applications (ainsi que celui de modèles comme Cauchy NMF).

Application à la spectroscopie de fluorescence

L'intérêt du modèle de distribution PαS est de traiter directement des données non-négatives. Les expériences précédentes ont montré le potentiel de la Lévy NMF, mais dans le cadre de données audio, qui sont intrinsèquement complexes. L'avantage par rapport à la méthode Cauchy NMF n'apparait donc pas clairement. Cette dernière possède en effet des pro-

2. On désigne par u. l'unité du logarithme de la divergence de Kullback-Leibler, par commodité d'écriture.

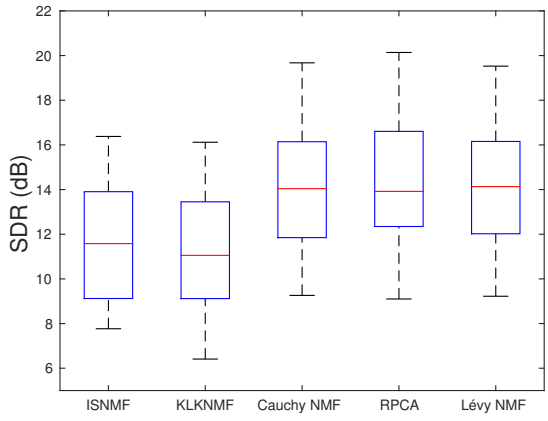


FIGURE 9.11 – Qualité du rehaussement de l'accompagnement (SDR en dB).

priétés similaires (stabilité, robustesse) et pourrait paraître plus adapté pour des applications audio, puisque la loi de Cauchy modélise directement des données complexes.

Le principal intérêt de la Lévy NMF est la modélisation et l'estimation de données non négatives, aussi nous proposons une application différente de l'audio, où son potentiel peut être encore davantage révélé. Dans un tel contexte, les modèles de type Cauchy NMF ou même ISNMF (gaussiennes complexes) ne sont pas appropriés. Nous avons choisi de considérer l'application de séparation de sources pour des données issues de la chimie, et plus spécifiquement de la spectroscopie de fluorescence LIU et al. (2013).

Pour cette application, les données considérées sont des spectres d'émission de certaines espèces chimiques : une fois une espèce excitée par une lumière incidente à une certaine longueur d'onde (365 nm pour nos données), elle émet avec une intensité variable selon la longueur d'onde d'émission considérée. En séparant le spectre d'émission d'un mélange d'espèces données, on peut alors estimer les espèces chimiques pures qui composent le mélange (spectres W), et dans quelles proportions (matrice de concentration H).

Le jeu de données dont nous disposons est celui employé dans GOBINET et al. (2004) [3]. Ce jeu de données est composé de spectres d'émission d'un grain de blé, dans la bande de longueurs d'onde allant de 350 à 670 nm (avec une précision d'environ 2.5 nm, soit $F = 128$ canaux fréquentiels). Ces spectres sont obtenus en différents points d'une section de grain (sur une grille de 20×20 points), ce qui conduit à $T = 400$ spectres en tout.

Les spectres d'émission de grains de blé sont composés principalement de 3 composantes, qui sont les spectres de l'acide férulique, de l'acide férulique libre et de l'acide paracoumarique. Nous avons à notre disposition les spectres de ces composés purs. Ne connaissant pas les concentrations exactes des espèces dans nos mélanges, on commence par appliquer une NMF dans un cas "Oracle", en supposant les spectres des composantes isolées connues et en n'estimant que les concentrations (c'est-à-dire la matrice H). Cela nous donne un point de référence qui servira par la suite à mesurer l'erreur de séparation de sources. La figure 9.12 illustre les cartes de concentrations obtenues pour les différents composants dans le cas Oracle (c'est donc une carte de 20×20 coordonnées spatiales).

Ensuite, nous effectuons sur les mélanges diverses NMF aveugles (en apprenant donc les spectres et les concentrations) utilisées dans la littérature dédiée à la spectroscopie de fluorescence : NMF avec distance Euclidienne (Euc) GOBINET et al. (2004); MONTCUQUET

3. Les mesure ont été effectuées à l'INRA Montpellier et nous ont été fournies par Cyril Gobinet, l'auteur de cet article. Nous en profitons donc pour le remercier.

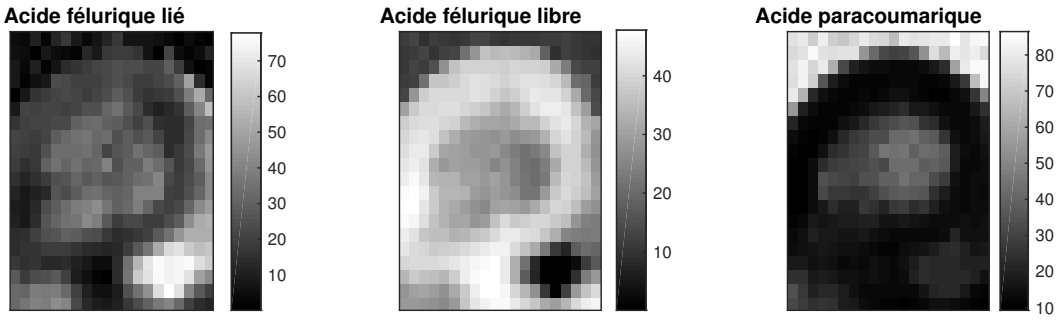


FIGURE 9.12 – Cartographie des concentrations des diverses espèces obtenues dans le cas Oracle.

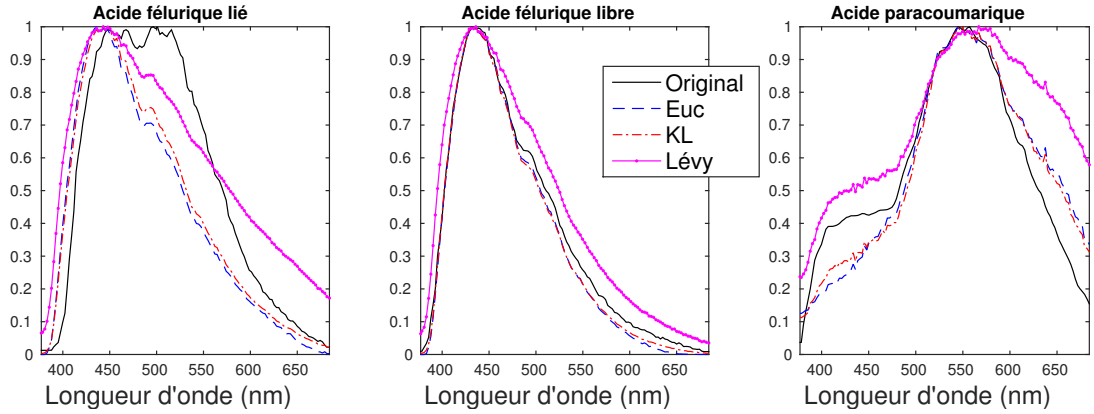


FIGURE 9.13 – Spectres d'émission (normalisés) réels et estimés par différentes méthodes pour les 3 composantes.

et al. (2009) et NMF avec divergence KL GOBINET et al. (2005). Les spectres obtenus sont représentés sur la figure 9.13. On constate le potentiel du modèle de Lévy NMF pour apprendre avec une certaine précision des spectres à partir de mélanges, avec des résultats similaires aux autres méthodes de NMF. On remarquera notamment que la Lévy NMF a tendance à approcher les spectres de référence par valeurs supérieures.

On peut alors estimer les sources par filtrage de Wiener généralisé. Pour chaque source et pour chaque méthode, on évalue la qualité de l'estimation par la corrélation entre les sources Oracle et estimées. Le résultat est présenté sur la figure 9.14. On constate que la reconstruction des sources par Lévy NMF conduit à un résultat plus proche de l'oracle que les autres approches. Cela confirme donc le potentiel de la Lévy NMF pour des applications de séparation de sources non-négatives.

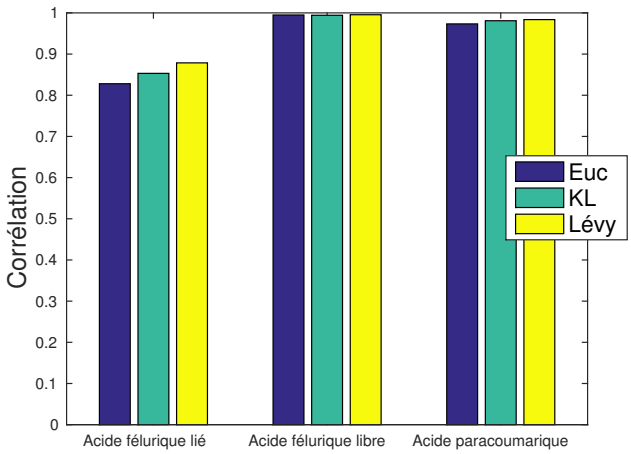


FIGURE 9.14 – Qualité de la reconstruction des sources pour différentes méthodes NMF.

9.5 Conclusion

Nous avons introduit dans ce chapitre les distributions positives α-stables pour représenter des données non-négatives. En prenant le cas particulier de la distribution de Lévy, nous avons mis au point une procédure d'estimation des paramètres des sources dans un mélange, par un modèle NMF estimé de différentes façons. Un estimateur des sources généralisant le filtrage de Wiener a également été obtenu. Expérimentalement, nous avons constaté le potentiel de cet outil en ce qui concerne sa robustesse au bruit impulsionnel dans le domaine TF, et vu qu'il était compétitif avec d'autres méthodes de l'état de l'art pour diverses tâches.

De nombreuses disciplines s'intéressent à la modélisation et à la séparation de données non-négatives par approches NMF, comme la physique appliquée SAJDA et al. (2004) ou la vision par ordinateur LEE et SEUNG (1999). Par ailleurs, la distribution de Lévy est utilisée dans divers domaines tels que l'optique ROGERS (2008) ou l'analyse du champ magnétique terrestre CARBONE et al. (2006).

Outre les nombreuses applications potentielles, la méthode présentée dans ce chapitre peut servir de fondation à des techniques plus avancées. En estimant les paramètres au sens MAP, il est possible d'intégrer certaines informations sur les matrices W (comme l'harmonicité) et H (parcimonie, continuité temporelle), et ainsi obtenir des résultats plus satisfaisants. Dans le cas général des distributions PαS, il serait intéressant de mettre en place une technique d'estimation (basée sur les MCMC par exemple) afin de pleinement exploiter le potentiel de ces distributions pour la modélisation robuste de données non négatives. Alternativement, on pourrait s'intéresser aux distributions inverse-gamma KOUNADES-BASTIAN et al. (2016), dont la loi de Lévy est un cas particulier, et dont la densité est exprimable analytiquement. Néanmoins, cette famille de lois n'est pas additive dans le cas général, et il n'est pas évident d'obtenir la loi d'une somme de variables inverse-gamma WITKOVSKY (2001), ce qui requiert la mise en place de méthodes variationnelles.

Nous avons considéré, dans nos expériences en audio, que les spectrogrammes d'amplitude suivaient une loi de Lévy : il est possible que de meilleurs résultats soient obtenu en modélisant non plus les spectrogrammes d'amplitude, mais des spectrogrammes fractionnaires, comme le suggèrent certains résultats récents sur le sujet LIUTKUS et BADEAU (2015); VORAN (2015). Enfin, ces modèles pourraient être combinés à un modèle de phase non uniforme (*cf.* chapitre 8) pour la séparation de sources audio.

Chapitre 10

Conclusion

Sommaire

Cet ultime chapitre vise à conclure ce travail de thèse. Nous dressons un bilan de nos contributions dans la section 10.1, et nous suggérons quelques directions de recherche qui prolongent et/ou complètent ce travail dans la section 10.2. Enfin, nous résumons les publications qui ont résulté de nos travaux dans la section 10.3.

10.1 Contributions

La première contribution de cette thèse a été de de montrer la nécessité de la reconstruction de phase dans la séparation de sources par approche NMF. Ce sujet a été durant longtemps considéré d'importance moindre que l'estimation des amplitudes. Notre étude comparative du chapitre 3 a montré que la qualité de la séparation de sources était limitée par celle de l'estimation des phases des composantes. Une direction prometteuse a été identifiée : la reconstruction de phase basée sur la modélisation de signaux.

Une partie importante de ce travail de thèse a consisté en l'utilisation de modèles sinusoïdaux qui fournissent des contraintes de phases dans le domaine TF. Nous avons contribué, par une nouvelle approche pour l'estimation des paramètres, à rendre cette technique très générale et applicable à une grande variété de signaux, notamment inharmoniques et non stationnaires. Nous avons intégré cette approche dans le cadre de deux applications réalistes : le débruitage audio ainsi que la séparation de sources, via une approche déterministe et un modèle probabiliste. Les résultats obtenus ont confirmé le potentiel des modèles de signaux pour la mise au point de techniques efficaces de reconstruction de phase.

Pour reconstruire les phases dans les trames d'attaque, nous avons proposé un modèle basé sur la répétition des évènements audio. Il s'agit là d'une contribution originale, puisque c'est, à notre connaissance, la première fois que le caractère répétitif des évènements qui composent les signaux sources est exploité pour la reconstruction de phase, alors que cette propriété est communément utilisée pour estimer les spectrogrammes. Nous avons également intégré cette propriété à un modèle de mélange pour la séparation de sources dans les trames d'attaques.

Nous sommes ensuite revenus à une problématique d'estimation conjointe d'amplitudes et de phases. Nous avons proposé un modèle inspiré de la NMF complexe, qui inclut des contraintes sur les phases des composantes, utilisant donc les modèles précédemment introduits. À la différence des approches similaires existantes, notre méthode ne requiert pas l'injection d'information externe (de type partition) ni la restriction à une classe de signaux particulière (harmoniques et stationnaires). Cette méthode est donc adaptée à la séparation aveugle de sources.

Dans le but de mettre au point un modèle complet de sources complexes à phases non-uniformes, nous avons étudié une famille de distributions à support non-négatif pour la modélisation des spectrogrammes d'amplitudes : les distributions positives α-stables. Celles-ci allient robustesse aux valeurs aberrantes et stabilité par somme. De façon quelque peu orthogonale au fil conducteur de cette thèse (la reconstruction de phase), nous avons ainsi mis au point un modèle de mélange basé sur ces distributions, et proposé un cadre complet pour la séparation de sources non-négatives. Outre les bonnes performances obtenues sur des données audio, ce type de modèles peut trouver des applications dans d'autres domaines de recherche.

10.2 Perspectives

10.2.1 Transformées multi-résolutions pour la reconstruction de phase

Les résultats du chapitre 4 ont montré que la technique de déroulé linéaire appliquée "à l'aveugle" (sans tenir compte d'une information supplémentaire comme la phase du mélange

dans le cadre de la séparation de sources) avait tendance à créer deux types d'artéfacts : le bruit musical et la perte de précision au niveau des transitoires. Ces artéfacts sont étroitement liés aux paramètres de la TFCT utilisés, et la recherche d'un compromis entre ces deux types de perturbations fait écho au compromis fondamental de l'analyse temps-fréquence entre résolution temporelle et fréquentielle. Pour pallier ce problème inhérent à la TFCT, des transformations alternatives ont été proposées. Elles reposent sur le principe de multi-résolution : c'est le cas par exemple de la transformée à Q-constant Fillon et Prado (2012), adaptée à l'audio, de la transformée en ondelettes Mallat (1998), ou encore de la TFCT à fenêtre variable Kwok et Jones (2000).

Il pourrait donc être intéressant de calculer la "phase" de mélanges de sinusoïdes dans de telles représentations. Des algorithmes similaires au notre pourraient être obtenus, mais appliqués à des représentations TF qui permettent de s'affranchir de cette recherche de compromis entre résolutions. Des versions sophistiquées du vocodeur de phase existent, et sont basées sur le fait de traiter spécifiquement les transitoires d'attaque afin de réduire le phénomène de *phasiness* Röbel (2003b). Ainsi, un cadre multi-résolution semble approprié pour lever ce type de verrous, et pourra notamment être utilisé pour la reconstruction de phase.

10.2.2 Phases dans les trames d'attaque

Modèles d'attaque Les résultats obtenus dans le chapitre 4 sur la reconstruction des phases de transitoires d'attaque étaient prometteurs, mais il apparaît clairement qu'il existe une nette marge de progression pour cette tâche. En effet, le modèle que nous avons proposé (impulsion) reste peu expressif de la diversité rencontrée dans les signaux réels. Il pourrait être intéressant d'étudier un modèle de mélange d'impulsions, dont il faudrait alors détecter les "temps d'attaque" multiples. Plus généralement, modéliser les dépendances entre canaux fréquentiels dans les trames d'attaque par filtrage ARMA Leglaive et al. (2016a) est une idée intéressante qui généralise le modèle d'impulsion.

