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.

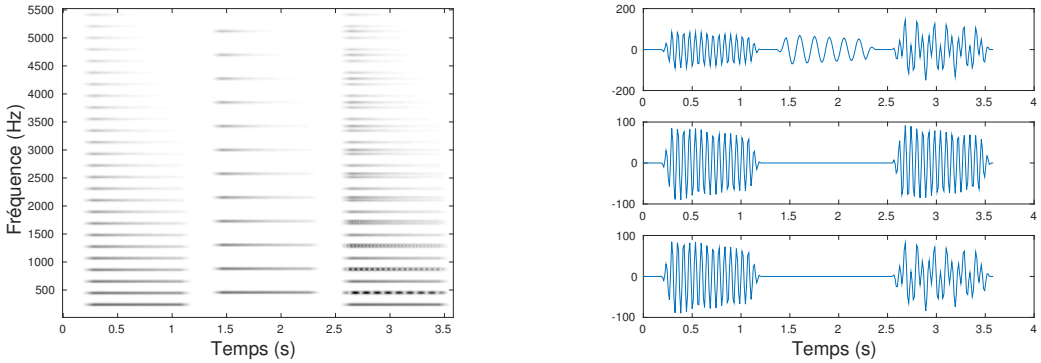

FIGURE 2.1 – Spectrogramme d'un mélange constitué de deux sources synthétiques (à gauche), et parties réelles de diverses composantes dans la bande de fréquences 730 Hz : mélange (en haut à droite), première source originale (au milieu à droite) et première source estimée par filtrage de Wiener (en bas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

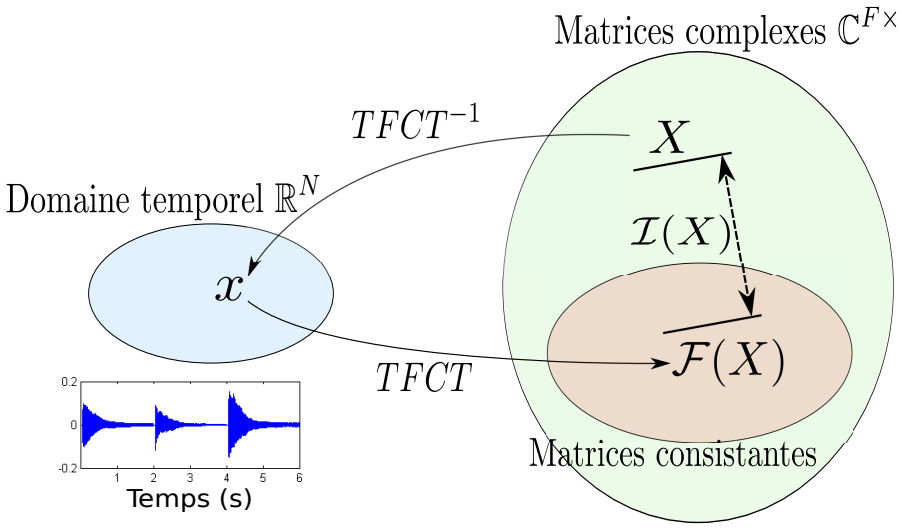

FIGURE 2.2 – Illustration de la notion de consistance : une matrice complexe quelconque n'est pas nécessairement égale à la TFCT de sa TFCT inverse. L'écart entre les deux est appelé inconsistance.

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.

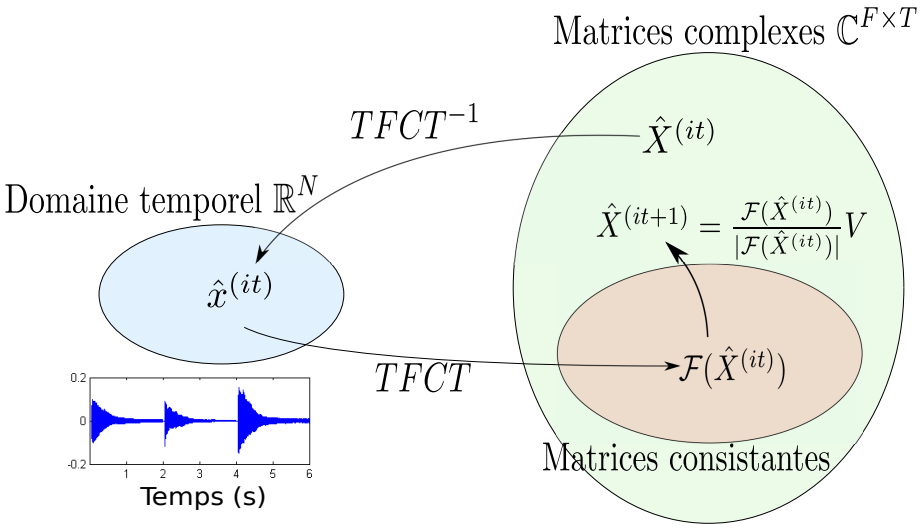

FIGURE 2.3 – Principe de l'algorithme GL : à chaque itération, on applique à la composante complexe estimée l'opérateur $\mathcal{F}$ puis on fixe son amplitude à la valeur objectif V.

