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

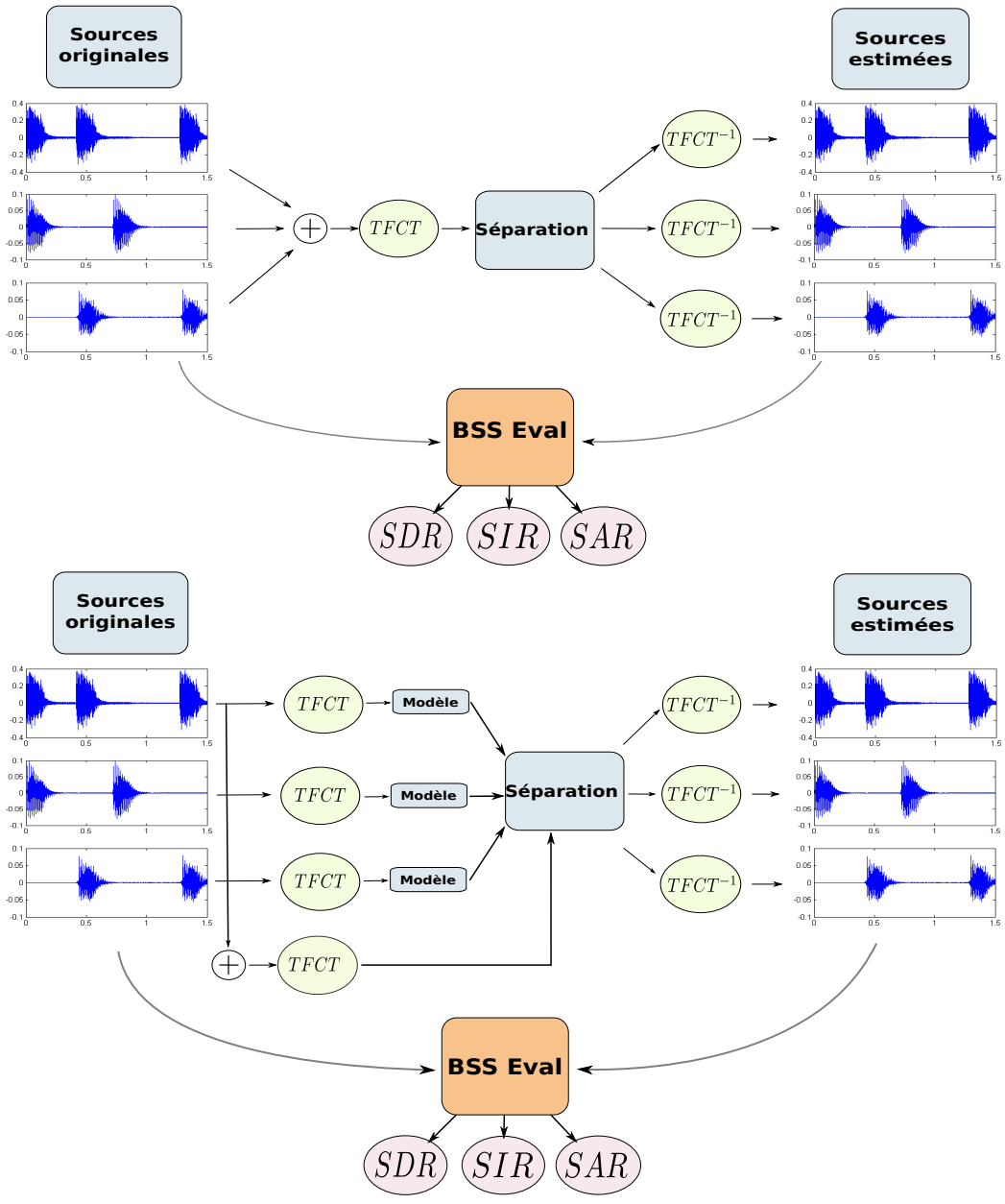

FIGURE 3.1 – Schéma de fonctionnement de notre étude. Deux approches complémentaires sont utilisées : une approche aveugle (en haut) et une approche Oracle (en bas).