Une approche différente pourrait également se trouver dans l'acoustique musicale, où les relations de phase entre partiels ont été étudiées Galembo et al. (2001), et appliquées avec de bons résultats préliminaires à la séparation de sources par NMF complexe contrainte Kirchhoff et al. (2014). Notons enfin qu'il serait intéressant d'étudier la pertinence de la phase des attaques en tant que descripteur du timbre : on sait en effet que les attaques sont un élément fondamental de la perception du timbre d'un instrument, aussi il pourrait être judicieux d'en étudier spécifiquement la phase, afin d'évaluer si celle-ci joue ou pas un rôle prépondérant dans la reconnaissance des instruments.

Enfin, les modèles que nous avons proposés dans cette thèse n'étaient pas dédiés aux sons percussifs. Nous avons constaté, dans les expériences du chapitre 4, que combiner un modèle d'impulsion avec celui de mélange de sinusoïdes ne donnait pas de résultats satisfaisants. Une direction intéressante pour traiter ces sons est d'utiliser des atomes spectraux avec un modèle de NMF convolutive, comme c'est proposé dans Laroche et al. (2017). Cette approche permet de rendre compte du caractère fortement non-stationnaire de ces signaux. En étendant ces modèles à des atomes complexes, donc en prenant en compte la phase de ces signaux, on peut espérer améliorer la qualité des résultats obtenus.

Répétition des phases d'attaque Outre la modélisation des transitoires d'attaque dans le domaine TF, nous avons également proposé d'utiliser leur caractère répétitif pour contraindre la phase (cf. chapitre 6). Ce modèle a donné quelques bons résultats dans le cas Oracle où les amplitudes sont égales à la vérité terrain, mais n'a pas vraiment amélioré les performances par rapport à une approche non contrainte dans un cas réaliste, celui du chapitre 7, où les amplitudes ne sont plus connues. Ce modèle pourrait être affiné en tenant compte du fait

que l'attaque ne concerne en réalité pas qu'une seule trame, mais plusieurs (en raison du caractère redondant de la TFCT). Ainsi, nous pourrions mettre au point un modèle inspiré de la notion de consistance "locale" (au niveau des attaques), qui serait contraint par un modèle de répétition.

10.2.3 Données non-négatives

Les distributions PαS introduites dans le chapitre 9 sont un outil prometteur pour la modélisation et la séparation de données non-négatives. Néanmoins, nous nous sommes limités, pour les applications pratiques, à la distribution de Lévy, qui est la seule à posséder une densité exprimable analytiquement. Cependant, nous pourrions envisager de mettre en place un modèle plus général de somme de sources PαS, dont les paramètres seraient par la suite estimés grâce à des techniques plus avancées, comme les méthodes MCMC Simsekli et al. (2015).

Dans une direction quelque peu similaire, on pourrait s'intéresser aux distributions inverse-gamma, dont la loi de Lévy est également un cas particulier. Cette famille de distributions, qui modélise des données non-négatives, a une densité qui s'écrit analytiquement de façon simple dans le cas général, néanmoins elle n'est pas additive. Aussi, des méthodes d'inférence variationnelle pourraient être mises en oeuvre pour estimer ces modèles Kounades-Bastian et al. (2016).

Plus spécifiquement en audio, il est pertinent de s'interroger non seulement sur la distribution la plus appropriée pour modéliser les données, mais également sur la nature même de ces données : la problématique du choix de l'exposant optimal de spectrogramme Voran (2015) résume bien cette question. Des expériences plus poussées pourraient être conduites pour obtenir une combinaison optimale (en ce qui concerne la fidélité aux données) d'une puissance de spectrogramme et d'un paramètre de forme de distribution (comme l'exposant α pour les lois stables). Cette problématique rejoint d'ailleurs une perspective précédente sur le travail sur une représentation multi-résolution, et plus généralement sur une représentation TF alternative à la TFCT. On sait notamment que la MDCT donne de bons résultats en audio, et que modéliser des coefficients MDCT de signaux audio par des lois de Student est un choix relativement précis Févotte et Godsill (2005). Aussi, il est prometteur de s'intéresser à des représentations de signaux audio alternatives à la TFCT, dans lesquelles nous avons une meilleure maîtrise du comportement de ces données.

10.2.4 Modèle de sources complexes à phase non-uniforme

Afin de mettre au point un modèle complet de sources décrivant aussi bien les phases que les amplitudes, on pourra reprendre l'idée du chapitre 8, qui consiste à supposer que la phase suit une loi de Von Mises (donc non-uniforme). On pourra compléter ce modèle en supposant alors que les amplitudes ne sont plus déterministes, mais sont à présent des variables aléatoires dont il deviendra nécessaire d'estimer les paramètres.

Modèle gaussien anisotrope Une première approche consiste à considérer que les amplitudes suivent une loi de Rayleigh. Ce choix est naturel car le modèle gaussien isotrope classique Févotte et al. (2009) revient à considérer une phase uniforme et une amplitude de Rayleigh. Nous proposons de conserver la modélisation de l'amplitude par une loi de Rayleigh, mais en modélisant à présent la phase par une loi de Von Mises. Cela conduit à un modèle dans lequel on ne sait pas exprimer analytiquement les densités des sources, aussi il est approché par un modèle gaussien anisotrope. On illustre sur la figure 10.1 la comparaison entre modèle originel (Rayleigh+Von Mises) et approché (gaussien anisotrope).

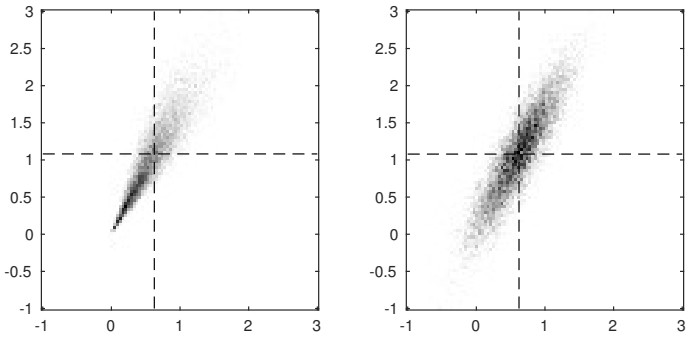


FIGURE 10.1 – Histogrammes en 2D d'échantillons générés par le modèle Von Mises + Rayleigh (gauche) et modèle équivalent gaussien (droite), pour $\sigma = 1$, $\mu = \pi/3$ et $\kappa = 100$. Les intersections des lignes en pointillés correspondent aux valeurs moyennes.

Dans ce nouveau modèle, les distributions des sources X_k dépendent de paramètres $\theta_k = \{\mu_k, W_k, H_k\}$ où μ_k est le paramètre de localisation de la variable de phase, et W_k, H_k correspondent aux paramètres du modèle NMF qui est utilisé pour structurer les paramètres de dispersion des lois de Rayleigh. Afin d'estimer les paramètres de ce modèle, on peut adopter une approche SAGE (*Space-Alternating Generalized EM*), similairement à BERTIN et al. (2010), qui fournit les mises à jour suivantes :

$$\theta_k \leftarrow \arg\max_{\theta_k} Q_k(\theta_k, \theta), \tag{10.1}$$

où le critère Q_k est défini par $Q_k(\theta_k, \theta) = \mathbb{E}_{X_k|X,\theta}(\log p(X_k|\theta_k))$. L'étape E consiste à calculer ce critère (ce qui revient à déterminer les moments à postériori des variables latentes X_k, ce qui est relativement aisé à faire dans le cas d'un modèle gaussien), et l'étape M consiste à maximiser celle-ci. Afin d'introduire un à priori sur la forme du paramètre de localisation μ_k, on pourra plutôt considérer le critère suivant :

$$Q_k^{MAP}(\theta_k, \theta) = \mathbb{E}_{X_k|X,\theta}(\log p(X_k|\theta_k)) + \log p(\mu_k), \tag{10.2}$$

l'à priori sur ce paramètre étant donné par une structure en chaînes de Markov, afin de garantir que la phase reste "proche" du modèle de déroulé linéaire. Il est à noter que ce modèle dépend de certains paramètres λ et ρ similaires à ceux du chapitre 8 qui reflètent la forme des distributions choisies (Von Mises et Rayleigh). Aussi, on pourrait aboutir à un modèle tout à fait similaire en modélisant les amplitudes non plus comme des variables de Rayleigh, mais comme des variables de Poisson ou Inverse-Gamma, et les phases comme des variables gaussiennes circulaires ou Cauchy circulaires : seules les expressions de λ et ρ changeraient.

Nos premiers calculs nous ont permis d'aboutir aux règles de mises à jour sur les paramètres θ_k. Par la suite, nous implémenterons celles-ci et effectuerons un certain nombre d'expériences pour attester du potentiel de ce modèle.

Modèle robuste de sources La mise au point d'un modèle de sources complexes dans le domaine de la TFCT peut être effectuée, comme on l'a proposé, en modélisant les phases (par une loi de Von Mises) et les amplitudes (par une loi de Rayleigh). Néanmoins, la loi de Rayleigh n'est peut-être pas le meilleur candidat pour représenter des spectrogrammes d'amplitude car ce n'est pas une distribution à queue lourde. Ainsi, la perspective précédente est intéressante car elle conduit à un modèle dans lequel on peut aisément effectuer un certain nombre de

calculs, mais n'est cependant pas la piste la plus prometteuse en matière de robustesse et de fidélité aux données.

On pourrait donc inclure les distributions stables dans ce contexte : les amplitudes suivraient alors une loi PαS. Cela pose néanmoins plusieurs problèmes. En particulier, dans ce modèle, on ne sait pas exprimer analytiquement la densité du mélange. Même les méthodes de type MCMC deviennent compliquées à mettre en oeuvre pour estimer ces modèles, dans lesquels on ne connaît pas la loi des variables latentes. En outre, si on souhaite approcher ce modèle par un modèle gaussien anisotrope, on se heurte au fait que l'on ne sait pas calculer les moments de lois PαS : il est donc impossible d'obtenir de cette façon les paramètres du modèle gaussien équivalent. Enfin, l'intérêt des distributions PαS est initialement de modéliser les données d'amplitudes de façon robuste : en approchant notre modèle par un modèle gaussien, on perd cette propriété de robustesse.

Aussi, on pourrait s'intéresser à une autre famille de distributions : les lois stables multivariées, ou lois elliptiques. Celles-ci sont aux lois normales multivariées ce que les lois stables sont aux gaussiennes classiques. Elles sont notamment paramétrées par une matrice de forme non-nécessairement diagonale : ainsi, en structurant cette matrice de forme, on pourrait obtenir un modèle de sources qui soit à la fois stable par additivité, robuste et à phase non-uniforme. Il s'agit là d'une piste intéressante car elle pose la question de la justification physique d'une structuration de la matrice de forme, et également de l'estimation des paramètres du modèle.

10.3 Publications

Articles de revues

Paul MAGRON, Roland BADEAU et Antoine LIUTKUS : Lévy NMF for robust nonnegative source separation, *IEEE Signal Processing Letters*, 2017 (submitted).

Paul MAGRON, Roland BADEAU et Bertrand DAVID : STFT phase recovery by sinusoidal modeling for audio source separation, *IEEE Transactions on Audio, Speech and Language Processing*, 2017 (submitted).

Articles de conférences

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Phase-dependent anisotropic Gaussian model for audio source separation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Complex NMF under phase constraints based on signal modeling : application to audio source separation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016.

Fabian-Robert STOTER, Antoine LIUTKUS, Roland BADEAU, Bernd EDLER et Paul MAGRON : Common fate models for unison source separation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Phase reconstruction of spectrograms based on a model of repeated audio events, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2015.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration, *European Signal Processing Conference (EUSIPCO)*, Nice, France, August 2015.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Phase recovery in NMF for audio source separation : an insightful benchmark, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.

Rapports techniques

Paul MAGRON, Roland BADEAU et Antoine LIUTKUS : Generalized Wiener filtering for positive alpha-stable random variables, *Télécom ParisTech 2016D004*, June 2016.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : An iterative algorithm for recovering the phase of complex components from their mixture, *HAL-01325625*, June 2016.

Paul MAGRON, Roland BADEAU et Bertrand DAVID : Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration, *Télécom ParisTech 2015D002*, February 2015.

Bibliographie

Abdallah, S. A. et M. D. Plumbley. 2006, «Unsupervised analysis of polyphonic music by sparse coding», *IEEE Transactions on Neural Networks*, vol. 17, n° 1, p. 179–196. (page 3)

Abe, M. et J. O. Smith. 2004a, «Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks», dans *Proc. Audio Engineering Society Convention 117*, Berlin, Germany. (page 61)

Abe, M. et J. O. Smith. 2004b, «Design Criteria for the Quadratically Interpolated FFT Method (I) : Bias due to Interpolation», Tech. Rep. STAN-M-117, Stanford University, Department of Music. (page 61)

Abe, M. et J. O. Smith. 2005, «AM/FM rate estimation for time-varying sinusoidal modeling», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Philadelphia, PA, USA. (page 60)

Achan, K., S. T. Roweis et B. J. Frey. 2003, «Probabilistic inference of speech signals from phaseless spectrograms», dans *Proc. Advances in Neural Information Processing Systems (NIPS)*, p. 1393–1400. (page 22)

Adler, A., V. Emiya, M. G. Jafari, M. Elad, R. Gribonval et M. D. Plumbley. 2012, «Audio inpainting», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, n° 3, p. 922–932. (page 70)

Agiomyrgiannakis, Y. et Y. Stylianou. 2009, «Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, n° 4, p. 775–786. (page 21, 130)

Alsteris, L. D. et K. K. Paliwal. 2006, «Further intelligibility results from human listening tests using the short-time phase spectrum», *Speech Communication*, vol. 48, n° 6, p. 727–736. (page 12)

Alsteris, L. D. et K. K. Paliwal. 2007, «Short-time phase spectrum in speech processing : a review and some experimental results», *Digital Signal Processing*, vol. 17, n° 3, p. 578–616. (page 12)

Andrieu, C., N. De Freitas, A. Doucet et M. I. Jordan. 2003, «An introduction to MCMC for machine learning», *Machine learning*, vol. 50, n° 1-2, p. 5–43. (page 53, 132)

Artin, E. 2015, *The gamma function*, Courier Dover Publications. (page 114)

Badeau, R. 2011, «Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 253–256. (page 26, 32, 37, 45, 70)

Badeau, R. 2012, *Méthodes à haute résolution*, Télécom ParisTech. (course manual). (page 37, 70)

Badeau, R., N. Bertin et E. Vincent. 2010, «Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization», *IEEE Transactions on Neural Networks*, vol. 21, n° 12, p. 1869–1881. (page 27, 152)

Badeau, R., B. David et G. Richard. 2006, «High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials», *IEEE Transactions on Signal Processing*, vol. 54, n° 4, p. 1341–1350. (page 37)

Badeau, R. et A. Dremeau. 2013, «Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, p. 6171–6175. (page 32, 38)

Badeau, R. et A. Liutkus. 2014, «Proof of Wiener-like linear regression of isotropic complex symmetric alpha-stable random variables», Tech. Rep. HAL-01069612, Paris, France. (page 160)

Badeau, R. et A. Ozerov. 2013, «Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco. (page 38)

Badeau, R. et M. D. Plumbley. 2013a, «Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 1–4. (page 38)

Badeau, R. et M. D. Plumbley. 2013b, «Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain», Tech. Rep. 2012D004, Queen Mary University, London, England. (page 38)

Badeau, R. et M. D. Plumbley. 2013c, «Probabilistic time-frequency source-filter decomposition of non-stationary signals», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Marrakech, Marocco, p. 1–5. (page 38)

Badeau, R. et M. D. Plumbley. 2014, «Multichannel High resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain», *IEEE Transactions on Audio Speech and Language Processing*, vol. 22, n° 11, p. 1670–1680. (page 38, 42, 45, 110)

Barker, J., E. Vincent, N. Ma, H. Christensen et P. Green. 2013, «The PASCAL CHiME speech separation and recognition challenge», *Computer Speech and Language*, vol. 27, n° 3, p. 621–633. (page 63, 123)

Barker, T. et T. Virtanen. 2013, «Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation», dans *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, p. 827–831. (page 35)

Bassiou, N., C. Kotropoulos et E. Koliopoulou. 2013, «Symmetric α-stable sparse linear regression for musical audio denoising», dans *Proc. IEEE International Symposium on Image and Signal Processing and Analysis (ISPA)*, Trieste, Italy, p. 382–387. (page 146)

Beal, M. J. et Z. Ghahramani. 2003, «The variational Bayesian EM algorithm for incomplete data : with application to scoring graphical model structures», *Bayesian statistics*, vol. 7, p. 453–464. (page 31)