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 :

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

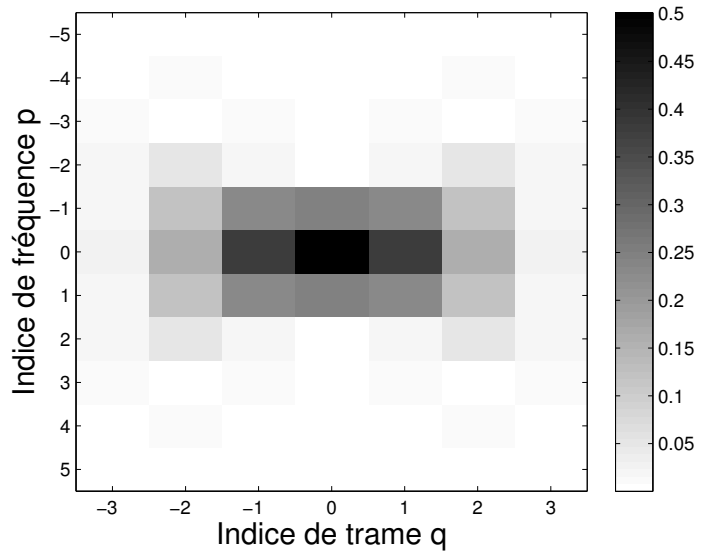

FIGURE 2.4 – Un exemple de module de noyau de consistance $|\alpha|$ obtenu à partir de racines carrées de fenêtres de Hann, de longueur 512 échantillons avec un recouvrement de 75 %.

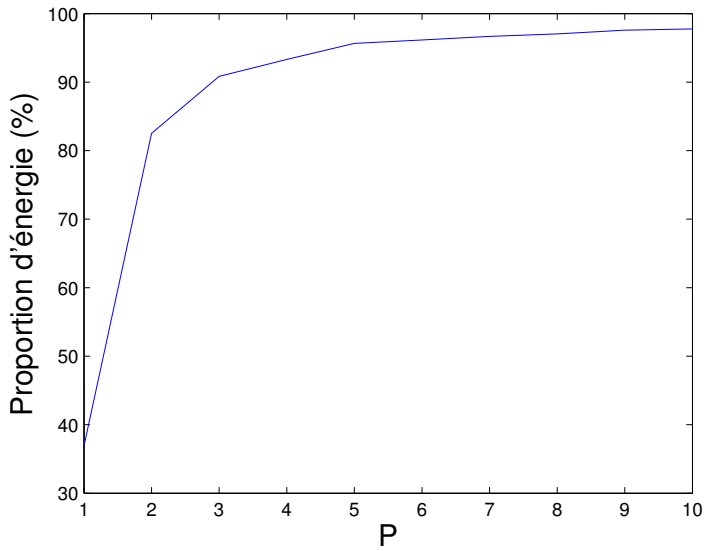

FIGURE 2.5 – Proportion d'énergie du noyau de consistance contenue dans les canaux fréquentiels d'indices $p \in [\![-P, P]\!]$ par rapport à son énergie totale.

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 $K(F+T) \ll FT$.