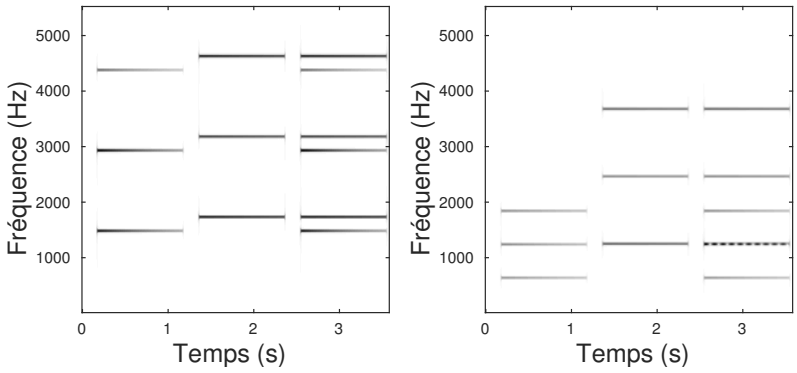

FIGURE 3.2 – Spectrogrammes de mélanges synthétiques : sans recouvrement TF (gauche) et avec recouvrement TF (droite).

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.

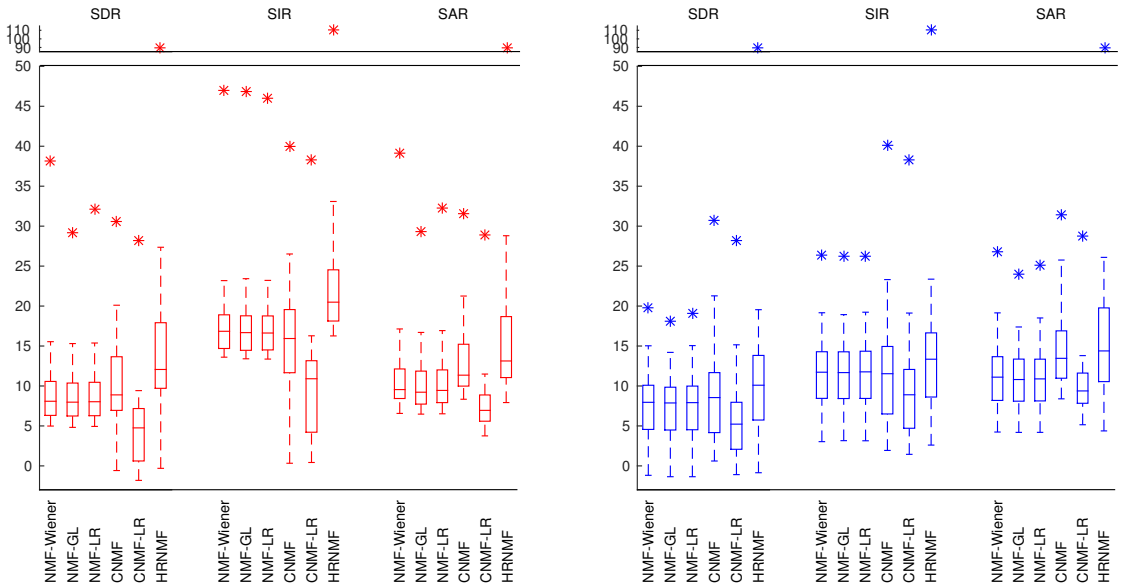

FIGURE 3.3 – Performance de la séparation de mélanges synthétiques sans recouvrement TF (gauche) et avec recouvrement TF (droite). Approches aveugle (boîtes à moustaches) et oracle (étoiles).

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

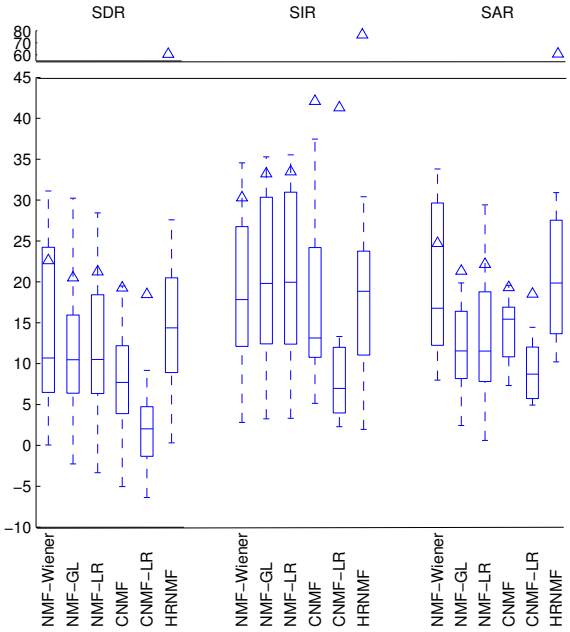

FIGURE 3.4 – Performance de la séparation de notes de pianos. Approches aveugle (boîtes à moustaches) et oracle (triangles).