Beauregard, G. T., M. Harish et L. L. Wyse. 2015, «Single pass spectrogram inversion», dans *Proc. IEEE International Conference on Digital Signal Processing (DSP)*, p. 427–431. (page 16)

Beckmann, P. 1962, «Statistical distribution of the amplitude and phase of a multiply scattered field», *Journal of Research of the National Bureau of Standards*, vol. 66D, n° 3, p. 231–240. (page 133)

Berry, M. W., M. Browne, A. N. Langville, V. P. Pauca et R. J. Plemmons. 2007, «Algorithms and applications for approximate nonnegative matrix factorization», *Computational statistics & data analysis*, vol. 52, n° 1, p. 155–173. (page 27)

Bertin, N., R. Badeau et E. Vincent. 2009a, «Fast bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 29–32. (page 33)

Bertin, N., R. Badeau et E. Vincent. 2010, «Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription», *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, n° 3, p. 538–549. (page 32, 33, 115, 177)

Bertin, N., C. Févotte et R. Badeau. 2009b, «A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, p. 1545–1548. (page 28)

Betser, M., P. Collen, G. Richard et B. David. 2008, «Estimation of frequency for AM/FM models using the phase vocoder framework», *IEEE Transactions on Signal Processing*, vol. 56, n° 2, p. 505–517. (page 64)

Bishop, C. M. 2006, *Pattern recognition and machine learning*, Springer. (page 31, 135)

Blondel, V. D., N.-D. Ho et P. van Dooren. 2007, «Algorithms for weighted non-negative matrix factorization», *Image and Vision Computing*. (page 34)

Bouboulis, P. 2010, «Wirtinger's calculus in general Hilbert spaces», *arXiv preprint arXiv :1005.5170*. (page 202)

Bouvier, D., N. Obin, M. Liuni et A. Roebel. 2016, «A source/filter model with adaptive constraints for nmf-based speech separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, p. 131–135. (page 34)

Bronson, J. et P. Depalle. 2014, «Phase constrained complex NMF : Separating overlapping partials in mixtures of harmonic musical sources», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy. (page 21, 36, 42, 50, 53, 58, 89, 113)

Candès, E. J., X. Li, Y. Ma et J. Wright. 2011, «Robust principal component analysis ?», *Journal of the ACM (JACM)*, vol. 58, n° 3, p. 11. (page 165)

Carbone, V., L. Sorriso-Valvo, A. Vecchio, F. Lepreti, P. Veltri, P. Harabaglia et I. Guerra. 2006, «Clustering of polarity reversals of the geomagnetic field», *Physical review letters*, vol. 96, n° 12, p. 128 501. (page 172)

Casey, M. A. et A. Westner. 2000, «Separation of mixed audio sources by independent subspace analysis», dans *Proc. International Computer Music Conference (ICMC)*, Berlin, Germany, p. 154–161. (page 35)

Cemgil, A. T. 2009, «Bayesian inference for nonnegative matrix factorisation models», *Computational Intelligence and Neuroscience*, vol. 2009, p. 17. (page 146)

Chacon, C. et P. Mowlaee. 2014, «Least squares phase estimation of mixed signals», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Singapore, p. 2705–2709. (page 21)

Chaigne, A. et J. Kergomard. 2008, *Acoustique des instruments de musique*, Belin, 704 p.. 704 pages. (page 73, 75)

Charbit, M. et O. Cappé. 1997, *Restauration d'enregistrements sonores anciens : Du traitement du signal aux applications grand public*, 6, Société de l'Electricité, de l'Electronique et des Technologies de l'Information et de la Communication (SEE), 40–43 p.. Pp. 40-43. (page 69)

Cichocki, A. et S.-i. Amari. 2010, «Families of alpha- beta- and gamma- divergences : Flexible and robust measures of similarities», *Entropy*, vol. 12, n° 6, p. 1532, ISSN 1099-4300. (page 24)

Cichocki, A., R. Zdunek, A. H. Phan et S.-i. Amari. 2009, *Nonnegative matrix and tensor factorizations : applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons. (page 22)

Corless, R. M., G. H. Gonnet, D. E. Hare, D. J. Jeffrey et D. E. Knuth. 1996, «On the lambert W function», *Advances in Computational mathematics*, vol. 5, n° 1, p. 329–359. (page 156)

Daudet, L. 2005, «A review on techniques for the extraction of transients in musical signals», dans *Proc. International Symposium on Computer Music Modeling and Retrieval*, Pisa, Italy, p. 219–232. (page 62)

Dempster, A. P., N. M. Laird et D. B. Rubin. 1977, «Maximum likelihood from incomplete data via the EM algorithm», *Journal of the royal statistical society. Series B (methodological)*, vol. 39, n° 1, p. 1–38. (page 31)

Dessein, A., A. Cont et G. Lemaitre. 2010, «Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence», dans *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, p. 489–494. (page 32)

Durrieu, J.-L. 2011, «A musically motivated mid-Level representation for pitch estimation and musical audio source separation», *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, n° 6, p. 1180–1191. (page 34)

Efron, B. 2012, *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press. (page 117)

ELDAR, Y. C., N. HAMMEN et D. G. MIXON. 2016, «Recent advances in phase retrieval», *IEEE Signal Processing Magazine*, vol. 33, n° 5, p. 158–162, ISSN 1053-5888. (page 3)

EMIYA, V., N. BERTIN, B. DAVID et R. BADEAU. 2010, «MAPS - A piano database for multipitch estimation and automatic transcription of music», Tech. Rep. 2010D017, Télécom ParisTech, Paris, France. (page 43, 62, 84, 96, 121, 164)

EMIYA, V., E. VINCENT, N. HARLANDER et V. HOHMANN. 2011, «Subjective and objective quality assessment of audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n° 7, p. 2046–2057. (page 39, 45)

EPHRAIM, Y. et D. MALAH. 1984, «Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, n° 6, p. 1109–1121. (page 12, 13)

ESQUEF, P., M. KARJALAINEN et V. VÄLIMÄKI. 2002, «Detection of clicks in audio signals using warped linear prediction», dans *Proc. International Conference on Digital Signal Processing*, vol. 2, Santorini, Greece, p. 1085–1088. (page 69)

EWERT, S., B. PARDO, M. MUELLER et M. D. PLUMBLEY. 2014a, «Score-informed source separation for musical audio recordings : An overview», *IEEE Signal Processing Magazine*, vol. 31, n° 3, p. 116–124, ISSN 1053-5888. (page 32)

EWERT, S., M. D. PLUMBLEY et M. SANDLER. 2014b, «Accounting for phase cancellations in non-negative matrix factorization using weighted distances», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, p. 5. (page 34)

FESSLER, J. A. et A. O. HERO. 1994, «Space-alternating generalized expectation-maximization algorithm», *IEEE Transactions on Signal Processing*, vol. 42, n° 10, p. 2664–2677. (page 32)

FÉVOTTE, C. 2011, «Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, p. 1980–1983. (page 152)

FÉVOTTE, C., N. BERTIN et J.-L. DURRIEU. 2009, «Nonnegative matrix factorization with the Itakura-Saito divergence : with application to music analysis», *Neural computation*, vol. 21, n° 3, p. 793–830. (page 5, 25, 26, 27, 32, 37, 42, 45, 84, 130, 137, 141, 148, 160, 164, 176)

FÉVOTTE, C. et A. T. CEMGIL. 2009, «Nonnegative matrix factorizations as probabilistic inference in composite models», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, p. 1913–1917. (page 25)

FÉVOTTE, C. et S. J. GODSILL. 2005, «A bayesian approach for blind separation of sparse sources», *IEEE Transactions on Speech and Audio Processing*, p. 1–15. (page 176)

FÉVOTTE, C., R. GRIBONVAL et E. VINCENT. 2005, «BSS_EVAL toolbox user guide», Tech. Rep. 1706, IRISA, Rennes, France. (page 13, 25)

FÉVOTTE, C. et J. IDIER. 2011, «Algorithms for nonnegative matrix factorization with the beta-divergence», *Neural Computation*, vol. 23, n° 9, p. 2421–2456. (page 27, 29, 152, 155)

Févotte, C. et M. Kowalski. 2014, «Low-rank time-frequency synthesis», dans *Proc. Advances in Neural Information Processing Systems (NIPS)*. (page 22)

Fillon, T. et J. Prado. 2012, «A flexible multi-resolution time-frequency analysis framework for audio signals», dans *Proc. International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, p. 1124–1129. (page 51, 175)

Fitzgerald, D., M. Cranitch et E. Coyle. 2005, «Shifted non-negative matrix factorisation for sound source separation», dans *Proc. IEEE/SP Workshop on Statistical Signal Processing*, p. 1132–1137. (page 35)

Fitzgerald, D., M. Cranitch et E. Coyle. 2008, «On the use of the beta divergence for musical source separation», dans *Proc. IET Irish Signals and Systems Conference (ISSC)*, Galway, Ireland. (page 25)

Flanagan, J. L. et R. M. Golden. 1966, «Phase vocoder», *Bell System Technical Journal*, vol. 45, p. 1493–1509. (page 20, 58)

Gaich, A. et P. Mowlaee. 2015, «On speech quality estimation of phase-aware single-channel speech enhancement», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, p. 216–220. (page 12)

Galembo, A., A. Askenfelt, L. L. Cudy et F. A. Russo. 2001, «Effects of relative phases on pitch and timbre in the piano bass range», *The Journal of the Acoustical Society of America*, vol. 110, n° 3, p. 1649–1666. (page 73, 75, 175)

Gerchberg, R. W. et W. O. Saxton. 1972, «A practical algorithm for the determination of phase from image and diffraction plane pictures», *Optik*, vol. 35, n° 2, p. 237–246. (page 3, 16)

Gerkmann, T. 2014, «MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, p. 5. (page 132, 138)

Gerkmann, T. et M. Krawczyk. 2013, «MMSE-optimal spectral amplitude estimation given the STFT-phase», *IEEE Signal Processing Letters*, vol. 20, n° 2, p. 129–132. (page 34)

Gerkmann, T., M. Krawczyk et R. Rehr. 2012, «Phase estimation in speech enhancement - Unimportant, important, or impossible ?», dans *Proc. IEEE Convention of Electrical Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, p. 1–5. (page 12, 21)

Gershman, S. J. et D. M. Blei. 2011, «A tutorial on Bayesian nonparametric models», *Journal of Mathematical Psychology*, vol. 56, n° 1, p. 1–12. (page 35)

Girin, L., S. Marchand, J. Di Martino, A. Röbel et G. Peeters. 2003, «Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 193–196. (page 20)

Gnann, V. et M. Spiertz. 2009, «Inversion of short-time fourier transform magnitude spectrograms with adaptive window lengths», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, p. 325–328. (page 16)

Gnann, V. et M. Spiertz. 2010, «Improving RTISI phase estimation with energy order and phase unwrapping», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, Graz, Austria. (page 16, 53)

Gobinet, C., A. Elhafid, V. Vrabie, R. Huez et D. Nuzillard. 2005, «About importance of positivity constraint for source separation in fluorescence spectroscopy», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, p. 1–4. (page 171)

Gobinet, C., E. Perrin et R. Huez. 2004, «Application of non-negative matrix factorization to fluorescence spectroscopy», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria. (page 170)

Godsill, S. et E. E. Kuruoglu. 1999, «Bayesian inference for time series with heavy-tailed symmetric α-stable noise processes», *Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics (Heavy Tails 99)*, p. 3–5. (page 146)

Godsill, S. J. et P. J. W. Rayner. 1998, *Digital Audio Restoration - A Statistical Model-Based Approach*, Springer-Verlag. (page 70)

Gonzalez, S. et M. Brookes. 2014, «PEFAC - A pitch estimation algorithm robust to high levels of noise», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, n° 2, p. 518–530, ISSN 2329-9290. (page 61)

Griffin, D. et J. S. Lim. 1984, «Signal estimation from modified short-time Fourier transform», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, n° 2, p. 236–243. (page 6, 15, 16, 42, 200)

Grosche, P. et M. Müller. 2011, «Tempogram Toolbox : MATLAB tempo and pulse analysis of music recordings», dans *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA. (page 62, 82, 96, 102, 118)

Guillamet, D. et J. Vitria. 2002, «Classifying faces with nonnegative matrix factorization», dans *Proc. Catalan conference for artificial intelligence*, Castellón, Spain, p. 24–31. (page 146)

Gunawan, D. et D. Sen. 2010, «Iterative phase estimation for the synthesis of separated sources from single-channel mixtures», *IEEE Signal Processing Letters*, vol. 17, n° 5, p. 421–424. (page 17, 79)

Hennequin, R. 2011, *Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale*, thèse de doctorat, Télécom ParisTech. (page 26)

Hennequin, R., R. Badeau et B. David. 2011a, «NMF with time–frequency activations to model nonstationary audio events», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n° 4, p. 744–753. (page 34, 113)

Hennequin, R., R. Badeau et B. David. 2011b, «Score informed audio source separation using a parametric model of non-negative spectrogram», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, p. 45–48. (page 32, 63)

Hoffman, M. D. 2012, «Poisson-uniform nonnegative matrix factorization», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, p. 5361–5364. (page 26)

Holighaus, N., M. Dörfler, G. A. Velasco et T. Grill. 2013, «A framework for invertible, real-time constant-Q transforms», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, n° 4, p. 775–785. (page 51)

Hoyer, P. O. 2004, «Non-negative matrix factorization with sparseness constraints», *Journal of Machine Learning Research*, vol. 5, p. 1457–1469. (page 33)

Hua, Y., A. B. Gershman et Q. Cheng. 2004, *High-resolution and robust signal processing*, Signal processing and communications, Marcel Dekker, New York, ISBN 0-8247-4752-6. (page 53, 101, 210)

Huang, P.-S., S. D. Chen, P. Smaragdis et M. Hasegawa-Johnson. 2012, «Singing-voice separation from monaural recordings using robust principal component analysis», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, p. 57–60. (page 5, 165)

Hunter, D. R. et K. Lange. 2004, «A tutorial on MM algorithms», *The American Statistician*, vol. 58, n° 1, p. 30–37. (page 28, 152)

Itakura, F. et S. Saito. 1968, «Analysis synthesis telephony based on the maximum likelihood method», dans *Proc. International Congress on Acoustics*, Tokyo, Japan, p. C17–C20. (page 24)

Iverson, P. et C. L. Krumhansl. 1993, «Isolating the dynamic attributes of musical timbre», *The Journal of the Acoustical Society of America*, vol. 94, n° 5, p. 2595–2603. (page 73)

Jaganathan, K., Y. C. Eldar et B. Hassibi. 2015, «Phase retrieval : An overview of recent developments», *CoRR*, vol. abs/1510.07713. (page 3)

Jaiswal, R., D. FitzGerald, D. Barry, E. Coyle et S. Rickard. 2011, «Clustering NMF basis functions using shifted NMF for monaural sound source separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, ISSN 1520-6149, p. 245–248. (page 35)

Janssen, A. J. E. M., R. N. J. Veldhuis et L. B. Vries. 1986, «Adaptive interpolation of discrete-time signals that can be modeled as AR processes», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, n° 2, p. 317–330. (page 69, 70)

Jaureguiberry, X., G. Richard, P. Leveau, R. Hennequin et E. Vincent. 2013, «Introducing a simple fusion framework for audio source separation», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, UK, ISSN 1551-2541, p. 1–6. (page 35)

Kahrs, M. et K. Brandenburg. 1998, *Applications of digital signal processing to audio and acoustics*, Springer Science & Business Media. (page 69)

Kameoka, H., N. Ono, K. Kashino et S. Sagayama. 2009, «Complex NMF : a new sparse representation for acoustic signals», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan. (page 6, 30, 35, 42, 50, 82, 112, 114, 115, 121, 122, 201, 204, 212)

Karp, T. et N. Fliege. 1999, «Modified DFT filter banks with perfect reconstruction», *IEEE Transactions on circuits and systems - II : Analog and digital signal processing*, vol. 46, n° 11, p. 1404–1414. (page 51)

Kehling, C., A. Jakob, D. Christian et S. Gerald. 2014, «Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, p. 8. (page 62, 96, 164, 168)

Kim, Y.-D. et S. Choi. 2009, «Weighted nonnegative matrix factorization», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, p. 1541–1544. (page 34)

King, B. et L. Atlas. 2010, «Single-channel source separation using simplified-training complex matrix factorization», dans *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, p. 4206–4209. (page 36)

King, B. et L. Atlas. 2011, «Single-channel source separation using complex matrix factorization», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n° 8, p. 2591–2597. (page 36)

King, B. et L. Atlas. 2012, «Complex Matrix Factorization Toolbox Version 1.0 for MATLAB», "https://sites.google.com/a/uw.edu/isdl/projects/cmf-toolbox". University of Washington. (page 36)