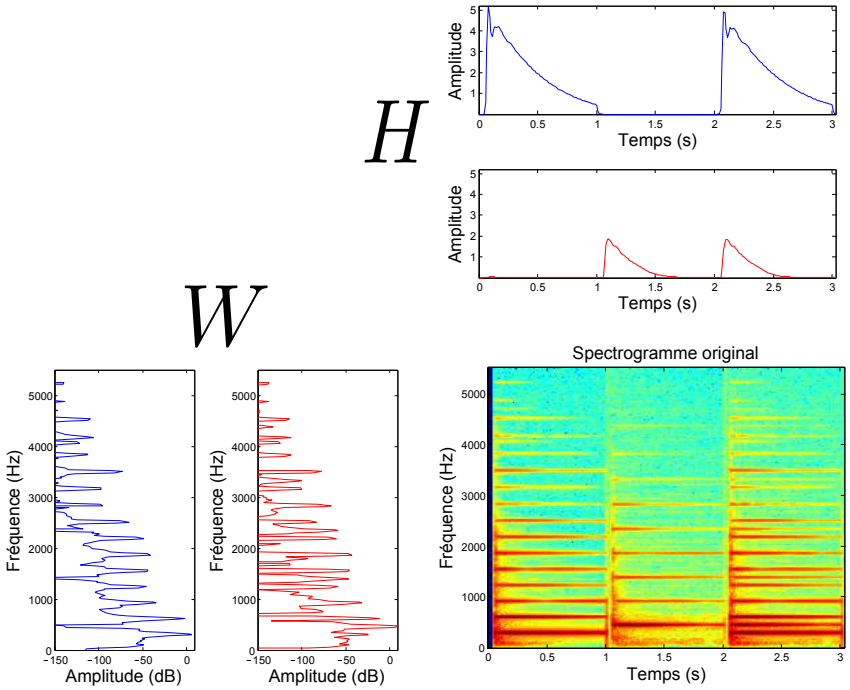

FIGURE 2.6 – Résultat de 100 itérations de règles multiplicatives d'une NMF euclidienne sur un spectrogramme constitué de deux notes de piano (E4 et B4).

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.

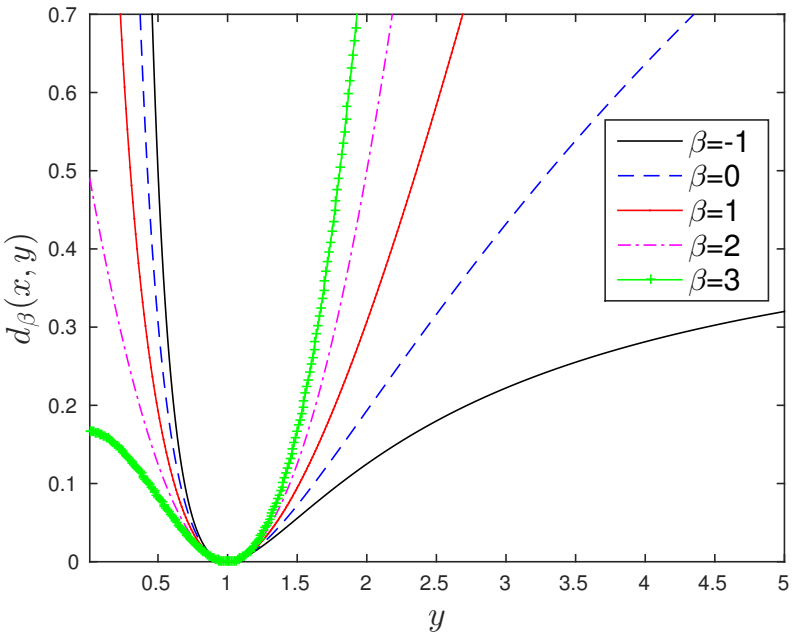

FIGURE 2.7 – Fonction de β-divergence $d_\beta(x, y)$ pour $x = 1$ et plusieurs valeurs de β.

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).

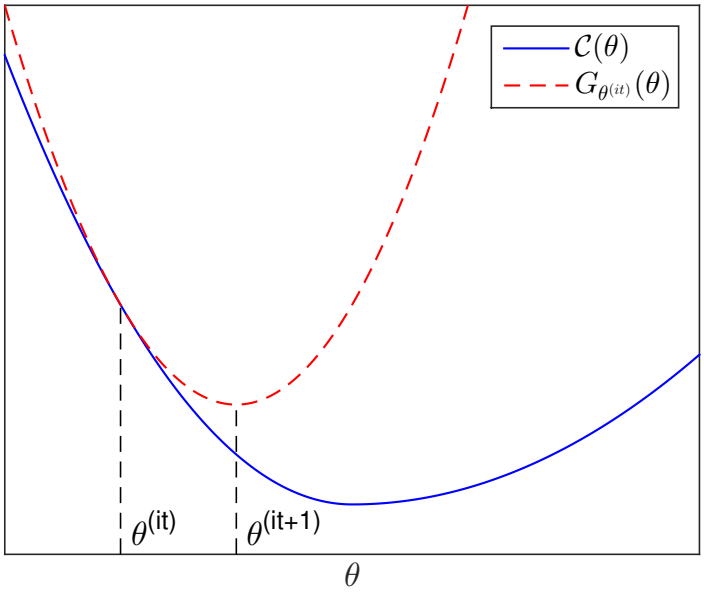

FIGURE 2.8 – Illustration de l'approche Majoration-Minimisation : la fonction de coût $\mathcal{C}$ est majorée à l'itération (it) par la fonction auxiliaire $G_{\theta^{(it)}}$, dont la minimisation conduit à un nouveau paramètre $\theta^{(it+1)}$.

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 ?