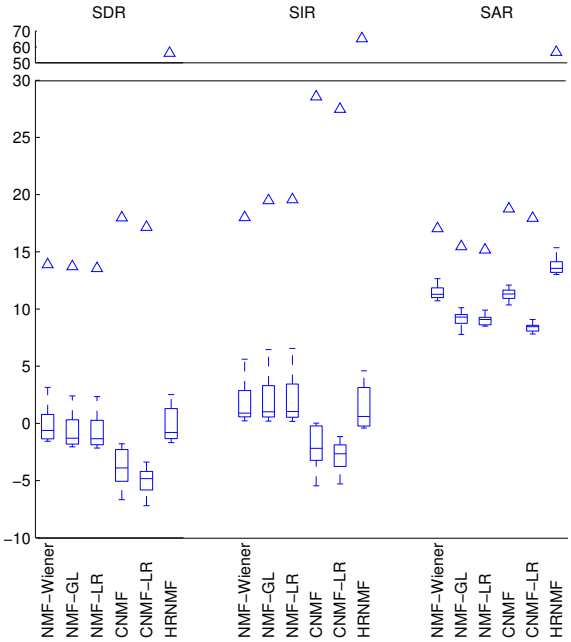

FIGURE 3.5 – Performance de la séparation des sources sur l'extrait MIDI. Approches aveugle (boîtes à moustaches) et oracle (triangles).

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

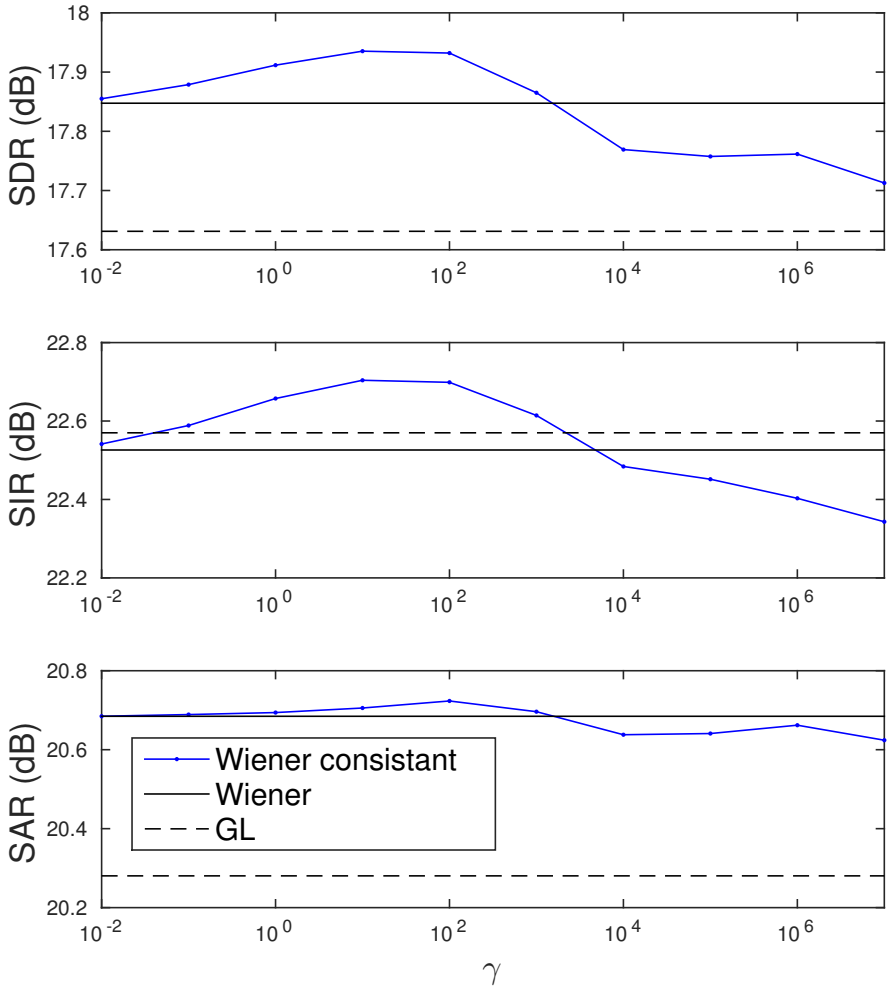

FIGURE 3.6 – Performance du filtrage de Wiener consistant en séparation de sources (mélanges de notes de piano).