King, B. J. 2012, *New Methods of Complex Matrix Factorization for Single-Channel Source Separation and Analysis*, thèse de doctorat, University of Washington. (page 36, 49)

Kirchhoff, H., R. Badeau et S. Dixon. 2014, «Towards complex matrix decomposition of spectrogram based on the relative phase offsets of harmonic sounds», dans *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy. (page 37, 42, 110, 175)

Kompass, R. 2007, «A generalized divergence measure for nonnegative matrix factorization», *Neural Computation*, vol. 19, n° 3, p. 780–791. Neural Computation. (page 27)

Kounades-Bastian, D., L. Girin, X. Alameda-Pineda, S. Gannot et R. Horaud. 2016, «An inverse-gamma source variance prior with factorized parameterization for audio source separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, p. 136–140. (page 172, 176)

Koutsogiannaki, M., O. Simantiraki, G. Degottex et Y. Stylianou. 2014, «The importance of phase on voice quality assessment», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Singapore. (page 12)

Krawczyk, M. et T. Gerkmann. 2012, «STFT phase improvement for single channel speech enhancement», dans *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, p. 1–4. (page 21)

Krawczyk, M. et T. Gerkmann. 2014, «STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, n° 12, p. 1931–1940. (page 21)

Krawczyk, M. et T. Gerkmann. 2015, «MMSE-optimal combination of wiener filtering and harmonic model based speech enhancement in a general framework», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 1–5. (page 21, 137)

Kullback, S. et R. A. Leibler. 1951, «On information and sufficiency», *The Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79–86. (page 24)

Kulmer, J. et P. Mowlaee. 2015, «Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, p. 5063–5067. (page 132, 138)

Kwok, H. K. et D. L. Jones. 2000, «Improved instantaneous frequency estimation using an adaptive short-time Fourier transform», *IEEE Transactions on Signal Processing*, vol. 48, n° 10, p. 2964–2972, ISSN 1053-587X. (page 175)

Laroche, C., M. Kowalski, H. Papadopoulos et G. Richard. 2015, «A structured non-negative matrix factorization for source separation», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Nice, France. (page 33)

Laroche, C., M. Kowalski, H. Papadopoulos et G. Richard. 2016, «Genre specific dictionaries for harmonic/percussive source separation», dans *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, New York, NY, USA. (page 32)

Laroche, C., H. Papadopoulos, M. Kowalski et G. Richard. 2017, «Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA. (page 175)

Laroche, J. et M. Dolson. 1997, «Phase-vocoder : about this phasiness business», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA. (page 66, 73)

Laroche, J. et M. Dolson. 1999, «Improved phase vocoder time-scale modification of audio», *IEEE Transactions on Speech and Audio Processing*, vol. 7, n° 3, p. 323–332. (page 3, 59, 60, 61, 63, 73)

Le Roux, J. 2009, *Exploiting Regularities in Natural Acoustical Scenes for Monaural Audio Signal Estimation, Decomposition, Restoration and Modification*, thèse de doctorat, The University of Tokyo & Université Paris VI – Pierre et Marie Curie. (page 17)

Le Roux, J., A. D. Cheveigné et L. C. Parra. 2008a, «Adaptive template matching with shift-invariant semi-NMF», dans *Advances in Neural Information Processing Systems 21*, édité par D. Koller, D. Schuurmans, Y. Bengio et L. Bottou, Curran Associates, Inc., p. 921–928. (page 22)

Le Roux, J., H. Kameoka, N. Ono, A. D. Cheveigné et S. Sagayama. 2008b, «Computational auditory induction by missing-data non-negative matrix factorization», dans *Proc. ITRW on Statistical and Perceptual Audio Processing*, Brisbane, Australia. (page 22)

Le Roux, J., H. Kameoka, N. Ono et S. Sagayama. 2010a, «Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, New Orleans, LA, USA, p. 397–403. (page 19)

Le Roux, J., H. Kameoka, E. Vincent, N. Ono, K. Kashino et S. Sagayama. 2009, «Complex NMF under spectrogram consistency constraints», dans *Proc. Acoustical Society of Japan Autumn Meeting*, Hukushima, Japan. (page 30, 36, 42, 89)

Le Roux, J., N. Ono et S. Sagayama. 2008c, «Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction», dans *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Brisbane, Australia, p. 23–28. (page 6, 14, 16, 17, 36, 42, 200)

Le Roux, J. et E. Vincent. 2013, «Consistent Wiener filtering for audio source separation», *IEEE Signal Processing Letters*, vol. 20, n° 3, p. 217–220. (page 6, 20, 49, 84, 87, 141)

Le Roux, J., E. Vincent, Y. Mizuno, H. Kameoka, N. Ono et S. Sagayama. 2010b, «Consistent Wiener filtering : generalized time-frequency masking respecting spectrogram consistency», dans *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Saint-Malo, France, p. 89–96. (page 19)

Le Roux, J., F. Weninger et J. Hershey. 2015, «Sparse NMF - half-baked or well done ?», Tech. Rep. TR2015-023, Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. (page 33)

Lee, D. D. et H. S. Seung. 1999, «Learning the parts of objects by non-negative matrix factorization», *Nature*, vol. 401, n° 6755, p. 788–791. (page 5, 22, 24, 27, 151, 172)

Lee, D. D. et H. S. Seung. 2001, «Algorithms for non-negative matrix factorization», dans *Advances in Neural Information Processing Systems 13*, MIT Press, p. 556–562. (page 27, 29, 45)

Leglaive, S., R. Badeau et G. Richard. 2016a, «Autoregressive moving average modeling of late reverberation in the frequency domain», dans *Proc. European Signal Processing Conference (EUSIPCO)*, édité par EURASIP, Budapest, Hungary. (page 175)

Leglaive, S., R. Badeau et G. Richard. 2016b, «Multichannel audio source separation with probabilistic reverberation priors», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 12, n° 99, p. 2453–2465, ISSN 2329-9290. (page 4)

Limem, A., G. Delmaire, M. Puigt, G. Roussel et D. Courcot. 2013, «Non-negative matrix factorization using weighted beta divergence and equality constraints for industrial source apportionment», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, United Kingdom, p. 1–6. (page 34, 168)

Liu, P., X. Zhou, Y. Li, M. Li, D. Yu et J. Liu. 2013, «The application of principal component analysis and non-negative matrix factorization to analyze time-resolved optical waveguide absorption spectroscopy data», *Analytical Methods*, vol. 5, p. 4454–4459. (page 22, 146, 170)

Liutkus, A. et R. Badeau. 2015, «Generalized Wiener filtering with fractional power spectrograms», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia. (page 13, 26, 78, 147, 163, 164, 172)

Liutkus, A., J.-L. Durrieu, L. Daudet et G. Richard. 2013, «An overview of informed audio source separation», dans *Proc. International Workshop on Image and Audio Analysis for Multimedia Interactive services (WIAMIS)*, Paris, France, p. 1–4. (page 32)

Liutkus, A., D. Fitzgerald et R. Badeau. 2015, «Cauchy nonnegative matrix factorization», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA. (page 26, 148, 165, 167)

Liutkus, A., J. Pinel, R. Badeau, L. Girin et G. Richard. 2012, «Informed source separation through spectrogram coding and data embedding», *Signal Processing*, vol. 92, n° 8, p. 1937–1949. (page 33)

Magron, P., R. Badeau et B. David. 2015a, «Phase reconstruction of spectrograms based on a model of repeated audio events», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA. (page 94, 100)

Magron, P., R. Badeau et B. David. 2015b, «Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Nice, France. (page 58)

Magron, P., R. Badeau et B. David. 2015c, «Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration», Tech. Rep. 2015D002, Télécom ParisTech, Paris, France. (page 58)

Magron, P., R. Badeau et B. David. 2015d, «Phase recovery in NMF for audio source separation : an insightful benchmark», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, p. 81–85. (page 42)

Magron, P., R. Badeau et B. David. 2016a, «An iterative algorithm for recovering the phase of complex components from their mixture», Tech. Rep. HAL-01325625, Paris, France. (page 78)

Magron, P., R. Badeau et B. David. 2016b, «Complex NMF under phase constraints based on signal modeling : application to audio source separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China. (page 112, 114, 121)

Magron, P., R. Badeau et B. David. 2017a, «Phase-dependent anisotropic Gaussian model for audio source separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA. (page 130)

Magron, P., R. Badeau et B. David. 2017b, «STFT phase recovery by sinusoidal modeling for audio source separation», *IEEE Transactions on Audio, Speech and Language Processing*. (submitted). (page 78)

Magron, P., R. Badeau et A. Liutkus. 2016c, «Generalized Wiener filtering for positive alpha-stable random variables», Tech. Rep. 2016D004, Télécom ParisTech, Paris, France. (page 146)

Magron, P., R. Badeau et A. Liutkus. 2017c, «Lévy NMF for robust nonnegative source separation», *IEEE Signal Processing Letters*. (submitted). (page 146)

Mallat, S. 1998, *A wavelet tour of signal processing*, Academic press. (page 175)

Mallat, S. et I. Waldspurger. 2015, «Phase retrieval for the Cauchy wavelet transform», *Journal of Fourier Analysis and Applications*, vol. 21, n° 6, p. 1251–1309. (page 16)

Mardia, K. V. et P. E. Jupp. 2000, *Directional Statistics*. (page 130)

Mardia, K. V. et P. Zemroch. 1975, «Algorithm AS 86 : The Von Mises distribution function», *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, n° 2, p. 268–272. (page 21, 116, 131, 134)

Marques, J. et L. Almeida. 1986, «A background for sinusoid based representation of voiced speech», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 11, Tokyo, Japan, p. 1233–1236, doi :10.1109/ICASSP.1986.1168731. (page 61)

McAuley, R. J. et T. F. Quatieri. 1986, «Speech analysis/synthesis based on a sinusoidal representation», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, n° 4, p. 744–754. (page 20, 58)

Montcuquet, A.-S., L. Herve, L. Guyon, J.-M. Dinten et J. I. Mars. 2009, «Non-negative matrix factorization : A blind sources separation method to unmix fluorescence spectra», dans *Proc. Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Grenoble, France, p. 1–4. (page 170)

Mowlaee, P. et J. Kulmer. 2015, «Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, n° 9, p. 1521–1532. (page 58, 132)

Mowlaee, P. et R. Martin. 2012, «On phase importance in parameter estimation for single-channel source separation», dans *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, p. 1–4. (page 12, 21)

Mowlaee, P. et R. Saeidi. 2013, «On phase importance in parameter estimation in single-channel speech enhancement», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada. (page 21)

Mowlaee, P. et R. Saeidi. 2014, «Time-frequency constraints for phase estimation in single-channel speech enhancement», dans *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, p. 337–341. (page 21)

Mowlaee, P., R. Saeidi et Y. Stylianou. 2014, «Phase importance in speech processing applications», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Singapore. (page 21)

Mowlaee, P., R. Saeidi et Y. Stylianou. 2016, «Advances in phase-aware signal processing in speech communication», *Speech Communication*, vol. 81, p. 1 – 29, ISSN 0167-6393. Phase-Aware Signal Processing in Speech Communication. (page 3)

Mowlaee, P., R. Saiedi et R. Martin. 2012, «Phase estimation for signal reconstruction in single-channel speech separation», dans *Proc. International Conference on Spoken Language Processing*, Portland, OR, USA, p. 1–4. (page 3, 21, 79)

Mowlaee, P., M. K. Watanabe et R. Saeidi. 2013, «Show & tell : phase-aware single-channel speech enhancement», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France. (page 21)

Mysore, G. J., P. Smaragdis et B. Raj. 2010, «Non-negative hidden Markov modeling of audio with application to source separation», dans *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, St. Malo, France, p. 140–148. (page 26)

Nawab, S. H., T. F. Quatieri et J. S. Lim. 1983, «Signal reconstruction from short-time Fourier transform magnitude», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, n° 4, p. 986–998. (page 15)

Nolan, J. P. 2015, *Stable Distributions - Models for Heavy Tailed Data*, Birkhauser, Boston. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan. (page 13, 26, 146, 147)

Ono, N., Z. Rafii, D. Kitamura, N. Ito et A. Liutkus. 2015, «The 2015 signal separation evaluation campaign», dans *Latent Variable Analysis and Signal Separation*, Springer, p. 387–395. (page 63, 84, 121, 140, 164, 165)

Ozerov, A. et C. Févotte. 2010, «Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 3, p. 550–563. (page 4)

Paliwal, K., K. Wójcicki et B. Shannon. 2011, «The importance of phase in speech enhancement», *Speech Communication*, vol. 53, n° 4, p. 465–494. (page 12)

Paliwal, K. K. et L. D. Alsteris. 2003, «Usefulness of phase spectrum in human speech perception», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Geneva, Switzerland. (page 12)

Paliwal, K. K. et L. D. Alsteris. 2005, «On the usefulness of STFT phase spectrum in human listening tests», *Speech Communication*, vol. 45, n° 2, p. 153–170. (page 12)

Parry, R. M. et I. Essa. 2007, «Incorporating phase information for source separation via spectrogram factorization», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Honolulu, Hawaii, USA. (page 34, 130)

Pauca, V. P., F. Shahnaz, M. W. Berry et R. J. Plemmons. 2004, «Text mining using non-negative matrix factorizations», dans *Proc. SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, USA, p. 452–456. (page 22, 146)

Paulus, J. et T. Virtanen. 2005, «Drum transcription with non-negative spectrogram factorisation», dans *Proc. European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, p. 1–4. (page 118)

Perraudin, N., P. Balazs et P. L. Sondergaard. 2013, «A fast Griffin-Lim algorithm», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 1–4. (page 16)

Plumbley, M. D., T. Blumensath, L. Daudet, R. Gribonval et M. E. Davies. 2010, «Sparse representations in audio and music : From coding to source separation», *Proc. of the IEEE*, vol. 98, n° 6, p. 995–1005. (page 51)

Princen, J. P. et A. B. Bradley. 1986, «Analysis/synthesis filter bank design based on time domain aliasing cancellation», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, n° 5, p. 1153–1161. (page 51)

Puckette, M. 1995, «Phase-locked vocoder», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 222–225. (page 73)

Raki, S. A., S. Makino, H. Sawada et R. Mukai. 2005, «Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Philadelphia, PA, USA, ISSN 1520-6149, doi :10.1109/ICASSP.2005.1415651. (page 51)

Rigaud, F., B. David et L. Daudet. 2013, «A parametric model and estimation techniques for the inharmonicity and tuning of the piano», *Journal of the Acoustical Society of America (JASA)*, vol. 133, n° 5, p. 3107–3118. (page 33, 99)

Röbel, A. 2003a, «A new approach to transient processing in the phase vocoder», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, London, United Kingdom, p. 344–349. (page 68)

Röbel, A. 2003b, «Transient detection and preservation in the phase vocoder», dans *Proc. International Computer Music Conference (ICMC)*, Singapore, p. 247–250. (page 68, 175)

Robert, C. et G. Casella. 2013, *Monte Carlo statistical methods*, Springer Science & Business Media. (page 132)

Rodriguez-Serrano, F. J., S. Ewert, P. Vera-Candeas et M. Sandler. 2016, «A score-informed shift invariant extension of complex matrix factorisation for improving the separation of overlapped partials in music recordings», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai, China. (page 113)

Rogers, G. L. 2008, «Multiple path analysis of reflectance from turbid media», *Journal of the Optical Society of America*, vol. 25, n° 11, p. 2879–2883. (page 172)

Sajda, P., S. Du, T. R. Brown, R. Stoyanova, D. C. Shungu, X. Mao et L. C. Parra. 2004, «Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain», *IEEE Transactions on Medical Imaging*, vol. 23, n° 12, p. 1453–1465. (page 146, 172)

Samoradnitsky, G. et M. S. Taqqu. 1994, *Stable non-Gaussian random processes : stochastic models with infinite variance*, vol. 1, CRC Press. (page 146, 161)

Sawada, H., H. Kameoka, S. Araki et N. Ueda. 2011, «Formulations and algorithms for multichannel complex NMF», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, p. 229–232. (page 36)

Schmidt, M. N. et H. Laurberg. 2008, «Nonnegative matrix factorization with Gaussian process priors», *Computational Intelligence and Neuroscience*, vol. 2008, n° 3, p. 3 :1–3 :10. (page 26)

Schorkhuber, C., A. Klapuri, N. Holighaus et M. Dorfler. 2014, «A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution», dans *Proc. Conference on Semantic Audio*, London, UK. (page 51)

Shannon, B. J. et K. K. Paliwal. 2006, «Role of phase estimation in speech enhancement.», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Pittsburgh, PA, USA. (page 12)