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.

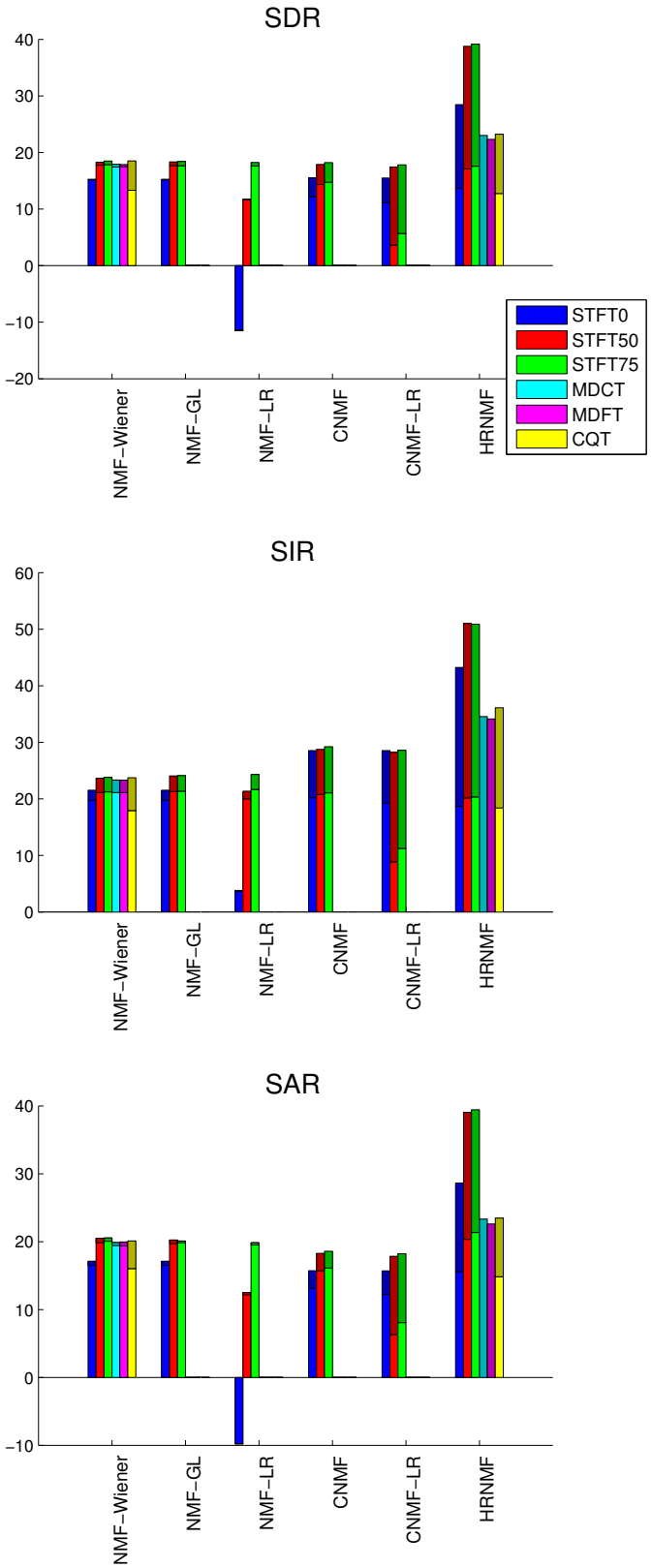

FIGURE 3.7 – Influence de la représentation temps-fréquence sur la performance de la séparation de sources (mélanges de notes de piano en relations harmoniques) : SDR (haut), SIR (milieu) et SAR (bas) en dB. Les barres claires représentent la performance aveugle, les barres foncées représentent la performance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