Shashanka, M., B. Raj et P. Smaragdis. 2008, «Probabilistic latent variable models as nonnegative factorizations», *Computational Intelligence and Neuroscience*, vol. 2008. (page 26)

Simsekli, U., A. Liutkus et A. T. Cemgil. 2015, «Alpha-stable matrix factorization», *IEEE Signal Processing Letters*, vol. 22, n° 12, p. 2289–2293. (page 148, 176)

Smaragdis, P., R. Bhiksha et S. Madhusudana. 2007, «Supervised and semi-supervised separation of sounds from single-channel mixtures», dans *Independent Component Analysis and Signal Separation*, Springer, p. 414–421. (page 26)

Smaragdis, P. et J. C. Brown. 2003, «Non-negative matrix factorization for polyphonic music transcription», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA. (page 22, 33)

Smaragdis, P., B. Raj et M. Shashanka. 2006, «A probabilistic latent variable model for acoustic modeling», dans *Proc. Workshop on Advances in Models for Acoustic Processing at NIPS*, Whistler, Canada. (page 5, 26)

Spiertz, M. et V. Gnann. 2009, «Source-filter based clustering for monaural blind source separation», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, Como, Italy. (page 35)

Stöter, F.-R., A. Liutkus, R. Badeau, B. Edler et P. Magron. 2016, «Common fate model for unison source separation», dans *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China. (page 46)

Sturmel, N. et L. Daudet. 2011, «Signal reconstruction from its STFT magnitude : a state of the art», dans *Proc. International Conference on Digital Audio Effects (DAFx)*, Paris, France. (page 20)

Sturmel, N. et L. Daudet. 2012, «Iterative phase reconstruction of Wiener filtered signals», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 101–104. (page 20)

Sturmel, N. et L. Daudet. 2013, «Informed source separation using iterative reconstruction», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, n° 1, p. 178–185. (page 6, 20)

Stylianou, Y. 2001, «Removing linear phase mismatches in concatenative speech synthesis», *IEEE Transactions on Speech and Audio Processing*, vol. 9, n° 3, p. 232–239, ISSN 1063-6676. (page 3, 20)

Sugiyama, A. et R. Miyahara. 2013a, «Phase randomization-a new paradigm for single-channel signal enhancement», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, p. 7487–7491. (page 12)

Sugiyama, A. et R. Miyahara. 2013b, «Tapping-noise suppression with magnitude-weighted phase-based detection», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 1–4. (page 74)

Sun, D. L. et I. J. O. Smith. 2012, «Estimating a signal from a magnitude spectrogram via convex optimization», dans *Proc. Audio Engineering Society Convention 133*, San Fransisco, CA, USA. (page 6, 19)

Sunnydayal, V. et T. K. Kumar. 2015, «Bayesian estimation for speech enhancement given a priori knowledge of clean speech phase», *International Journal of Speech Technology*, vol. 18, n° 4, p. 593–607. (page 21)

Tan, V. Y. et C. Févotte. 2005, «A study of the effect of source sparsity for various transforms on blind audio source separation performance», dans *Proc. Workshop on Signal*

Processing with Adaptive Sparse Structured Representations (SPARS), Rennes, France, p. 16–18. (page 51)

Vincent, E. 2010, «An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation», dans *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Saint-Malo, France, p. 157–164. (page 6, 14)

Vincent, E. 2012, «Improved perceptual metrics for the evaluation of audio source separation», dans *Latent Variable Analysis and Signal Separation*, Springer, p. 430–437. (page 39)

Vincent, E., R. Gribonval et C. Févotte. 2006, «Performance measurement in blind audio source separation», *IEEE Transactions on Speech and Audio Processing*, vol. 14, n° 4, p. 1462–1469. (page 38, 45, 63, 84, 97, 121, 140, 168)

Vincent, E., M. G. Jafari, S. A. Abdallah, M. D. Plumbley et M. E. Davies. 2010, «Probabilistic modeling paradigms for audio source separation», dans *Machine Audition : Principles, Algorithms and Systems*, édité par W. Wang, IGI Global, p. 162–185. (page 25)

Vincent, E., H. Sawada, P. Bofill, S. Makino et J. P. Rosca. 2007, «First stereo audio source separation evaluation campaign : data, algorithms and results», dans *Independent Component Analysis and Signal Separation*, Springer, p. 552–559. (page 38)

Virtanen, T. 2007, «Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 3, p. 1066–1074. (page 13, 22, 33)

Virtanen, T., A. T. Cemgil et S. Godsill. 2008, «Bayesian extensions to non-negative matrix factorisation for audio signal modelling», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, p. 1825–1828. (page 25, 146, 164)

Virtanen, T. et A. Klapuri. 2006, «Analysis of polyphonic audio using source-filter model and non-negative matrix factorization», dans *Proc. Neural Information Processing Systems Workshop Advances in Models for Acoustic Processing,*. (page 34)

Voran, S. 2015, «Exploration of the additivity approximation for spectral magnitudes», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, p. 1–5. (page 26, 172, 176)

Wang, B. et M. D. Plumbley. 2005, «Musical audio stream separation by nonnegative matrix factorization», dans *Proc. UK Digital Music Research Network (DMRN) Summer Conference*, Glasgow, United Kingdom, p. 23–4. (page 3, 22)

Wang, D. L. et J. S. Lim. 1982, «The unimportance of phase in speech enhancement», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, n° 4, p. 679–681. (page 3, 12)

Watanabe, M. K. et P. Mowlaee. 2013, «Iterative sinusoidal-based partial phase reconstruction in single-channel source separation», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, p. 832–836. (page 20)

Watson, G. N. 1995, *A treatise on the theory of Bessel functions*, Cambridge university press. (page 131, 137)

Weron, R. 2010, «STABLERND : MATLAB function to generate random numbers from the stable distribution», Statistical Software Components, Boston College Department of Economics. (page 166)

Wiener, N. 1949, *The extrapolation, interpolation and smoothing of stationary time series with engineering applications*, Wiley, John and sons, Inc. (page 13)

Wilson, K. W., B. Raj, P. Smaragdis et A. Divakaran. 2008, «Speech denoising using nonnegative matrix factorization with priors», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, p. 4029–4032. (page 22)

Wirtinger, W. 1927, «Zur formalen theorie der funktionen von mehr komplexen veränderlichen», *Mathematische Annalen*, vol. 97, n° 1, p. 357–375. (page 202)

Witkovsky, V. 2001, «Computing the distribution of a linear combination of inverted gamma variables», *Kybernetika*, vol. 37, n° 1, p. 79–90. (page 172)

Yilmaz, Ö. et S. Rickard. 2004, «Blind separation of speech mixtures via time-frequency masking», *IEEE Transaction on Signal Processing*, vol. 52, n° 7, p. 1830–1847. (page 13)

Yoshii, K., R. Tomioka, D. Mochihashi et M. Goto. 2013, «Beyond NMF : Time-domain audio source separation without phase reconstruction», dans *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, Curitiba, Brazil. (page 22, 46)

Zhang, Y. et Y. Fang. 2007, «A NMF algorithm for blind separation of uncorrelated signals», dans *Proc. International Conference on Wavelet Analysis and Pattern Recognition*, vol. 3, Beijing, China, p. 999–1003. (page 33)

Zhu, X., G. T. Beauregard et L. L. Wyse. 2007, «Real-time signal estimation from modified short-time Fourier transform magnitude spectra», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 5, p. 1645–1653. (page 16, 73)

Annexe A

Transformée de Fourier à Court Terme

Le cadre approprié pour l'analyse de signaux musicaux est une représentation Temps-Fréquence (TF), dans laquelle nous pouvons exploiter la propriété de parcimonie de ces signaux. Nous rappelons ici quelques définitions et propriétés élémentaires de la Transformée de Fourier à Court Terme (TFCT) qui est la principale transformation utilisée dans cette thèse.

Définition L'expression de la TFCT X d'un signal x, pour chaque bande de fréquences $f \in [\![0, F-1]\!]$ et pour chaque trame temporelle $t \in \mathbb{Z}$ est :

$$X(f,t) = \sum_{n=0}^{N_w-1} x(n+tS) w_a(n) e^{-2i\pi \frac{f}{F} n}, \tag{A.1}$$

où w_a est une fenêtre d'analyse de longueur N_w (elle est nulle en dehors de son support, donc on peut sans perte de généralité écrire la somme sur $\mathbb{Z}$) et S est le décalage temporel (exprimé en échantillons) entre deux trames successives.

TFCT inverse Pour resynthétiser un signal temporel à partir d'une TFCT Y, on applique la Transformée de Fourier Discrète (TFD) inverse à chaque trame, et celle-ci est multipliée par une fenêtre appelée fenêtre de synthèse et notée w_s :

$$y_s(t,n) = \sum_{f=0}^{F-1} Y(f,t) w_s(n) e^{2i\pi \frac{f}{F} n}. \tag{A.2}$$

On ajoute ensuite les signaux obtenus y_s pour reconstituer le signal de départ par addition-recouvrement (procédure *Overlap-Add* ou OLA) :

$$y(n) = \sum_{t \in \mathbb{Z}} y_s(t, n - tS). \tag{A.3}$$

Il est à noter que plusieurs conventions d'écriture de la TFCT inverse existent. En particulier, la plupart des expressions utilisées dans la littérature utilisent un facteur $1/F$ dans l'expression de la transformée inverse.

Reconstruction parfaite La reconstruction parfaite est obtenue lorsque $\forall n \in \mathbb{Z}$, $y(n) = x(n)$, où y est la TFCT inverse de la TFCT de x. On a :

$$\begin{aligned} y(n) &= \sum_{t\in\mathbb{Z}} y_s(t, n - tS) \\ &= \sum_{t\in\mathbb{Z}} \sum_{f=0}^{F-1} Y(f,t) w_s(n - tS) e^{2i\pi \frac{f}{F}(n-tS)} \\ &= \sum_{t\in\mathbb{Z}} \sum_{f=0}^{F-1} \sum_{m\in\mathbb{Z}} x(m + tS) w_a(m) e^{-2i\pi \frac{f}{F} m} w_s(n - tS) e^{2i\pi \frac{f}{F}(n-tS)} \\ &= \sum_{t\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} x(m + tS) w_a(m) w_s(n - tS) \sum_{f=0}^{F-1} e^{2i\pi \frac{f}{F}(n-tS-m)}. \end{aligned}$$

Or, $\sum_{f=0}^{F-1} e^{2i\pi \frac{f}{F}(n-tS-m)} = 0$ en général, sauf dans le cas où $n - tS - m = rF$ (c'est-à-dire $m = n - tS + rF$), $r \in \mathbb{Z}$, auquel cas chaque terme de la somme vaut 1, donc la somme vaut F. On a donc [1] :

$$\begin{aligned} y(n) &= \sum_{t\in\mathbb{Z}} x(n - tS + tS) F w_a(n - tS) w_s(n - tS) \\ &= x(n) \sum_{t\in\mathbb{Z}} F w_a(n - tS) w_s(n - tS). \end{aligned}$$

Il y a donc reconstruction parfaite lorsque, $\forall n \in \mathbb{Z}$:

$$\sum_{t\in\mathbb{Z}} w_a(n - tS) w_s(n - tS) = \frac{1}{F}. \tag{A.4}$$

Il est à noter qu'en pratique, les signaux audio que nous traitons n'ont pas une durée infinie. Aussi, $n \in [\![0, N_x - 1]\!]$ et il n'y a donc qu'un certain nombre de trames T qui sont retenues pour la TFCT. On peut alors voir une TFCT X comme une matrice à F lignes et T colonnes dont les entrées sont des nombres complexes : $X \in \mathbb{C}^{F\times T}$. La condition de reconstruction parfaite peut alors être réécrite : en notant $Q = \frac{N_w}{S}$, on a $\forall n \in [\![0, N_x - 1]\!]$:

$$\sum_{q=0}^{Q-1} w_a(n - tS) w_s(n - tS) = \frac{1}{F}, \tag{A.5}$$

condition qui a largement été décrite et exploitée dans la littérature Griffin et Lim (1984); Le Roux et al. (2008c).

Dans tout ce manuscrit, on utilise la même fenêtre pour l'analyse et la synthèse, (c'est-à-dire $w_a = w_s$), et celles-ci sont normalisées de sorte à vérifier la condition de reconstruction parfaite (A.5). Il est intéressant de voir que la condition de reconstruction parfaite dépend non seulement de la nature de la fenêtre, mais également du taux de recouvrement de la transformée (via le paramètre Q). On peut vérifier expérimentalement (pour trois types de fenêtres : Hann, Hamming et Blackman) que des taux de recouvrement vérifiant que N_w est un multiple de $2S$ permettent d'atteindre la reconstruction parfaite. En particulier, les taux de 50 %, 75 % et 87.5 % conduisent à une reconstruction parfaite pour ces trois types de fenêtres. Dans la quasi-totalité de ce manuscrit, on utilisera un taux de 75 %, à l'exception de certaines expériences du chapitre 4 où l'on adopte un taux de 50 ou de 87.5 %.

1. On ne retient en pratique que le cas $r = 0$ car la fenêtre d'analyse étant nulle en dehors de son support, les autres valeurs de r ont une contribution nulle.

Annexe B

Procédure itérative de séparation de sources par la méthode de la fonction auxiliaire

La procédure itérative qui permet la minimisation de la fonction de coût introduite dans le chapitre 5 peut être obtenue de façon rigoureuse par l'application de la méthode de la fonction auxiliaire, employée notamment dans Kameoka et al. (2009). On considère la fonction de coût suivante (on retire les indices (f,t) pour plus de lisibilité, ce qui ne nuit pas à la rigueur de la preuve puisque tous les points TF peuvent être traités indépendemment) :

$$\mathcal{C}(\theta) = |E|^2 = |X - \sum_k \hat{X}_k|^2, \tag{B.1}$$

avec $\theta = \{\hat{X}_k, k \in [\![1, K]\!]\}$, sous la contrainte $|\hat{X}_k| = V_k$ pour tout k.

Fonction auxiliaire L'idée est d'introduire une fonction *auxiliaire* $g(\theta, \tilde{\theta})$ dépendant de nouveaux paramètres auxiliaires $\tilde{\theta}$, et qui vérifie :

$$\mathcal{C}(\theta) = \min_{\tilde{\theta}} g(\theta, \tilde{\theta}). \tag{B.2}$$

On peut alors montrer (comme on l'a rappelé dans la section 2.2.4 du chapitre 2) que f est décroissante sous les règles de mises à jour suivantes :

$$\tilde{\theta} \leftarrow \arg\min_{\tilde{\theta}} g(\theta, \tilde{\theta}) \text{ et } \theta \leftarrow \arg\min_{\theta} g(\theta, \tilde{\theta}). \tag{B.3}$$

Nous cherchons donc à construire une fonction auxiliaire à $\mathcal{C}$. On introduit les variables auxiliaires $\tilde{\theta} = \{Y_k, k \in [\![1, K]\!]\}$ telles que $\sum_k Y_k = X$. On a alors :

$$|X - \sum_k \hat{X}_k|^2 = |\sum_k (Y_k - \hat{X}_k)|^2. \tag{B.4}$$

On introduit alors des termes λ_k positifs qui vérifient $\sum_k \lambda_k = 1$, et on peut écrire :

$$|X - \sum_k \hat{X}_k|^2 = \left| \sum_k \lambda_k \left(\frac{Y_k - \hat{X}_k}{\lambda_k} \right) \right|^2. \tag{B.5}$$

En appliquant l'inégalité de Jensen à la fonction convexe $|.|^2$, on a alors :

$$|X - \sum_k \hat{X}_k|^2 \leq \sum_k \frac{|Y_k - \hat{X}_k|^2}{\lambda_k}. \tag{B.6}$$

Ainsi, $\mathcal{C}(\theta) \leq g(\theta, \tilde{\theta})$ avec :

$$g(\theta, \tilde{\theta}) = \sum_k \frac{|Y_k - \hat{X}_k|^2}{\lambda_k}, \tag{B.7}$$

et le problème devient alors celui de la minimisation de g sous contraintes $\sum_k Y_k = X$ et $\forall k$, $|\hat{X}_k| = V_k$. Montrons que g est une fonction auxiliaire de f, c'est-à-dire qu'elle vérifie (B.2). Pour ce faire, on introduit la contrainte sur les variables auxiliaires dans g via la méthode des multiplicateurs de Lagrange :

$$\mathcal{L}(\theta, \tilde{\theta}, \gamma) = g(\theta, \tilde{\theta}) + \gamma(\sum_k \bar{Y}_k - \bar{X}). \tag{B.8}$$

où $\bar{.}$ désigne ici la conjugaison complexe. La minimisation de g par rapport à $\tilde{\theta}$ se fait alors par la recherche d'un point selle de $\mathcal{L}$. On va donc calculer les dérivées partielles de $\mathcal{L}$ par rapport aux variables complexes Y_k, c'est-à-dire en calculant des dérivées au sens de Wirtinger Wirtinger (1927). En pratique on dérive par rapport à la variable $\bar{Y}_k$ que l'on traite comme une variable réelle, et en considérant Y_k comme une constante Bouboulis (2010). Cela donne :

$$\frac{\partial \mathcal{L}}{\partial Y_k}(\theta, \tilde{\theta}, \gamma) = \frac{1}{\lambda_k}(Y_k - \hat{X}_k) + \gamma, \tag{B.9}$$

ce qui, une fois annulé, conduit à :

$$Y_k = \hat{X}_k + \lambda_k \gamma. \tag{B.10}$$

Par ailleurs, l'annulation de la dérivée partielle de $\mathcal{L}$ par rapport au multiplicateur de Lagrange γ conduit à la contrainte $\sum_k Y_k = X$. En sommant alors les égalités (B.10) et en utilisant cette contrainte, on a :

$$X = \sum_k Y_k = \sum_k \hat{X}_k + \gamma \sum_k \lambda_k, \tag{B.11}$$

et comme les λ_k doivent sommer à 1, on obtient :

$$\gamma = X - \sum_k \hat{X}_k, \tag{B.12}$$

ce qui mène, en réinjectant cette relation dans (B.10) à :

$$Y_k = \hat{X}_k + \lambda_k(X - \sum_k \hat{X}_k). \tag{B.13}$$

Ainsi, la valeur minimale de $g(\theta, \tilde{\theta})$ est obtenue pour un jeu de paramètres auxiliaires $\tilde{\theta}_m$ donné par (B.13). g vaut alors :

$$\begin{aligned} g(\theta, \tilde{\theta}_m) &= \sum_k \frac{|\hat{X}_k + \lambda_k(X - \sum_k \hat{X}_k) - \hat{X}_k|^2}{\lambda_k} \\ &= \sum_k \lambda_k |X - \sum_k \hat{X}_k|^2 \\ &= |X - \sum_k \hat{X}_k|^2 \sum_k \lambda_k \\ &= |X - \sum_k \hat{X}_k|^2 \\ &= \mathcal{C}(\theta), \end{aligned}$$

ce qui montre que g est bien une fonction auxiliaire de $\mathcal{C}$.