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

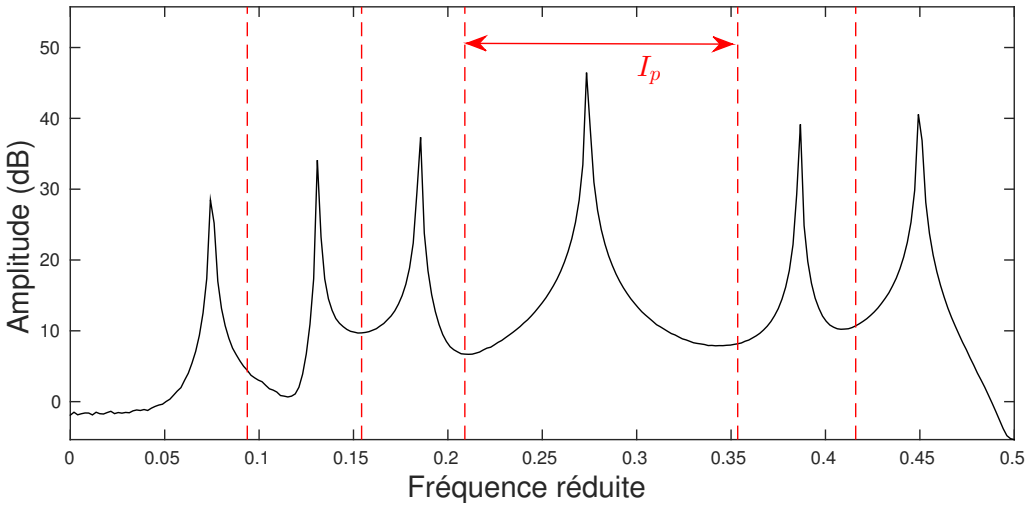

FIGURE 4.1 – Découpage en régions d'influence : un spectre (courbe en traits pleins) est segmenté en régions d'influence, qui sont d'autant plus larges qu'un pic est plus important que ses voisins. Les traits en pointillés représentent les frontières entre ces régions.

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) :

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

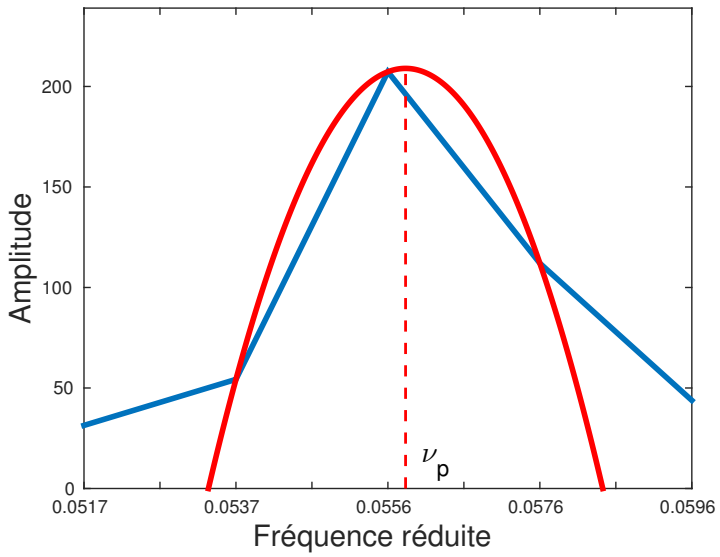

FIGURE 4.2 – Illustration de la QIFFT : un pic d'amplitude est approché par une parabole, et le calcul du maximum de cette parabole conduit à une estimation de la fréquence instantanée.

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.

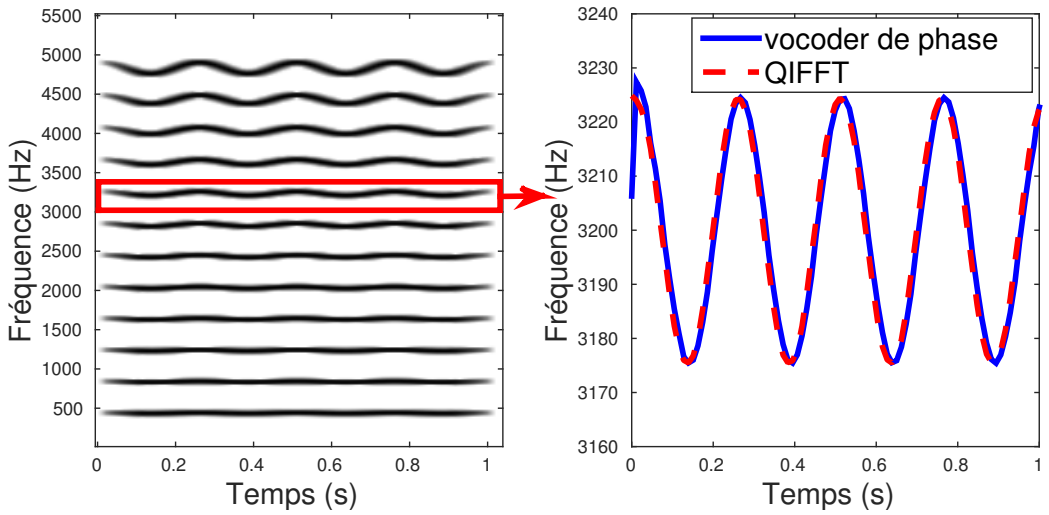

FIGURE 4.3 – Spectrogramme d'un mélange synthétique avec vibrato (gauche) et fréquences instantanées correspondant au partiel oscillant autour de 3200 Hz (droite).

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.

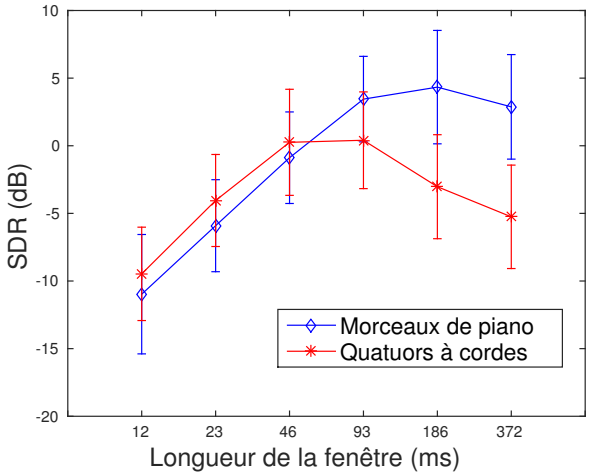

FIGURE 4.4 – Influence de la longueur de fenêtre d'analyse sur la qualité de reconstruction du signal (SDR en dB). Les marqueurs centraux représentent la moyenne et les barres horizontales l'écart-type calculés sur le jeu de données correspondant.

— 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.

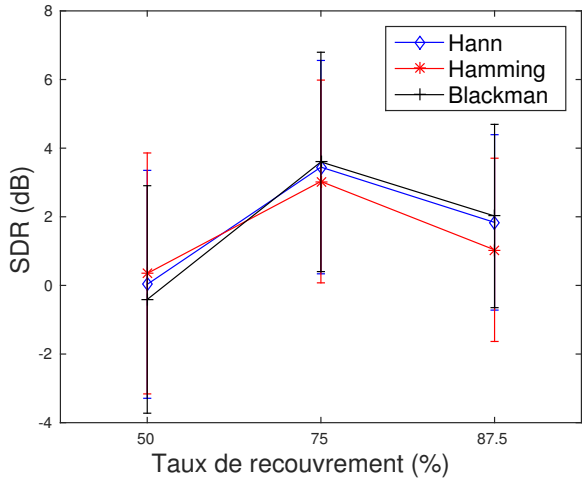

FIGURE 4.5 – Influence du taux de recouvrement de la TFCT sur la qualité de reconstruction du signal (SDR en dB).

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

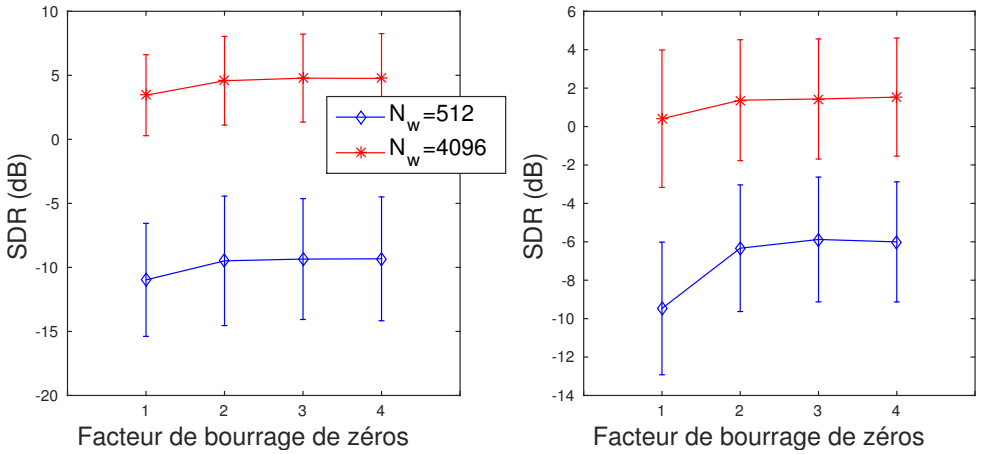

FIGURE 4.6 – Influence du facteur de bourrage de zéros sur la qualité de reconstruction du signal (SDR en dB) pour des morceaux de piano (gauche) et des quatuors à cordes (droite).