Mises à jour On peut alors obtenir les mises à jours des paramètres grâce à (B.3). La mise à jour des Y_k est donnée par (B.13), comme on l'a déjà montré. Pour obtenir les mises à jour des $\hat{X}_k$, on introduit, $\forall k$, les contraintes $|\hat{X}_k| = V_k$ à nouveau par la méthode des multiplicateurs de Lagrange :

$$\mathcal{H}(\theta, \tilde{\theta}, \gamma) = g(\theta, \tilde{\theta}) + \sum_k \delta_k (|\hat{X}_k|^2 - V_k^2). \tag{B.14}$$

On calcule ensuite les dérivées partielles de $\mathcal{H}$ par rapport aux variables $\hat{X}_k$ comme précédemment (par la méthode des dérivées de Wirtinger) :

$$\frac{\partial \mathcal{H}}{\partial \hat{X}_k}(\theta, \tilde{\theta}) = \frac{1}{\lambda_k}(\hat{X}_k - Y_k) + \delta_k \hat{X}_k, \tag{B.15}$$

et annuler cette dérivée conduit à :

$$\hat{X}_k = \frac{Y_k}{1 + \lambda_k \delta_k}. \tag{B.16}$$

Par ailleurs, l'annulation de la dérivée partielle de $\mathcal{H}$ par rapport aux multiplicateurs de Lagrange δ_k conduit là encore aux contraintes $|\hat{X}_k| = V_k$. En prenant le module dans (B.16) et en utilisant cette contrainte, on obtient donc :

$$V_k = |\hat{X}_k| = \frac{|Y_k|}{|1 + \lambda_k \delta_k|}, \tag{B.17}$$

soit :

$$1 + \lambda_k \delta_k = \pm \frac{|Y_k|}{V_k}, \tag{B.18}$$

et donc finalement, en réinjectant cette relation dans (B.16) :

$$\hat{X}_k = \pm V_k \frac{Y_k}{|Y_k|}. \tag{B.19}$$

Afin de lever l'ambiguïté sur le signe dans (B.19), on peut calculer la valeur de g dans chacun des cas. On constate aisément que, $\forall k \in [\![1, K]\!]$:

$$|Y_k - V_k \frac{Y_k}{|Y_k|}| = ||Y_k| - V_k|, \tag{B.20}$$

et :

$$|Y_k + V_k \frac{Y_k}{|Y_k|}| = ||Y_k| + V_k|. \tag{B.21}$$

Comme $|Y_k| \geq 0$ et $V_k \geq 0$, il est évident que $||Y_k| - V_k| \leq ||Y_k| + V_k|$. Ainsi, on retient la solution conduisant à la valeur minimale de g, donc :

$$\hat{X}_k = V_k \frac{Y_k}{|Y_k|}. \tag{B.22}$$

Remarque : On aurait pu obtenir la même mise à jour sans tenir compte explicitement de la contrainte $|\hat{X}_k| = V_k$ que l'on a introduite via les multiplicateurs de Lagrange. On aurait alors obtenu la mise à jour $\hat{X}_k = Y_k$, qui, une fois projetée sur l'espace contrainte (via une normalisation), aurait de nouveau donné (B.22). C'est ce type de procédé qui est utilisé dans Kameoka et al. (2009) pour tenir compte de la normalisation des colonnes W dans le modèle de NMF complexe.

En conclusion, on peut, en vertu de (B.3), alterner les mises à jour sur Y_k et $\hat{X}_k$ en utilisant (B.13) et (B.22), ce qui donne lieu à la procédure que nous avons présentée dans l'Algorithme 3 au chapitre 5, avec une garantie de décroissance de la fonction de coût associée.

Impact d'une initialisation avec la phase du mélange Une idée intuitive consiste à initialiser les composantes en leur donnant la phase du mélange. Nous montrons ici qu'il s'agit alors d'un point fixe de la procédure. Avec une telle initialisation, on a :

$$\hat{X}_k^{(0)} = V_k \frac{X}{|X|}, \tag{B.23}$$

et l'erreur vaut :

$$E^{(0)} = X - \sum_l V_l \frac{X}{|X|} = \frac{X}{|X|}\left(|X| - \sum_l V_l\right), \tag{B.24}$$

ce qui mène à :

$$Y_k^{(1)} = \hat{X}_k^{(0)} + \lambda_k E^{(0)} = V_k \frac{X}{|X|} + \lambda_k \frac{X}{|X|}\left(|X| - \sum_l V_l\right) = \frac{X}{|X|}\left(V_k + \lambda_k\left(|X| - \sum_l V_l\right)\right). \tag{B.25}$$

Après normalisation, on obtient finalement :

$$\hat{X}_k^{(1)} = sg(\alpha_k) V_k \frac{X}{|X|} = sg(\alpha_k)\hat{X}_k^{(0)}, \tag{B.26}$$

où sg désigne la fonction signe, et $\alpha_k = V_k + \lambda_k(|X| - \sum_l V_l)$. Ainsi, les composantes sont soit inchangées par la procédure, soit déphasées d'un angle π après une itération, selon le signe du terme α_k.

Dans le cas où $|X| \geq \sum_l V_l$, il est clair que $\alpha_k \geq 0$. Ainsi, les composantes sont inchangées par application de notre procédure. Considérons à présent le cas contraire [1], c'est-à-dire tel que $|X| \leq \sum_l V_l$. La condition pour qu'une composante k ne soit pas modifiée par la procédure est donc :

$$V_k + \lambda_k(|X| - \sum_l V_l) \geq 0, \tag{B.27}$$

1. C'est notamment le cas dans le scénario Oracle : $X = \sum_k X_k = \sum_k V_k e^{i\phi_k}$ donc par inégalité triangulaire, on a $|X| = |\sum_k X_k| \leq \sum_k |X_k| = \sum_k V_k$.

ce qui est équivalent à :

$$\lambda_k \leq \frac{V_k}{\sum_l V_l - |X|}. \tag{B.28}$$

On considère en effet le cas où le dénominateur ne s'annule pas : lorsque c'est le cas, il est aisé de vérifier que les composantes sont inchangées par la procédure (l'erreur est nulle).

On définit l'ensemble $\mathcal{P} \subset [\![1, K]\!]$ des indices pour lesquels la condition (B.28) est respectée, et $\mathcal{Q} = [\![1, K]\!] \backslash \mathcal{P}$ l'ensemble des autres indices (qui ne vérifient pas la condition). Si $\mathcal{Q}$ est vide, tous les indices respectent la condition (B.28) auquel cas notre conclusion est inchangée (le schéma d'initialisation choisi est un point fixe de l'algorithme). Par ailleurs, $\mathcal{P}$ ne peut pas être vide, car cela signifierait qu'aucun des poids λ_k ne respecte (B.28). On aurait alors $\forall k$:

$$\lambda_k > \frac{V_k}{\sum_l V_l - |X|}, \tag{B.29}$$

ce qui en sommant conduirait à :

$$\sum_k \lambda_k > \frac{\sum_k V_k}{\sum_l V_l - |X|} > 1, \tag{B.30}$$

ce qui est contradictoire avec le fait que les poids doivent sommer à 1. On considère donc le cas où ni $\mathcal{P}$ ni $\mathcal{Q}$ ne sont vides. On aurait donc, en fin de première itération, pour tout $p \in \mathcal{P}$ et $q \in \mathcal{Q}$:

$$\hat{X}_p^{(1)} = V_p \frac{X}{|X|} \text{ et } \hat{X}_q^{(1)} = -V_q \frac{X}{|X|}. \tag{B.31}$$

A l'itération suivante, on a alors, pour $p \in \mathcal{P}$:

$$Y_p^{(2)} = \frac{X}{|X|} \left(V_p + \lambda_p (|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q) \right). \tag{B.32}$$

A partir de là, distinguons deux cas :

— Si $|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q \geq 0$, alors $V_p + \lambda_p (|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q) \geq 0$;

— Sinon, alors en utilisant la relation (B.28) sur λ_p, on a, $\forall p \in \mathcal{P}$:

$$V_p + \lambda_p (|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q) \geq V_p \left(1 + \frac{|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q}{\sum_l V_l - |X|} \right). \tag{B.33}$$

Or, il est aisé de constater que :

$$\begin{aligned} \sum_{q \in \mathcal{Q}} V_q &\geq -\sum_{q \in \mathcal{Q}} V_q \\ -|X| + \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q &\geq -|X| + \sum_{p' \in \mathcal{P}} V_{p'} - \sum_{q \in \mathcal{Q}} V_q \\ -|X| + \sum_l V_l &\geq -|X| + \sum_{p' \in \mathcal{P}} V_{p'} - \sum_{q \in \mathcal{Q}} V_q \\ \frac{|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q}{\sum_l V_l - |X|} &\geq -1 \end{aligned}$$

Ainsi, on a tout de même $V_p + \lambda_p (|X| - \sum_{p' \in \mathcal{P}} V_{p'} + \sum_{q \in \mathcal{Q}} V_q) \geq 0$.

Donc, comme le terme qui multiplie $\frac{X}{|X|}$ dans (B.32) est positif, la mise à jour après normalisation donne bien :

$$\hat{X}_p^{(2)} = V_p \frac{X}{|X|}, \tag{B.34}$$

elle n'est donc toujours pas modifiée.

On procède alors à un calcul similaire pour un indice $q \in \mathcal{Q}$. On a :

$$Y_q^{(2)} = \frac{X}{|X|} \left(-V_q + \lambda_p (|X| - \sum_{p \in \mathcal{P}} V_p + \sum_{q' \in \mathcal{Q}} V_{q'}) \right). \tag{B.35}$$

Encore une fois, on peut distinguer les deux cas de figure. Si $|X| - \sum_{p \in \mathcal{P}} V_p + \sum_{q' \in \mathcal{Q}} V_{q'}$ est négatif, alors tout le terme devant $\frac{X}{|X|}$ dans l'équation (B.35) est négatif. Dans le cas contraire, on a, en utilisant la non validité de la condition (B.28) sur λ_q :

$$-V_q + \lambda_q (|X| - \sum_{p \in \mathcal{P}} V_p + \sum_{q' \in \mathcal{Q}} V_{q'}) \leq V_q \left(-1 + \frac{|X| - \sum_{p \in \mathcal{P}} V_p + \sum_{q' \in \mathcal{Q}} V_{q'}}{\sum_l V_l - |X|} \right) \tag{B.36}$$

Un calcul similaire au précédent montre alors que ce terme est négatif. Ainsi, dans les deux cas, le terme devant $\frac{X}{|X|}$ dans l'équation (B.35) est négatif et la mise à jour après normalisation donne donc :

$$\hat{X}_q^{(2)} = -V_q \frac{X}{|X|}. \tag{B.37}$$

En fin de compte, on observe que l'algorithme se stabilise après une itération, puisque, $\forall k \in \mathcal{P} \bigcup \mathcal{Q} = [\![1, K]\!]$:

$$\hat{X}_k^{(2)} = \hat{X}_k^{(1)}. \tag{B.38}$$

En conclusion, lorsqu'on initialise l'algorithme avec cette technique (donner la phase du mélange à chaque source), le résultat dépend du choix qui est fait pour les poids λ_k :

— Soit tous les poids vérifient (B.28), auquel cas les valeurs initiales ne sont pas modifiées par l'algorithme ;

— Soit ce n'est pas le cas. Alors, les composantes ne respectant pas cette condition verront leur phase décalée de π par rapport à leur valeur initiale (i.e. $\angle X$). La procédure est ensuite stabilisée après cette première itération.

Annexe C

Estimation du modèle de CNMF à phase contrainte

Nous fournissons ici le détail mathématique de l'obtention des règles de mise à jour pour les paramètres du modèle de NMF complexe à phase contrainte décrite dans le chapitre 7. Deux méthodes d'estimation sont possibles : la méthode de relaxation, présentée dans la section C.1, qui conduit à l'algorithme 8, et la méthode de la fonction auxiliaire, décrite dans la section C.2, qui mène à l'algorithme 9.

C.1 Méthode de relaxation

La méthode de relaxation consiste à minimiser successivement la fonction de coût par rapport à chaque variable.

C.1.1 Estimation de H

On commence par calculer la dérivée partielle de chacun des termes composant la fonction de coût par rapport à $H(k,t)$.

Terme NMF Réécrivons tout d'abord D en isolant les termes dépendant de cette variable (on raisonne donc ici à k et t fixés) :

$$\begin{aligned} D(X,\hat{X}) &= \sum_{f,t'} |X(f,t') - \sum_{l=1}^{K} W(f,l)H(l,t')e^{i\phi_l(f,t')}|^2 \\ &= \sum_{f} |X(f,t) - \sum_{l=1}^{K} W(f,l)H(l,t)e^{i\phi_l(f,t)}|^2 + \sum_{f,t'\neq t} |X(f,t') - \sum_{l=1}^{K} W(f,l)H(l,t')e^{i\phi_l(f,t')}|^2 \\ &\overset{c}{=} \sum_{f} |X(f,t) - \sum_{l\neq k} W(f,l)H(l,t)e^{i\phi_l(f,t)} - W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2. \end{aligned}$$

On note :

$$B_k(f,t) = X(f,t) - \sum_{l\neq k} W(f,l)H(l,t)e^{i\phi_l(f,t)}. \tag{C.1}$$

Le terme D s'écrit alors, en notant $\Re$ la partie réelle :

$$\begin{aligned} D(X,\hat{X}) &\stackrel{c}{=} \sum_f |B_k(f,t) - W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2 \\ &\stackrel{c}{=} \sum_f W(f,k)^2H(k,t)^2 - 2W(f,k)H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)}) + B_k(f,t)^2 \\ &\stackrel{c}{=} \sum_f W(f,k)^2H(k,t)^2 - 2W(f,k)H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)}). \end{aligned}$$

Nous pouvons à présent aisément calculer la dérivée de D par rapport à $H(k,t)$:

$$\begin{aligned} \frac{\partial D}{\partial H(k,t)} &= \sum_f 2W(f,k)^2H(k,t) - 2W(f,k)\Re(B_k(f,t)e^{-i\phi_k(f,t)}) \\ &= 2\left[H(k,t)\sum_f W(f,k)^2 - \sum_f W(f,k)\Re(B_k(f,t)e^{-i\phi_k(f,t)})\right]. \end{aligned} \tag{C.2}$$

Terme de parcimonie Le seul autre terme dépendant de H est le terme de parcimonie, qui se réécrit en isolant l'unique terme en $H(k,t)$:

$$\begin{aligned} \mathcal{C}_s(H) &= 2\sum_{l,t'} H(l,t')^p \\ &= 2H(k,t)^p + 2\sum_{(l,t')\neq(k,t)} H(l,t')^p \\ &\stackrel{c}{=} 2H(k,t)^p. \end{aligned}$$

Sa dérivée partielle est donnée par :

$$\frac{\partial \mathcal{C}_s}{\partial H(k,t)} = 2pH(k,t)^{p-1}. \tag{C.3}$$

Dérivée complète La dérivée partielle de $\mathcal{C}$ par rapport à $H(k,t)$ est alors obtenue en combinant (C.2) et (C.3) :

$$\begin{aligned} \frac{\partial \mathcal{C}}{\partial H(k,t)} =& 2\left[H(k,t)\sum_f W(f,k)^2 - \sum_f W(f,k)\Re(B_k(f,t)e^{-i\phi_k(f,t)})\right] \\ &+ 2p\sigma_s H(k,t)^{p-1}. \end{aligned}$$

On peut alors simplifier cette expression :

$$\begin{aligned} \frac{1}{2}\frac{\partial \mathcal{C}}{\partial H(k,t)} =& H(k,t)\left[\sigma_s H(k,t)^{p-2} + \sum_f W(f,k)^2\right] \\ &- \sum_f W(f,k)\Re(B_k(f,t)e^{-i\phi_k(f,t)}). \end{aligned}$$

L'annulation de cette dérivée partielle conduit alors à la règle de mise à jour suivante pour $H(k,t)$:

$$H(k,t) = \frac{\sum_f W(f,k)\Re(B_k(f,t)e^{-i\phi_k(f,t)})}{\sigma_s H(k,t)^{p-2} + \sum_f W(f,k)^2}. \tag{C.4}$$

On pose :

$$\beta_k = \Re(B_k \odot \overline{\Phi}_k). \tag{C.5}$$

et on note α la matrice de dimensions $F \times T$ dont tous les éléments valent 1. On peut à présent écrire de façon compacte les mises à jour de H sous forme vectorielle à partir de (C.4) :

$$H_k = \frac{(W_k)^T \beta_k}{p\sigma_s (H_k)^{\odot p-2} + ((W_k)^{\odot 2})^T \alpha}. \tag{C.6}$$

Cette mise à jour est ensuite projetée sur l'orthant positif afin de prendre en compte le fait que les H doivent être positifs.

C.1.2 Estimation de W

Pour obtenir la règle de mise à jour sur W, nous allons procéder de la même façon que pour obtenir celle sur H. Le calcul est similaire, le seul terme dépendant de W étant D, qui peut se réécrire en isolant les termes dépendant de $W(f,k)$:

$$\begin{aligned} D(X,\hat{X}) &\stackrel{c}{=} \sum_t |B_k(f,t) - W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2 \\ &\stackrel{c}{=} \sum_t W(f,k)^2 H(k,t)^2 - 2W(f,k)H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)}). \end{aligned}$$

La dérivée partielle de D par rapport à $W(f,k)$ est alors :

$$\frac{\partial D}{\partial W(f,k)} = 2\left[W(f,k)\sum_t H(k,t)^2 - \sum_f H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)})\right]. \tag{C.7}$$

soit :

$$\frac{\partial \mathcal{C}}{\partial W(f,k)} = 2\left[W(f,k)\sum_t H(k,t)^2 - \sum_f H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)})\right]. \tag{C.8}$$

L'annulation de cette dérivée partielle conduit alors à la règle de mise à jour suivante pour $W(f,k)$:

$$W(f,k) = \frac{\sum_t H(k,t)\Re(B_k(f,t)e^{-i\phi_k(f,t)})}{\sum_t H(k,t)^2}. \tag{C.9}$$

En utilisant les notations vectorielles, on a :

$$W_k = \frac{\beta_k (H_k)^T}{\alpha((H_k)^{\odot 2})^T}. \tag{C.10}$$

On effectue également une projection sur l'orthant positif car les W doivent être positifs.

C.1.3 Estimation de ψ

Nous allons à présent estimer le paramètre de phase de référence ψ. Le seul terme de la fonction de coût qui dépend de ce paramètre est $\mathcal{C}_r$. La démarche est alors similaire à celle conduite dans le chapitre 6 quand nous avions considéré une contrainte relaxée. Nous devons tout d'abord réécrire ce terme $\mathcal{C}_r$ en isolant les dépendances en $\psi_k(f)$, et faire apparaître un terme cosinusoïdal, aisé à maximiser. Le calcul est le même, à la différence que les amplitudes, supposées connues dans le chapitre 6, ne le sont plus. On a alors :

$$\psi_k(f) = \angle \left(\sum_{t \in \Omega_k} |X(f,t)|^2 e^{i\phi_k(f,t)} e^{-i\lambda_k(t)f} \right). \tag{C.11}$$

En reprenant la règle de mise à jour de ψ (C.11) et en utilisant les notations matricielles introduites dans le chapitre 7 (section 7.2), on obtient :

$$\Psi_k = \frac{\texttt{diag}_\texttt{m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)}{|\texttt{diag}_\texttt{m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)|}. \tag{C.12}$$

C.1.4 Estimation de λ

Minimiser $\mathcal{C}$ par rapport à $\lambda_k(t)$ revient à minimiser uniquement $\mathcal{C}_r$ puisque seul ce terme dépend de cette variable. Comme pour l'estimation de $\psi_k(f)$, nous nous inspirons donc ce que qui a été fait dans le chapitre 6. L'idée était d'adapter l'algorithme ESPRIT Hua et al. (2004) pour estimer ce paramètre. En posant, $\forall (k,f)$ et $\forall t \in \Omega_k$:

$$\begin{aligned} \gamma_k(f,m) &= |X(f,t)| e^{i\phi_k(f,t)} e^{-i\psi_k(f)}, \\ \underline{\gamma}(t,k) &= [\gamma_k(0,t), ..., \gamma_k(F-1,t)]^T, \end{aligned}$$

On obtient alors l'estimation de $\lambda_k(t)$ pour les trames d'attaques $t \in \Omega_k$:

$$\lambda_k(t) = \angle \left(\underline{\gamma}(t,k)^H_\downarrow \underline{\gamma}(t,k)_\uparrow \right). \tag{C.13}$$

On réécrit alors cette mise à jour :

$$\lambda_k = \angle \left[(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H (|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow}) \right], \tag{C.14}$$

et en utilisant la notation Λ, cela s'écrit :

$$\Lambda_k = \texttt{vand} \left(\frac{(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H (|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})}{|(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H (|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})|} \right) \texttt{diag}_\texttt{v}(\mathbb{1}_k). \tag{C.15}$$

C.1.5 Estimation de ϕ

Pour pouvoir calculer la dérivée de la fonction de coût $\mathcal{C}$ par rapport à $\phi_k(f,t)$, il faudrait distinguer les trames d'attaque $t \in \Omega_k$ des autres. Afin de simplifier la formulation et d'éviter de devoir systématiquement faire cette distinction, nous introduisons la fonction indicatrice de l'ensemble Ω_k, que nous notons de la façon suivante :

$$\mathbb{1}_k(t) = \begin{cases} 1 & \text{si } t \in \Omega_k \\ 0 & \text{sinon.} \end{cases} \tag{C.16}$$

Nous notons de façon similaire $\overline{\mathbb{1}}_k = 1 - \mathbb{1}_k$ qui est la fonction indicatrice du complémentaire de Ω_k. L'estimation du paramètre $\phi_k(f,t)$ repose sur le calcul des dérivées partielles des différents

termes de la fonction de coût. Ce calcul a déjà été conduit pour les termes D et $\mathcal{C}_r$ dans le chapitre 6. Ainsi :

$$\begin{aligned}\frac{\partial D}{\partial \phi_k(f,t)} &= -i\overline{B}_k(f,t)W(f,k)H(k,t)e^{i\phi_k(f,t)} + iB_k(f,t)W(f,k)H(k,t)e^{-i\phi_k(f,t)} \\ &= -iW(f,k)H(k,t)(\overline{B}_k(f,t)e^{i\phi_k(f,t)} - iB_k(f,t)e^{-i\phi_k(f,t)}),\end{aligned}$$

et :

$$\frac{\partial \mathcal{C}_r}{\partial \phi_k(f,t)} = -i\mathbb{1}_k(t)|X(f,t)|^2(e^{i\phi_k(f,t)}e^{-i\lambda_k(t)f}e^{-i\psi_k(f)} - e^{-i\phi_k(f,t)}e^{i\lambda_k(t)f}e^{i\psi_k(f)}).$$

On en déduit, de façon analogue, la dérivée partielle du terme $\mathcal{C}_u$ (notons qu'il faut considérer deux contributions du terme $\phi_k(f,t)$ puisque celui-ci est présent dans 2 termes de la somme, faisant ainsi intervenir les valeurs en $t-1$ et en $t+1$) :

$$\begin{aligned}\frac{\partial \mathcal{C}_u}{\partial \phi_k(f,t)} &= -i\overline{\mathbb{1}}_k(t)|X(f,t)|^2(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t-1)}e^{-2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}) \\ &- i\overline{\mathbb{1}}_k(t+1)|X(f,t+1)|^2(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t+1)}e^{2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t+1)}e^{-2i\pi S\nu_k(f)}).\end{aligned} \tag{C.17}$$

On peut donc obtenir la dérivée partielle de la fonction de coût total par rapport à $\phi_k(f,t)$:

$$\begin{aligned}i\frac{\partial \mathcal{C}}{\partial \phi_k(f,t)} &= |X(f,t)|^2\sigma_r\mathbb{1}_k(t)(e^{i\phi_k(f,t)}e^{-i\lambda_k(t)f}e^{-i\psi_k(f)} - e^{-i\phi_k(f,t)}e^{i\lambda_k(t)f}e^{i\psi_k(f)}) \\ &+ |X(f,t)|^2\sigma_u\overline{\mathbb{1}}_k(t)(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t-1)}e^{-2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}) \\ &+ |X(f,t+1)|^2\sigma_u\overline{\mathbb{1}}_k(t+1)(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t+1)}e^{2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t+1)}e^{-2i\pi S\nu_k(f)}) \\ &+ W(f,k)H(k,t)(\overline{B}_k(f,t)e^{i\phi_k(f,t)} - iB_k(f,t)e^{-i\phi_k(f,t)}).\end{aligned}$$

Ainsi, on aboutit à l'estimation de $\phi_k(f,t)$:

$$\begin{aligned}\phi_k(f,t) = \angle[&B_k(f,t)W(f,k)H(k,t) + |X(f,t)|^2\sigma_r\mathbb{1}_k(t)e^{i\lambda_k(t)f}e^{i\psi_k(f)} \\ &+ |X(f,t)|^2\sigma_u\overline{\mathbb{1}}_k(t)e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)} + |X(f,t+1)|^2\sigma_u\overline{\mathbb{1}}_k(t+1)e^{i\phi_k(f,t+1)}e^{-2i\pi S\nu_k(f)}].\end{aligned} \tag{C.18}$$

Remarque : Comme on l'a fait remarquer dans le chapitre 6, il faudrait en toute rigueur considérer deux solutions (celle-ci et la même modulo π). En examinant la dérivée seconde, on trouve alors que c'est bien la solution donnée par (C.18) qui minimise la fonction de coût.

Nous définissons la variable $\rho_k \in \mathbb{C}^{F\times T}$:

$$\rho_k = \sigma_r(\Psi_k\mathbb{1}_k)\odot\Lambda_k\odot|X|^{\odot 2} + \sigma_u(\mu_k\overline{\mathbb{1}}_k)\odot\Phi_{k,\rightarrow}\odot|X|^{\odot 2} + \sigma_u(\overline{\mu}_k\overline{\mathbb{1}}_{k,\leftarrow})\odot\Phi_{k,\leftarrow}\odot|X_{\leftarrow}|^{\odot 2}. \tag{C.19}$$

Ainsi, la mise à jour sur ϕ (C.18) se réécrit :

$$\phi_k = \angle(B_k \odot (W_kH_k) + \rho_k). \tag{C.20}$$

On utilisera plutôt la notation Φ et donc la mise à jour :

$$\Phi_k = \frac{B_k \odot (W_kH_k) + \rho_k}{|B_k \odot (W_kH_k) + \rho_k|}. \tag{C.21}$$

C.1.6 Estimation de ν

Les derniers paramètres à estimer sont les fréquences instantanées $\nu_k(f)$. Cette estimation est faite en utilisant la méthode présentée dans le chapitre 4 : sur chaque colonne de W (c'est-à-dire sur chaque spectre constituant le dictionnaire d'atomes), on effectue une localisation de pics, puis on procède à une QIFFT, et on découpe l'ensemble des fréquences en régions d'influence. Il est à noter que comme cette estimation est faite à chaque itération, les paramètres $\nu_k(f)$ ne sont pas fixes. Ainsi, cela peut impacter la décroissance de la fonction de coût.

Remarque : Nous avons écrit les mises à jour pour chacune des composantes k, et donc nous n'avons pas tout écrit sous forme matricielle. Ceci est dû au fait que la méthode utilisée (méthode de relaxation) implique de mettre à jour chaque composante θ_k en fonction des valeurs actualisées des précédentes $\theta_1, ..., \theta_{k-1}$. Aussi, il est nécessaire de faire cette mise à jour séquentiellement.

C.2 Méthode de la fonction auxiliaire

Nous fournissons ici la dérivation mathématique de l'obtention des règles de mise à jour pour les paramètres du modèle de NMF complexe à phase contrainte décrite dans le chapitre 7, par la méthode de la fonction auxiliaire. Cela conduit à l'Algorithme 9.

C.2.1 Principe général

Rappelons l'expression de la fonction de coût :

$$\mathcal{C}(\theta) = D(X, \hat{X}) + \sigma_u \mathcal{C}_u(\phi), + \sigma_r \mathcal{C}_r(\phi, \psi, \lambda) + \sigma_s \mathcal{C}_s(H), \tag{C.22}$$

où les différents termes sont définis dans le chapitre 7. L'idée est de majorer $\mathcal{C}(\theta)$ par une fonction auxiliaire $g(\theta, \tilde{\theta})$ dépendant de nouveaux paramètres $\tilde{\theta}$, et qui vérifie :

$$\mathcal{C}(\theta) = \min_{\tilde{\theta}} g(\theta, \tilde{\theta}). \tag{C.23}$$

Comme on l'a rappelé dans la section 2.2.4 du chapitre 2, la fonction de coût $\mathcal{C}$ est alors décroissante sous les règles de mise à jour suivantes :

$$\tilde{\theta} \leftarrow \arg\min_{\tilde{\theta}} g(\theta, \tilde{\theta}) \text{ et } \theta \leftarrow \arg\min_{\theta} g(\theta, \tilde{\theta}). \tag{C.24}$$

C.2.2 Obtention de la fonction auxiliaire

L'obtention de la fonction auxiliaire se fait de façon similaire à Kameoka et al. (2009), à laquelle sont simplement ajoutés les termes de contraintes de phase. Ainsi, en considérant les paramètres auxiliaire $B_k \in \mathbb{C}^{F \times T}$ tels que $\sum_k B_k = X$ et en introduisant les termes $G_k \in [0, 1]^{F \times T}$ tels que $\forall (f, t)$, $\sum_k G_k(f, t) = 1$, on peut montrer en utilisant l'inégalité de Jensen sur la fonction convexe $|.|^2$ que :

$$D(X, \hat{X}) \leq \sum_{f,t,k} \frac{|B_k(f,t) - W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2}{G_k(f,t)}. \tag{C.25}$$

Le cas d'égalité est obtenu en recherchant le minimum du terme de droite par rapport à B_k. Pour ce faire, on introduit la contrainte $\sum_k B_k = X$ par la méthode des multiplicateurs de

Lagrange, et la recherche du point selle du Lagrangien fournit la solution suivante (comme on l'a déjà similairement prouvé dans l'Annexe B) :

$$B_k = \hat{X}_k + G_k(X - \hat{X}), \tag{C.26}$$

où $\hat{X}_k(f,t) = W(f,k)H(k,t)e^{i\phi_k(f,t)}$ et $\hat{X} = \sum_k \hat{X}_k$. Par ailleurs, en introduisant les variables auxiliaire $\bar{H} \in \mathbb{R}_+^{K\times T}$, on montre également que :

$$\mathcal{C}_s(H) \leq \sum_{t,k} p\bar{H}(k,t)^{p-2}H(k,t)^2 + (2-p)\bar{H}(k,t)^p. \tag{C.27}$$

Cette inégalité est en effet montrée pour chaque terme de la somme en utilisant le fait que la dérivée du terme de droite par rapport à $\bar{H}(k,t)$ est nulle lorsque $\bar{H}(k,t) = H(k,t)$. La valeur du terme de droite vaut alors $2H(k,t)^p$. Il est enfin aisé de montrer que c'est un minimum, en examinant par exemple la dérivée seconde en ce point, qui s'avère être positive.