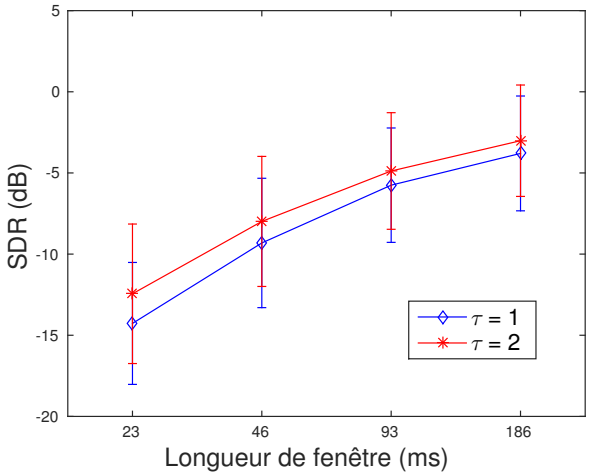

FIGURE 4.7 – Influence de la longueur de fenêtre et du facteur de bourrage de zéros sur la qualité de reconstruction du signal (SDR en dB). Les marqueurs centraux représentent la moyenne et les barres horizontales l'écart-type calculés sur le jeu de données correspondant (jeu de données E).

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.

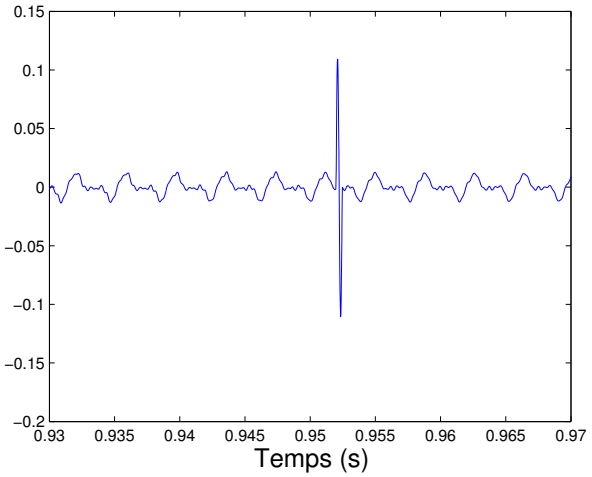

FIGURE 4.8 – Exemple d'un signal de piano en présence d'un craquement.

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") ;

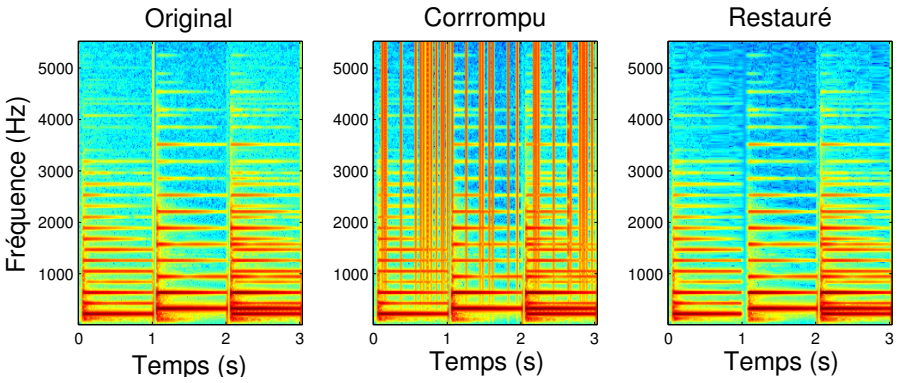

FIGURE 4.9 – Restauration de spectrogrammes par interpolation linéaire de log-amplitude sur un mélange de notes de piano : original (gauche), corrompu par des clics (centre) et restauré (droite).

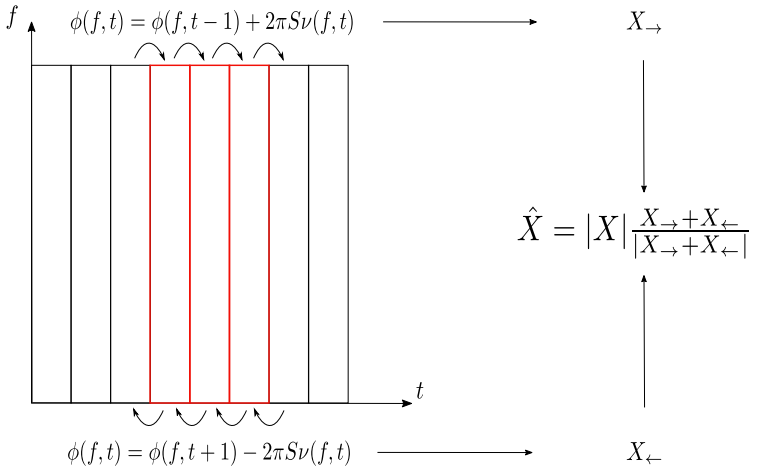

FIGURE 4.10 – Méthode de déroulé de phase pour reconstruire des trames temporelles corrompues : on combine un déroulé avant et un déroulé arrière que l'on moyenne ensuite.

— 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.

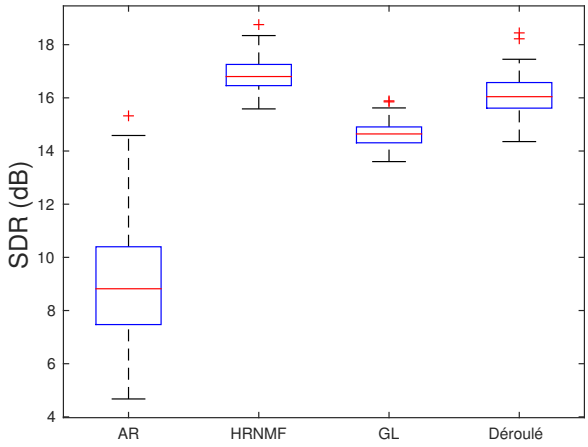

FIGURE 4.11 – Performance de la suppression de clics (SDR en dB) sur le jeu de données E pour plusieurs méthodes.

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.

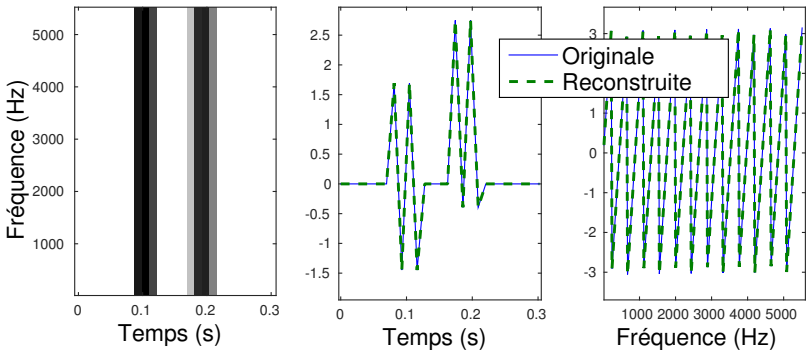

FIGURE 4.12 – Mélange d'impulsions : spectrogramme (gauche) et reconstruction de phase par déroulé linéaire en fonction du temps dans la bande de fréquences à 1800 Hz (centre) et en fonction des fréquences dans la première trame d'attaque (droite).

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.

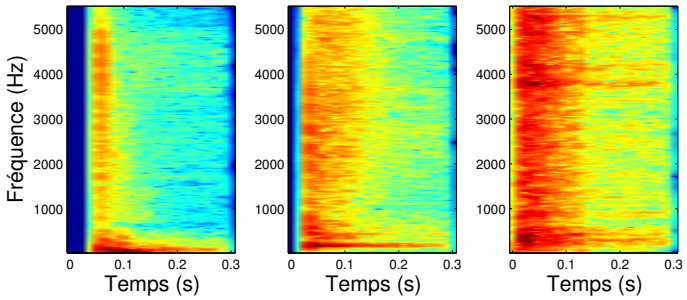

FIGURE 4.13 – Spectrogrammes de signaux percussifs : grosse caisse (gauche), caisse claire (centre) et cymbale Charleston fermée (droite).

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.