En fin de compte, on a $\mathcal{C}(\theta) \leq g(\theta, \tilde{\theta})$ avec :

$$\begin{aligned} g(\theta,\tilde{\theta}) = \sum_{f,t,k} \frac{|B_k(f,t) - W(f,k)H(k,t)e^{i\phi_k(f,t)}|^2}{G_k(f,t)} + \sigma_s \sum_{t,k} p\bar{H}(k,t)^{p-2}H(k,t)^2 + (2-p)\bar{H}(k,t)^p \\ + \sigma_u \sum_{f,k}\sum_{t\notin\Omega^k} |X(f,t)|^2 |e^{i\phi_k(f,t)} - e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}|^2 + \sigma_u \sum_{f,k}\sum_{t\in\Omega^k} |X(f,t)|^2 |e^{i\phi_k(f,t)} - e^{i\psi_k(f)}e^{i\lambda_k(t)f}|^2. \end{aligned} \tag{C.28}$$

La fonction g est bien une fonction auxiliaire de $\mathcal{C}$ puisque sa valeur minimale en fonction des paramètres auxiliaires $\tilde{\theta} = \{B_k, \bar{H}\}$ est exactement $\mathcal{C}(\theta)$.

C.2.3 Mise à jour des paramètres auxiliaires

L'obtention des règles de mise à jour des paramètres auxiliaires se fait en minimisant g par rapport à ces paramètres. On a déjà calculé ces valeurs minimales pour prouver que g était une fonction auxiliaire, on les synthétise donc ici :

$$B_k = \hat{X}_k + G_k(X - \sum_k \hat{X}_k), \tag{C.29}$$

$$\bar{H} = H. \tag{C.30}$$

C.2.4 Mise à jour des paramètres principaux

L'estimation des paramètres principaux du modèle se fait de façon similaire à la méthode de relaxation (cf. section C.1), c'est-à-dire que l'on va successivement minimiser g par rapport à chacune des variables θ.

Estimation de H Pour estimer H, on calcule la dérivée de g par rapport à cette variable et on annule celle-ci. Le calcul est similaire à celui présenté dans le cas de la méthode de relaxation, et conduit à :

$$\begin{aligned} \frac{\partial g}{\partial H(k,t)} =& 2\left[H(k,t)\sum_f \frac{W(f,k)^2}{G_k(f,t)} - \sum_f \frac{W(f,k)}{G_k(f,t)}\Re(B_k(f,t)e^{-i\phi_k(f,t)})\right] \\ &+ 2p\sigma_s H(k,t)\bar{H}(k,t)^{p-2}, \end{aligned}$$

et son annulation conduit à :

$$H(k,t) = \frac{\sum_f \frac{W(f,k)}{G_k(f,t)} \Re(B_k(f,t)e^{-i\phi_k(f,t)})}{\sigma_s \bar{H}(k,t)^{p-2} + \sum_f \frac{W(f,k)^2}{G_k(f,t)}}. \tag{C.31}$$

On pose $\beta_k = \Re(B_k \odot \overline{\Phi}_k)$, et on obtient la mise à jour de H sous forme vectorielle :

$$H_k = \frac{(W_k)^T \frac{\beta_k}{G_k}}{p\sigma_s(\bar{H}_k)^{\odot p-2} + ((W_k)^{\odot 2})^T \frac{1}{G_k}}. \tag{C.32}$$

Estimation de W Comme précédemment, on obtient la mise à jour de W en annulant la dérivée partielle de g par rapport à cette variable :

$$\frac{\partial g}{\partial W(f,k)} = 2\left[W(f,k)\sum_t \frac{H(k,t)^2}{G_k(f,t)} - \sum_f \frac{H(k,t)}{G_k(f,t)} \Re(B_k(f,t)e^{-i\phi_k(f,t)})\right], \tag{C.33}$$

ce qui conduit à :

$$W(f,k) = \frac{\sum_t \frac{H(k,t)}{G_k(f,t)} \Re(B_k(f,t)e^{-i\phi_k(f,t)})}{\sum_t \frac{H(k,t)^2}{G_k(f,t)}}. \tag{C.34}$$

Avec les notations vectorielles, on peut donc écrire :

$$W_k = \frac{\frac{\beta_k}{G_k}(H_k)^T}{\frac{1}{G_k}((H_k)^{\odot 2})^T}. \tag{C.35}$$

Paramètres de phases d'attaque Les paramètres du modèle de phase dans les trames d'attaques ψ et λ n'apparaissent que dans le terme $\mathcal{C}_r$ de la fonction de coût, et donc de la fonction auxiliaire. Aussi, la minimisation de g par rapport à ces paramètres est identique à la minimisation de $\mathcal{C}$ par méthode de relaxation. On a donc les mêmes mises à jour :

$$\Psi_k = \frac{\mathtt{diag_m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)}{|\mathtt{diag_m}((\Phi_k \odot X^{\odot 2})(\Lambda_k)^H)|}, \tag{C.36}$$

$$\Lambda_k = \mathtt{vand}\left(\frac{(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H(|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})}{|(\overline{\Psi}_{k,\downarrow} \odot \Psi_{k,\uparrow})^H(|X|_\downarrow \odot |X|_\uparrow \odot \overline{\Phi}_{k,\downarrow} \odot \Phi_{k,\uparrow})|}\right) \mathtt{diag_v}(\mathbb{1}_k). \tag{C.37}$$

Estimation de la phase ϕ La phase est estimée de façon similaire au cas par relaxation. Néanmoins, il faut prendre ici en compte la présence des poids G_k :

$$\begin{aligned}
i\frac{\partial \mathcal{C}}{\partial \phi_k(f,t)} &= |X(f,t)|^2\sigma_r \mathbb{1}_k(t)(e^{i\phi_k(f,t)}e^{-i\lambda_k(t)f}e^{-i\psi_k(f)} - e^{-i\phi_k(f,t)}e^{i\lambda_k(t)f}e^{i\psi_k(f)}) \\
&+ |X(f,t)|^2\sigma_u \overline{\mathbb{1}}_k(t)(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t-1)}e^{-2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)}) \\
&+ |X(f,t+1)|^2\sigma_u \overline{\mathbb{1}}_k(t+1)(e^{i\phi_k(f,t)}e^{-i\phi_k(f,t+1)}e^{2i\pi S\nu_k(f)} - e^{-i\phi_k(f,t)}e^{i\phi_k(f,t+1)}e^{-2i\pi S\nu_k(f)}) \\
&+ \frac{W(f,k)H(k,t)}{G_k(f,t)}(\overline{B}_k(f,t)e^{i\phi_k(f,t)} - iB_k(f,t)e^{-i\phi_k(f,t)}).
\end{aligned}$$

Ainsi, on aboutit à l'estimation de $\phi_k(f,t)$:

$$\phi_k(f,t) = \angle[B_k(f,t)\frac{W(f,k)H(k,t)}{G_k(f,t)} + |X(f,t)|^2\sigma_r \mathbb{1}_k(t)e^{i\lambda_k(t)f}e^{i\psi_k(f)} + |X(f,t)|^2\sigma_u\overline{\mathbb{1}}_k(t)e^{i\phi_k(f,t-1)}e^{2i\pi S\nu_k(f)} + |X(f,t+1)|^2\sigma_u\overline{\mathbb{1}}_k(t+1)e^{i\phi_k(f,t+1)}e^{-2i\pi S\nu_k(f)}]. \tag{C.38}$$

En notant :

$$\rho_k = \sigma_r(\Psi_k\mathbb{1}_k)\odot\Lambda_k\odot|X|^{\odot 2}+\sigma_u(\mu_k\overline{\mathbb{1}}_k)\odot\Phi_{k,\rightarrow}\odot|X|^{\odot 2}+\sigma_u(\overline{\mu}_k\overline{\mathbb{1}}_{k,\leftarrow})\odot\Phi_{k,\leftarrow}\odot|X_{\leftarrow}|^{\odot 2}, \tag{C.39}$$

la mise à jour se réécrit :

$$\phi_k = \angle(B_k \odot \frac{W_kH_k}{G_k} + \rho_k), \tag{C.40}$$

soit, avec la notation Φ :

$$\Phi_k = \frac{B_k \odot \frac{W_kH_k}{G_k} + \rho_k}{|B_k \odot \frac{W_kH_k}{G_k} + \rho_k|}. \tag{C.41}$$

Remerciements

Car ce travail n'aurait jamais pu être mené à bien sans l'aide, le support et l'existence de vous tous.

Mes remerciements vont tout d'abord à mes directeurs de thèse, Roland Badeau et Bertrand David. Votre pari était risqué, mais vos conseils, votre disponibilité, et surtout votre patience ont rendu cela possible. Merci pour votre confiance et pour la qualité de votre encadrement.

Merci aux membres du jury, les rapporteurs Philippe Depalle et Laurent Girin, et les examinateurs Yannis Stylianou et Jonathan Le Roux, pour vous être intéressés à mon travail, pour avoir joué le jeu de la soutenance, et pour avoir, par vos remarques et conseils, amélioré la qualité de ce manuscrit même après le grand oral.

Je remercie également tout le microcosme de Télécom, qui est l'univers dans lequel j'ai pris plaisir à évoluer ces trois dernières années.

Un grand merci aux permanents de l'équipe AAO : Gaël, Slim, Alexandre, Yves, et le petit dernier Umut. Chacun à votre manière, que ce soit par un conseil technique, une blague du vendredi, ou autour d'un verre, vous avez contribué au bon déroulement de ce travail.

Merci à Florence, Marie-Laure et Laurence pour les coups de main logistiques et administratifs. Vous avez toujours été aimables, disponibles et d'une aide précieuse.

Je tiens naturellement à remercier tous les autres, doctorants, post-docs ou stagiaires avec qui j'ai (un peu) souffert et (beaucoup) ri ces dernières années.

Merci aux anciens qui m'ont accueilli avant de s'envoler vers de nouvelles aventures : Xabier, François, Nicolas, Mounira et Aymeric. Pour la façon dont vous m'avez permis d'amorcer cette thèse.

Merci à mes contemporains : Simon D., Floriane, Clément, Simon L., Victor, Arthur, Umut, Dogac, Jaïr et tous les autres. Sans vous tous, cette thèse ne serait qu'une accumulation d'équations sans âme.

Un merci spécial à Xabier, pour ton soutien face aux drones quand j'étais seul dans mon bureau, ton éternelle présence au Diamant, et parce que tu fusionnes plus que tu ne sépares. À Clément, pour ton esthétique de la symphonie, tes performances au DC, et les cafés-clopes sans clope. À Glavio, pour ta patience tout sauf variationnelle, pour les poules à la peau dorée, et pour avoir été un super co-bureau.

Mes remerciements vont bien sûr à tous mes amis, qui ont fait de la vie au-delà des murs de Télécom ce qu'elle a été, et continuera d'être.

À la bande de viets & assimilés : Qassem, Philippe, Louis, Dimitri, Bastien, Jackal, Yann, Alban et Alexis. Depuis le temps que vous m'entendez proférer des idioties, il est temps de vous dire merci.

À Philippe et Elsa, pour votre humour, votre personnalité et votre générosité naturelle.

À Jackal et Anne-Lise, pour votre amitié précieuse et sans concession.

À Pierre, pour nos répètes endiablées, pour tes théories ampoulées, et pour tout le reste.

À François, pour l'enfer.

À Hugo, pour les cours de guitare, les ondes sensuelles et les comparaisons absurdes.

Au Djo et Clem, les pois chiches délurés et délirants. Aux petites qui squattent et à la petite qui stats'. Pour être, depuis tant d'années, une lumière dans un couloir obscur.

Et surtout à Qassem, qui a supporté l'obésité de mes inepties. Pour m'avoir permis de tenir ces trois années. Pour être un aussi bon coloc que pote.

Mes remerciements vont, enfin mais en premier lieu, à ma famille.

À mes parents, sans qui je ne serais pas ici (ni ailleurs, à vrai dire). Pour votre amour depuis les balbutiements et votre bienveillance tout au long du voyage.

À mes quatre fantastiques frères, Victor, Charles, Louis et Simon, pour être chacun une source d'inspiration et de bonheur.

Merci de m'avoir montré le chemin, et de l'avoir fait avec moi.

À ceux que j'oublie, et à ceux que je n'oublie pas.

RECONSTRUCTION DE PHASE PAR MODÈLES DE SIGNAUX : APPLICATION À LA SÉPARATION DE SOURCES AUDIO

Paul MAGRON

RESUME : De nombreux traitements appliqués aux signaux audio travaillent sur une représentation Temps-Fréquence (TF) des données. Lorsque le résultat de ces algorithmes est un champ spectral d'amplitude, la question se pose, pour reconstituer un signal temporel, d'estimer le champ de phase correspondant. C'est par exemple le cas dans les applications de séparation de sources, qui estiment les spectrogrammes des sources individuelles à partir du mélange ; la méthode dite de filtrage de Wiener, largement utilisée en pratique, fournit des résultats satisfaisants mais est mise en défaut lorsque les sources se recouvrent dans le plan TF.

Cette thèse aborde le problème de la reconstruction de phase de signaux dans le domaine TF appliquée à la séparation de sources audio. Une étude préliminaire révèle la nécessité de mettre au point de nouvelles techniques de reconstruction de phase pour améliorer la qualité de la séparation de sources. Nous proposons de baser celles-ci sur des modèles de signaux. Notre approche consiste à exploiter des informations issues de modèles sous-jacents aux données comme les mélanges de sinusoïdes. La prise en compte de ces informations permet de préserver certaines propriétés intéressantes, comme la continuité temporelle ou la précision des attaques. Nous intégrons ces contraintes dans des modèles de mélanges pour la séparation de sources, où la phase du mélange est exploitée. Les amplitudes des sources pourront être supposées connues, ou bien estimées conjointement dans un modèle inspiré de la factorisation en matrices non-négatives complexe. Enfin, un modèle probabiliste de sources à phase non-uniforme est mis au point. Il permet d'exploiter les à priori provenant de la modélisation de signaux et de tenir compte d'une incertitude sur ceux-ci.

Ces méthodes sont testées sur de nombreuses bases de données de signaux de musique réalistes. Leurs performances, en termes de qualité des signaux estimés et de temps de calcul, sont supérieures à celles des méthodes traditionnelles. En particulier, nous observons une diminution des interférences entre sources estimées, et une réduction des artéfacts dans les basses fréquences, ce qui confirme l'intérêt des modèles de signaux pour la reconstruction de phase.

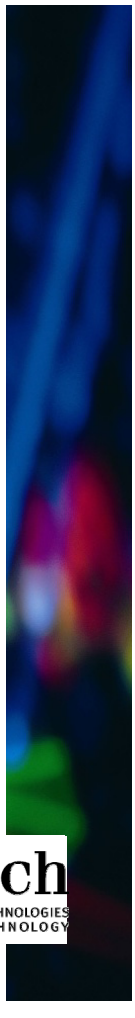
MOTS-CLEFS : reconstruction de phase, modèles de signaux, séparation de sources audio, musique, mélanges de sinusoïdes, factorisation en matrices non-négatives, analyse temps-fréquence, modèles probabilistes

ABSTRACT: A variety of audio signal processing techniques act on a Time-Frequency (TF) representation of the data. When the result of those algorithms is a magnitude spectrum, it is necessary to reconstruct the corresponding phase field in order to resynthesize time-domain signals. For instance, in the source separation framework the spectrograms of the individual sources are estimated from the mixture ; the widely used Wiener filtering technique then provides satisfactory results, but its performance decreases when the sources overlap in the TF domain.

This thesis addresses the problem of phase reconstruction in the TF domain for audio source separation. From a preliminary study we highlight the need for novel phase recovery methods. We therefore introduce new phase reconstruction techniques that are based on music signal modeling : our approach consists in exploiting phase information that originates from signal models such as mixtures of sinusoids. Taking those constraints into account enables us to preserve desirable properties such as temporal continuity or transient precision. We integrate these into several mixture models where the mixture phase is exploited ; the magnitudes of the sources are either assumed to be known, or jointly estimated in a complex nonnegative matrix factorization framework. Finally we design a phase-dependent probabilistic mixture model that accounts for model-based phase priors.

Those methods are tested on a variety of realistic music signals. They compare favorably or outperform traditional source separation techniques in terms of signal reconstruction quality and computational cost. In particular, we observe a decrease in interferences between the estimated sources and a reduction of artifacts in the low-frequency components, which confirms the benefit of signal model-based phase reconstruction methods.

KEY-WORDS: phase recovery, signal modeling, audio source separation, music, mixtures of sinusoids, nonnegative matrix factorization, time-frequency analysis, probabilistic modeling



TELECOM ParisTech

EDITE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY