Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.

Nancy Bertin

Thèse

présentée pour obtenir le grade de docteur

de l'Ecole Nationale Supérieure des Télécommunications

Spécialité : Signal et Images

Nancy BERTIN

Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.

Soutenue le 2 octobre 2009 devant le jury composé de

Frédéric Bimbot	Rapporteurs
Laurent Daudet	
Bruno Torrésani	Examinateurs
Tuomas Virtanen	
Roland Badeau	Directeurs de thèse
Gaël Richard	

Je dédie ce mémoire
à Thomas Quinot.

Remerciements

Le mémoire et la soutenance de thèse ne marquent pas seulement la fin d'un travail ou l'obtention d'un diplôme. Ils sont l'aboutissement d'un projet mûri, porté, choyé, depuis la naissance d'une vocation jusqu'à la poignée de main finale avec le président du jury ; un projet tant scientifique que personnel, parfois affectif. Jamais solitaire, il est jalonné de moments-clefs et de temps qui s'écoule, de rencontres décisives et de présences réconfortantes. Que soient remerciés ici les acteurs de ces rencontres.

Je remercie en premier lieu et de tout cœur mes directeurs de thèse, **Roland Badeau** et **Gaël Richard**. **Gaël**, superviseur bienveillant, m'a accordé sa confiance dès notre rencontre, en me recommandant en 2003 pour mon stage d'ingénieur. Au cours de cette thèse, il a toujours su garder en tête la vision globale et cohérente de mes travaux, leurs grandes lignes, leurs orientations, et les éclairer pour moi. Je me souviendrai du carnet d'adresses de ce catalyseur de relations sociales, de l'énergie et de l'enthousiasme qu'il déploie pour insérer pleinement ses doctorants dans la communauté des chercheurs. Un œil précieux sur les cordons de la bourse et le calendrier, compréhensif dans les moments difficiles, il m'a offert le temps et les conditions matérielles pour finir sereinement cette thèse. **Roland**, toujours disponible pour moi malgré un emploi du temps parfois acrobatique, a été présent au quotidien pour suivre l'avancement de mes travaux. Je le remercie du temps qu'il m'a accordé et de sa qualité d'écoute. La qualité rédactionnelle de ce mémoire et de mes publications doit beaucoup à l'œil de lynx de ce redoutable relecteur. Modeste et abordable, il m'a toujours reçue d'égal à égale, bien que nous n'ayions pas été du même côté de la soutenance. Je le remercie pour sa disponibilité, son ouverture d'esprit, pour avoir su aussi bien me laisser la bride sur le cou qu'être présent, toujours au moment approprié.

Je remercie l'ensemble des membres du jury de me faire l'honneur de juger mon travail, en cette période pourtant chargée. Je remercie particulièrement **Laurent Daudet** et **Frédéric Bimbot**, rapporteurs, qui ont bien voulu emporter mon manuscrit sur la plage à la place de lectures bien plus divertissantes.

Je n'aurais sans doute jamais acquis les savoirs, les capacités, l'envie et la passion nécessaires pour mener cette thèse à bien sans de cruciales rencontres avec des enseignants formidables, tout au long de ma vie. Je pense à **Claudie Kerrich** et **Christiane N'Guyen**, institutrices, **Pierre Salvaing** et **Éric Merle**, enseignants du secondaire et de classes préparatoires. Je pense également à ceux qui m'ont accompagnée dans l'apprentissage de la musique : **Francine Thivet**, **Agnès Davan**, **Philippe Coupry** et **Richard Bachand**. Dans un passé plus proche, je voudrais remercier particulièrement **Bertrand David** et **Denis Matignon**, enseignants-chercheurs à Télécom ParisTech, pour leur soutien lors de mes candidatures de Master, et pour m'avoir donné une grande leçon de professionnalisme.

Ces rencontres, et les envies qu'elles ont fait naître, n'auraient jamais trouvé leur concrétisation sans la découverte de l'IRCAM, visité en 1993 lors du festival Agora. Parmi tous les chercheurs qui partageaient ce jour-là leur passion, je garde un souvenir particulier de **Stephen MacAdams** et **René Caussé**. J'aimerais aussi remercier l'anonyme qui répondit par téléphone à ma lettre d'adolescente, écrite comme une bouteille à la mer et qui posait la question naïve mais sincère : « comment puis-je arriver à faire le même travail que vous ? » Plus tard, mon année de Master fut largement à la hauteur de ces espérances longuement nourries. Je remercie toute l'équipe pédagogique du Master ATIAM et

mes coreligionnaires de la promotion 2004-2005, qui ont fait de ce Master une année d'épanouissement.

Puis vint la thèse. Outre mes directeurs déjà mentionnés, je tiens à remercier chaleureusement deux personnes qui, plus qu'une simple collaboration, m'ont offert de leur temps, de leur attention, des pistes de réflexion précieuses et des orientations majeures pour mon travail. Que grâces soient rendues à **Emmanuel Vincent** et **Cédric Févotte** pour m'avoir associée à leurs travaux. Ce manuscrit serait tout autre sans leur participation.

Il serait également bien différent sans la participation amicale, les relectures et les remarques de **Laurent Defours**, , **Nicolas Moreau** et **Arnaud Picard**.

Au cours de cette thèse, j'ai également eu la chance de m'adonner à des activités d'enseignement, qui m'ont considérablement enrichie. Je remercie les personnes qui m'ont accueillie dans leurs équipes : **Pierre Audin**, **Guillaume Reuiller**, **Robin Jamet**, **Caroline Bulf**, **Anne Hervé-Minvielle** au Palais de la Découverte ; **Maryse Pelletier**, **Brigitte de la Passardière**, **Ludovic Perret** à l'Université Pierre et Marie Curie. Sans oublier mes visiteurs et étudiants, qui m'ont apporté sans doute beaucoup plus qu'ils ne l'imaginent.

Ces remerciements ne seraient pas complets sans mentionner des rencontres plus personnelles, au sein du laboratoire comme ailleurs. À Télécom ParisTech, j'ai été entourée de compagnons, dont j'ai vu partir certains ; d'autres me succèderont. Ils ont tous contribué tant à la bonne ambiance de travail qu'à d'autres moments plus conviviaux, incluant apéros et éclats de rire. Je citerai dans un ordre aléatoire **Valentin** (grand frère de thèse, du BDT et de la CJC), **Slim** (recel de tickets FIAP), **Cléo**, **Chloé**, **Grégoire**, **Christophe** (rock slovaque et rap mongol), **Thomas T.** (marsupilami), **Loïs** (il est où l'aspirateur ?), **Maria**, **Teodora** (le rat ronronne rarement), **Miguel** (docteur ès-guacamole), **Mathieu G.**, **Nicolas**, **Lionel**, **Pierre L.**, **Jean-Louis**, **Sarah**, **Julien** (PhD comique), **Alexey**, **Mathieu L.**, **Thomas F.**, **Damien** et les membres du BDT. Mais la vie de doctorant ne s'arrête, heureusement, pas aux frontières du laboratoire. Je serai aussi brève que possible ; que les oubliés me pardonnent. Amis d'école et d'ailleurs : **Marie** (rousse), **Louis** (caméraman), **Pablo**, **Pierre-Aimé**, **Alric**, **Ali** & **Éric** (fillots), **Aurélie**, **Alexandra**, **Lionel**, **Grégoire**, **Bastien**, **Alice** & **Jérôme** (sushis périgourdins), **Stéphane**, **Guillaume** (37 points au scrabble), **Pierre P.** & **Isabelle**, **Yacine** & **Zinedine** (carburant). Au rayon cyber-ludique : **Angandilas** (humour nocturne), **Lunean** (naughty girl), **Eithlenn** (rousse quantique), **Zian** (prière anti-péché), **Tirodem** (porteur d'espoir non-euclidien). Les amis des grandes dents : **Carol**, **Immi**, **Rattounette**, **Lizzy**. Discussions musicales et plus si affinités : **Sandra** (et sa hache), **Corinne**, **Karine**, **Delphine**, **Ingrid**, **Kita**.

Au cours de cette thèse, comme cela arrive parfois dans la vie, j'ai dû aussi composer avec ma propre santé ; je dois beaucoup aux équipes médicales compétentes et humaines qui m'ont suivie. Je remercie particulièrement les Dr. **Limal**, **Roullier**, **Bourgeois** et **Papeix**. Recevoir le titre de docteur sera pour moi la meilleure façon de leur témoigner ma gratitude.

Il me reste bien sûr à remercier ma famille. Mes parents, **Martine** et **Yves**, m'ont toujours encouragée et soutenue dans mes choix, faits librement. Ma plus grande fierté est d'être la leur. Je remercie aussi mon frère **Cédric**, pour son soutien pudique ; le respect qu'il me porte est un honneur pour moi. Ma tendresse va à ma tante **Marie-France** et à ma belle-sœur **Hélène** ; elles sont les sœurs que je n'ai pas eues. J'ai aussi une pensée émue pour ma grand-mère **Rolande**, disparue avant de m'avoir vue docteur. Elle m'a transmis deux choses inestimables, qui sont mes armes aujourd'hui : sa force de caractère devant les épreuves, et la foi.

Enfin, le dernier de cette longue mais profondément sincère liste de remerciements est pour **Thomas Quinot**, à qui je dédie ce mémoire. Ami fidèle, soutien précieux, professionnel exemplaire, à l'humanité et la bienveillance inestimables, aux encouragements toujours renouvelés, il est pour moi le modèle de ce que doit être un homme de science, et un homme de bien. Future docteur, j'ai fait mienne sa devise d'ancien doctorant :

« Savoir, savoir faire et faire savoir. »

Table des matières

Table des figures

Liste des tableaux

Acronymes

Pour des raisons de lisibilité, la signification d'un acronyme ou d'une abréviation n'est en général rappelée que lors de sa première utilisation dans le texte d'un chapitre. Par ailleurs, nous employons le terme français ou le terme anglais suivant l'usage le plus répandu.

BSS	*Blind Source Separation* (séparation aveugle de sources)
CRO	*Closeness to rank one* (proximité au rang 1)
DSP	Densité spectrale de puissance
ERB	*Equivalent Rectangular Bandwidth* (bandes de fréquences rectangulaires équivalentes)
EUC	Euclidien(nne), distance euclidienne
EM	*Expectation-Maximization* (Espérance-Maximisation)
fdp	Fonction Densité de Probabilité
FN	Faux négatif
FP	Faux positif
GMM	*Gaussian Mixture Model* (Modèle de mélange gaussien)
HMM	*Hidden Markov Model* (Modèle de Markov Caché)
ICA	*Independant Component Analysis* (analyse en composantes indépendantes)
iid	Indépendantes et identiquement distribuées
IS	Itakura-Saito, divergence d'Itakura-Saito
KKT	Conditions de Karush-Kuhn-Tucker
KL	Kullback-Leibler, divergence de Kullback-Leibler
MAP	Maximum A Posteriori
MMSE	*Minimum Mean Square Error* (moindres carrés)
MV	Maximum de Vraisemblance
NMF	*Non-negative matrix factorization* (factorisation en matrices non-négatives)
PCA	*Principal Component Analysis* (analyse en composantes principales)
SAGE	*Space Alternating Expectation-Maximization*
ssi	Si et seulement si
SSL	Stationnaire au sens large
SVD	*Singular Value Decomposition* (décomposition en valeurs singulières)
SVM	*Support Vector Machine* (machine à vecteur support)
TFCT	Transformée de Fourier à court terme
TP	*True Positive* (vrai positif)

Notations

Les principales conventions de notations employées dans ce document sont résumées ici. Dans la mesure du possible, nous avons tenté de conserver les mêmes lettres et polices de caractères pour désigner les mêmes objets. Les éventuelles collisions de notation sont explicitées lorsque cela est nécessaire, sauf si la quantité désignée peut être clairement déduite du contexte.

$\mathbf{1}$	Matrice dont tous les coefficients sont égaux à 1
$\mathbf{I}$	Matrice identité
$\mathbf{x}$	Vecteur
$\mathbf{X}$	Matrice
$\mathbf{X}_{ij}$, $[\mathbf{X}]_{ij}$, x_{ij}	Coefficient d'indice i, j de la matrice $\mathbf{X}$
$\mathbf{x}_j$	j^e Colonne de la matrice $\mathbf{X}$
x_i	i^e Ligne de la matrice $\mathbf{X}$
$(.)^T$	Transposé
$(.)^H$	Conjugué hermitien
$\otimes$	Produit de Hadamard (terme à terme)
$\oslash$	Division terme à terme
$\mathbf{M}^{[\alpha]}$	Matrice de coefficients $([\mathbf{M}]_{ij})^\alpha$
$det(.)$	Déterminant d'une matrice carrée
$rg(.)$	Rang d'une matrice
ℓ	Numéro de l'itération courante d'un algorithme
n, N	Indice de trame temporelle, nombre de trames temporelles
f, F	Indice de point fréquentiel, nombre de points fréquentiels
k, K	Indice de composante élémentaire, nombre de composantes élémentaires
$\mathbf{V}$	Représentation temps-fréquence à valeurs positives ou nulles

$\stackrel{\text{def}}{=}$	Définition
$\stackrel{\text{c}}{=}$	Égalité à une constante additive près
$\propto$	Proportionnel à
∇f	Gradient de la fonction multivariée f
$\nabla_x f$	Dérivée partielle de la fonction f en sa variable x
$\nabla^2 f$	Matrice hessienne de la fonction f
$\lfloor . \rfloor$	Partie entière
δ	Symbole de Kronecker
$\#\mathcal{C}$	Cardinal d'un ensemble $\mathcal{C}$
$\mathcal{N}(\mu, \sigma)$	Loi normale de moyenne μ et de variance σ
$\mathcal{P}(\lambda)$	Loi de Poisson de paramètre λ
$\mathcal{G}(\alpha, \beta)$	Loi Gamma de paramètre de forme α et de paramètre d'échelle β
$p_X(x)$, $p(x)$	Densité de probabilité de la variable aléatoire X pour une réalisation x
$var(.)$	Variance d'une variable aléatoire
$\mathbb{E}[.]$	Espérance mathématique

Tous les indices désignant des éléments de vecteurs ou de matrices sont numérotés à partir de 1, en accord avec la convention du logiciel Matlab (©MathWorks) utilisé pour les simulations informatiques.

Par égard pour le lecteur non musicien, un glossaire des termes musicaux est proposé en fin de mémoire. Les termes inclus dans ce glossaire sont signalés dans le texte par une étoile * les précédant.

Introduction

Évanescente par nature, la musique peut être fixée sur de nombreux supports de nature très différente : la mémoire humaine, la partition classique, les formats audionumériques tel que le CD, les formats de fichiers symboliques comme le MIDI... La nécessité de conserver la musique sur un support pérenne répond à de nombreux besoins apparus au fil du temps, et comporte de nombreuses difficultés conceptuelles et technologiques. La *transcription*, du latin *trans*, au-delà, et *scribere*, écrire, est l'opération qui consiste à produire un tel support, mais pas n'importe lequel : la forme de représentation de la musique produite par cette opération doit contenir du sens, dire quelque chose sur la musique qu'elle contient. Il faut pour cela disposer d'un format qui permet la représentation de telles informations sémantiques, et de techniques pour convertir la musique vers ce format.

Écrire la musique

S'il existe des arguments, notamment archéologiques, indiquant que nos ancêtres pratiquèrent la musique dès le paléolithique [1], la notation musicale est évidemment postérieure à l'écriture elle-même. Elle lui est pourtant quasi contemporaine : alors que la première écriture, l'écriture cunéiforme des Sumériens, apparaît aux alentours de 3300 av. J.-C., une tablette babylonienne datée du XVI^{e} siècle av. J.-C. atteste de premiers essais de notation musicale [Dablemont, 2008]. La notation est alors très inspirée de l'écriture elle-même : les notes sont représentées par des lettres, surmontées d'accents. Cette notation alphabétique se retrouve chez les Grecs, qui la développent considérablement à partir du V^{e} siècle av. J.-C., puis au Moyen-Âge, où l'écriture de la musique prend véritablement son essor. Sur la figure 1, nous pouvons voir un manuscrit du XI^{e} siècle caractéristique de cette notation largement basée sur l'alphabet, chaque note étant représentée par une seule lettre ; l'idée de contour mélodique ou de notation verticale apparaît, sans être clairement dissociée des paroles elles-mêmes.

Ces notations archaïques semblent relever davantage de l'aide-mémoire que de l'écriture fidèle de la musique : il paraît difficile d'interpréter la musique écrite ici sans l'avoir jamais entendue. [Cullin, 2008] souligne d'ailleurs que les fonctions de la musique écrite vont bien au-delà de cette fonction d'aide-mémoire : elle possède notamment un aspect religieux, rituel voire sacré, comme dépositaire de la parole divine (les neumes, apparus au IX^{e} siècle, sont utilisés dans les monastères) et de support à la méditation. L'aspect esthétique est également indéniable, avec des ornementations et illustrations accompagnant la notation informative. D'abord allusive (contours mélodiques, accents...), l'information

1. Des sifflets creusés dans des phalanges de rennes ont été retrouvés en Charente et datés à 100.000 ans avant notre ère. On a également découvert des flûtes (grotte d'Isturitz au pays basque) et des peintures représentant des hommes jouant de la musique (grotte des Trois Frères en Ariège), datées à 15.000 ans av. J.-C.

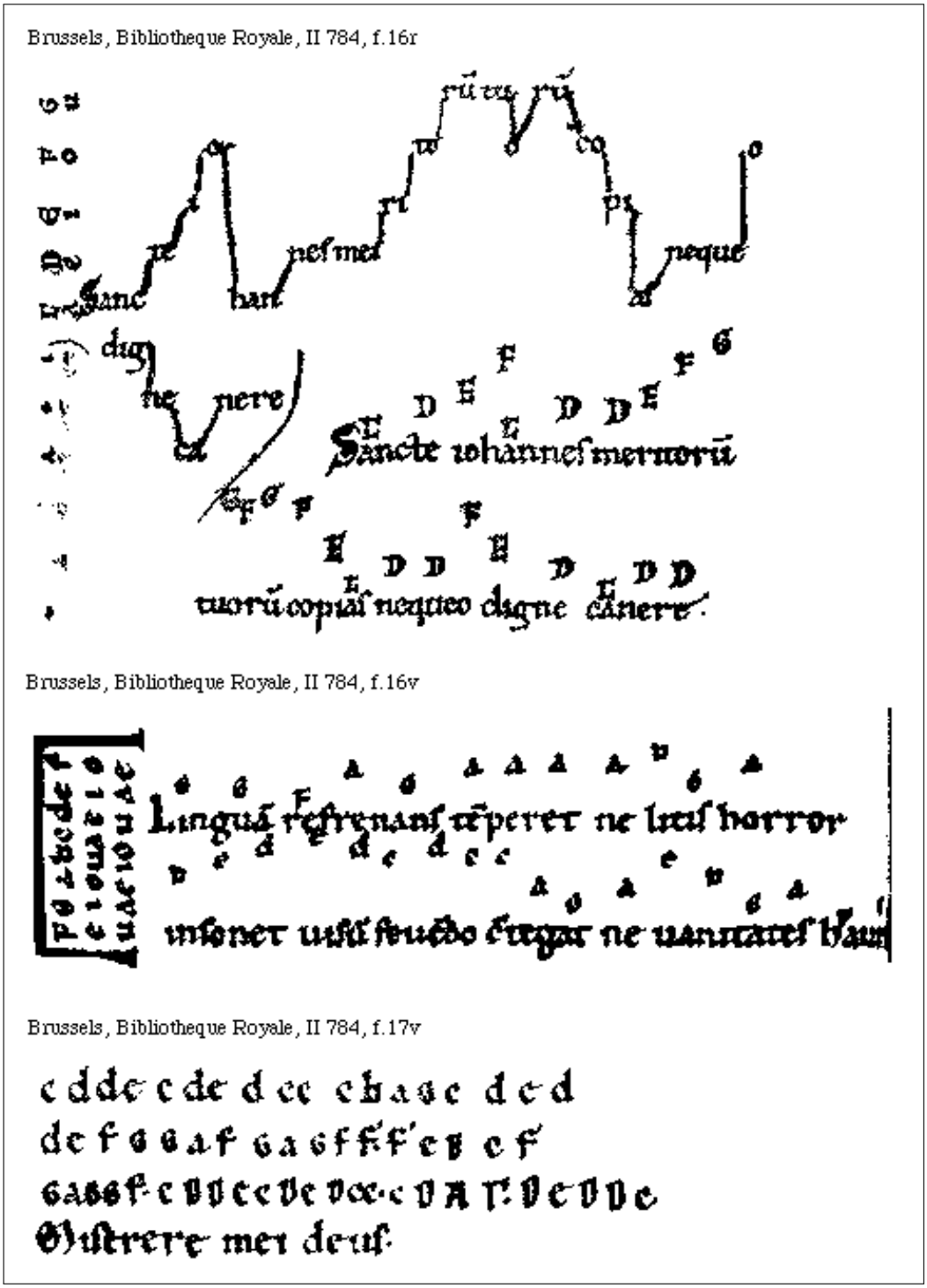


FIGURE 1 – *Micrologus*, manuscrit du XI^e^ siècle attribué à Guido d'Arezzo. Domaine public, Bibliothèque Royale de Bruxelles.

acquiert au fil des siècles de plus en plus de précision : apparition de la *portée [2] et des hauteurs relatives, puis des clefs et de la notation de la hauteur absolue ; normalisation des valeurs rythmiques (rondes, blanches, noires...). On peut voir sur la figure 2 un manuscrit du XV^{e}, l'enrichissement des informations contenues dans la partition à cette époque.

La notation évolue alors progressivement jusqu'au XVII^{e} pour adopter la forme de la partition classique que nous connaissons.

Transcrire la musique

Jusqu'au milieu du XX^{e} siècle, la question de la transcription ne se pose évidemment qu'en termes d'expertise humaine. Puis, avec l'émergence de sciences et surtout de technologies capables de s'attaquer au problème, l'étude automatisée de la musique devient un domaine de recherche très actif, présentant à la fois de grands défis et des enjeux applicatifs. Cela n'est pas étonnant : les scientifiques s'intéressent à l'étude de la musique depuis Pythagore (580-498 av. J.-C.), qui formalise une construction de la gamme (« cycle des quintes ») et observe le rapport entre la longueur d'une corde, la fréquence de sa vibration et la hauteur du son produit. Avec la démocratisation de la pratique musicale et l'apparition de moyens de plus en plus perfectionnés d'enregistrement et de reproduction du son, le défi scientifique et technologique que représente la production automatisée d'une représentation symbolique de la musique devient également un préalable pour remplir de nouveaux besoins. Transcrire automatiquement la musique n'est pas seulement un but en soi, mais aussi un objectif dont le résultat peut déboucher sur de très nombreuses applications, avec des enjeux commerciaux, légaux, artistiques, pédagogiques...

En effet, la manipulation de représentations symboliques est bien plus aisée que celle du son brut, et peut bénéficier de nombreux travaux antérieurs et contemporains sur les séquences de symboles de manière générale, comme par exemple les recherches en traitement du language naturel ou en bio-informatique (étude des séquences ADN). Si l'on dispose d'une représentation symbolique de la musique contenue dans un fichier son, de très nombreuses applications sont envisageables, parmi lesquelles :

- Indexation et navigation dans de grandes bases de données sonores, telles qu'elles sont disponibles aujourd'hui en grande quantité, par exemple sur Internet ou dans nos bibliothèques personnelles ;
- Recherche par similarité, recommandation musicale, recherche de reprises ;
- Protection des droits d'auteur, recherche de plagiat, de copies illégales ;
- Aide au musicologue : recherche de thèmes, d'influences d'un compositeur sur un autre ;
- Aide à l'apprentissage, apprentissage interactif de la musique ;
- Écoute active de la musique.

Avant d'en arriver à ces applications, il faut bien sûr parvenir à produire cette représentation à partir de la seule information de la forme d'onde. C'est l'objet des systèmes de transcription automatique de la musique étudiés ici.

Factoriser le spectrogramme

Pour s'atteler à cette tâche, certaines approches sont basées sur des connaissances a priori (modèles de signaux, utilisation de bases de données d'apprentissage, par exemple [Ryynänen et Klapuri,

2. Dans la suite de ce manuscrit, les termes musicaux sont signalés par une astérisque les précédant, et leurs définitions sont regroupées dans un glossaire à la page 199.

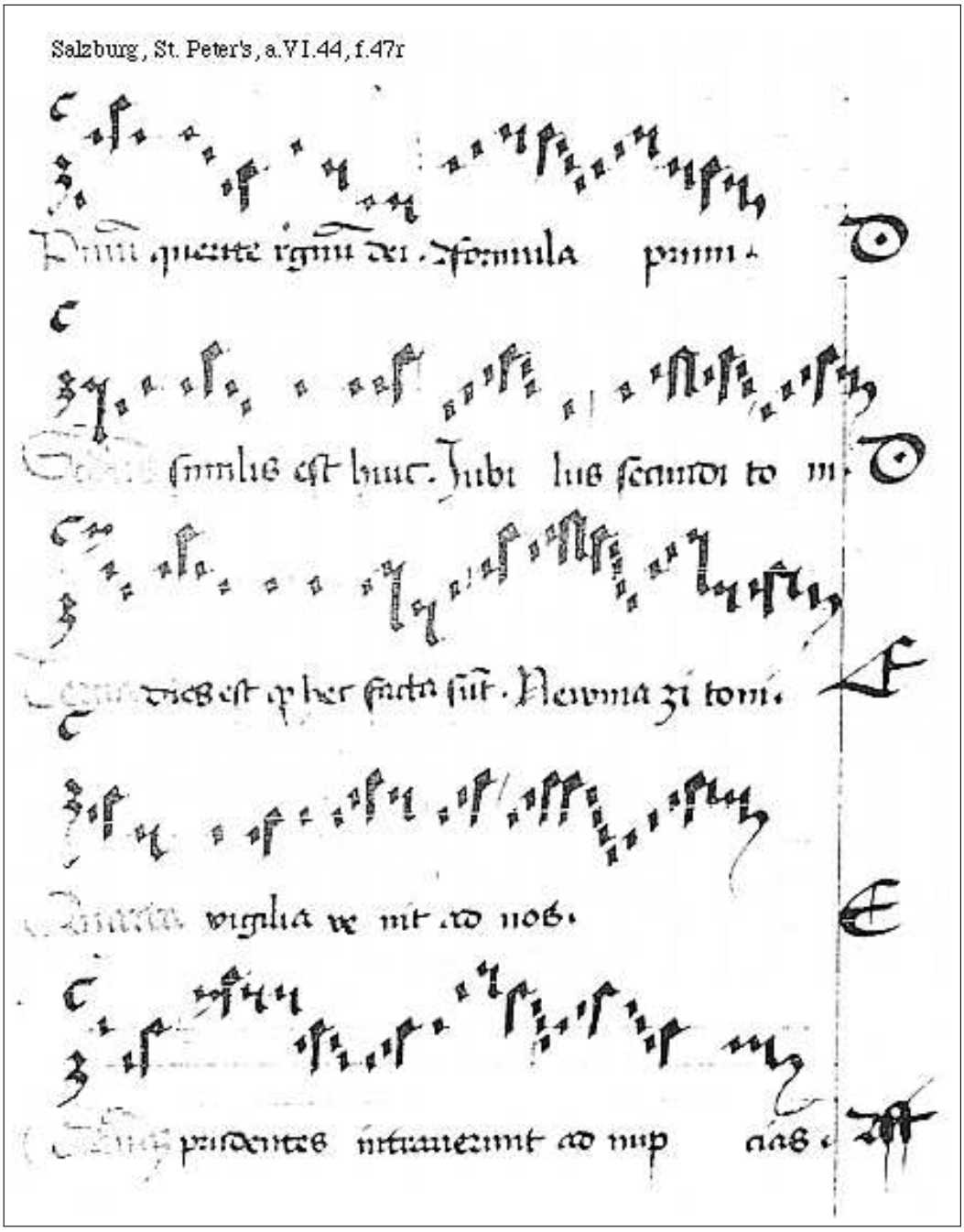


FIGURE 2 – *De modis cantandi*, manuscrit anonyme du XV^e^ siècle. Domaine public, collection de l'Abbaye de Saint-Pierre, Salzburg.

2008, Marolt, 2004]). La contrepartie de ce type de méthodes est leur faible capacité d'adaptation à des signaux s'éloignant trop du modèle ou des données initiales. Pour s'affranchir de cette contrainte, une nouvelle famille d'approches consiste à introduire le moins d'a priori possible sur l'audio de départ, et à tenter de séparer les notes jouées « à l'aveugle ». Parmi celles-ci, les représentations parcimonieuses [Abdallah et Plumbley, 2004, Leveau, 2007, Aharon *et al.*, 2005], les techniques de factorisation en matrices non-négatives [Smaragdis et Brown, 2003], de séparation aveugle (comme l'analyse en composantes indépendantes), et leurs variantes font des hypothèses faibles et réduites sur les signaux. Elles ont montré des résultats prometteurs en transcription de musique polyphonique [Bertin *et al.*, 2007].

Les représentations du signal généralement utilisées dans les tâches d'indexation audio sont issues de techniques d'analyse temps-fréquence ou temps-échelle (transformée de Fourier à court terme, transformée de Wigner-Ville, transformée en ondelettes, etc.). Bien que certaines structures du son soient distinguables sur leurs visualisations (les transitoires percussifs et les partiels d'un son musical sont visibles sur un spectrogramme par exemple), elles ne sont généralement pas clairement extraites par les techniques de séparation de sources, ni par les algorithmes d'analyse de paramètres de haut niveau (tels les outils d'estimation de hauteurs multiples ou de détection de tempo). L'objectif de cette thèse est d'obtenir une représentation orientée-objet du signal, exhibant clairement ces structures, qui serait un intermédiaire entre le signal et une représentation plus haut niveau de type MIDI, par exemple. Une telle représentation aura l'avantage de simplifier les tâches d'indexation et de transcription automatique de musique. Pour extraire ces structures du signal, l'approche que nous visons repose sur des techniques d'analyse matricielle.

Les décompositions de matrices en valeurs propres et en valeurs singulières sont des techniques classiques d'algèbre linéaire utilisées dans un grand nombre d'applications de traitement du signal. Elles permettent de représenter efficacement les données observées en utilisant un nombre limité d'atomes élémentaires. Contrairement à d'autres techniques de représentations du signal, ces atomes ne sont pas recherchés au sein d'un dictionnaire pré-défini, mais sont extraits des données elles-mêmes. La factorisation en matrices non-négatives (NMF) est une technique analogue d'algèbre linéaire, qui réduit le rang tout en fournissant des atomes à valeurs exclusivement positives, donc plus facilement interprétables, et sémantiquement plus pertinentes pour la représentation de données elles-mêmes à valeurs positives [Lee et Seung, 1999]. Alors que d'autres travaux se concentrent soit sur la mise au point de dictionnaires, soit sur la décomposition de signaux à partir de ces dictionnaires, la NMF fournit conjointement un dictionnaire extrait des données et la décomposition de ces mêmes données dans ce dictionnaire.

Contributions de la thèse

Au cours de cette thèse, nous nous sommes particulièrement intéressée à cette technique et à ses variantes. À notre connaissance, il s'agit de la première thèse entièrement consacrée à son étude. Nous avons cherché à établir ses propriétés et à examiner différentes améliorations pouvant mener à une meilleure performance de transcription. Dans la littérature, les questions d'existence et de non-unicité des solutions, de problèmes de minima locaux et de convergence des algorithmes sont peu traitées, voire considérées comme des faits acquis. Notre première contribution est l'étude bibliographique, théorique et expérimentale de cette question. Nous montrons que toutes les fonctions de coût usuelles du problème de NMF sont sujettes aux minima locaux, et que les algorithmes multiplicatifs prédominants dans l'état de l'art échouent à les éviter, et ce même lorsqu'on adopte des stratégies d'initialisation structurée. Nous discutons en particulier du rapport entre fonction de coût et sémantique des données, et développons plusieurs arguments en faveur de l'usage de la divergence d'Itakura-Saito pour la décomposition NMF

de signaux audio, ainsi que de l'usage de contraintes supplémentaires, en sus de la non-négativité.

Nous proposons également un cadre bayésien permettant d'intégrer deux contraintes utiles : l'harmonicité du dictionnaire et la régularité temporelle des enveloppes. Nous montrons l'équivalence entre la NMF par minimisation de la divergence d'Itakura-Saito et l'estimation du maximum de vraisemblance dans un certain modèle de signal. Ce cadre théorique offre des propriétés de convergence que ne possèdent pas les algorithmes multiplicatifs.

Une partie importante de la thèse est dévolue au développement et à l'évaluation d'algorithmes originaux de transcription musicale fondés sur la NMF, qui constituent une part importante de nos contributions. Huit algorithmes de NMF, dont sept mis au point pendant la thèse et un appliqué pour la première fois à un problème de transcription musicale, sont évalués au sein d'un système original de transcription « WAV vers MIDI » complet.

Contributions en collaboration

À nos contributions personnelles viennent s'ajouter des travaux réalisés en collaboration. Le modèle harmonique proposé à la section V.3 (page 84), la modélisation probabiliste introduite dans la section VI.3 (page 97) et l'a priori de régularité temporelle développé dans la section VII.2 (page 106) sont issus de travaux en collaboration dans laquelle nos co-auteurs ont pris une part prépondérante. Ils sont par conséquent et à juste titre les premiers auteurs des publications qui en ont découlé [Vincent *et al.*, 2007, Vincent *et al.*, 2008, Févotte *et al.*, 2009].

Structure du document

Le mémoire est organisé en quatre parties. La première partie présente la tâche de transcription musicale, et dresse un état de l'art des différentes approches de cette tâche. La deuxième partie est consacrée à l'approche déterministe de la NMF tandis que la troisième l'aborde dans un paradigme probabiliste. La dernière partie présente l'application et l'évaluation des algorithmes proposés dans les parties précédentes.

Première partie : *Transcription automatique de la musique* (p.11)

L'observation de la partition musicale nous conduit au **chapitre I** à définir la tâche de transcription musicale, à constater sa richesse à plusieurs niveaux de lecture et de sens, et à mesurer ses enjeux en traitement du signal. Nous nous intéressons à la note comme objet musical élémentaire et examinons sa nature et ses propriétés, et présentons quelques éléments simples de théorie musicale permettant de motiver les approches scientifiques. Le **chapitre II** dresse un état de l'art des systèmes de transcription automatique, envisagé sous les angles de l'exploitation de la redondance temporelle et de l'usage de connaissances extérieures, de manière à motiver notre approche, dont la problématique est levée en fin de chapitre.

Deuxième partie : *Approche déterministe* (p.47)

Le travail effectué s'oriente suivant deux axes principaux. Dans le premier, le problème standard de NMF est étudié sous sa formulation usuelle, c'est-à-dire comme problème d'optimisation d'une

fonction de coût matricielle, dans un cadre purement déterministe. Au **chapitre III**, les fonctions et algorithmes usuels sont présentés, et leurs propriétés sont discutées d'un point de vue théorique : question de l'existence et de l'unicité des solutions, choix des fonctions de coût, propriétés de convergence des algorithmes, existence de minima locaux. Cette étude nous amène au **chapitre IV** à proposer et étudier expérimentalement deux stratégies d'évitement des minima locaux, d'une part par le biais de l'initialisation, d'autre part *via* un algorithme original, à fonction de coût variable, permettant d'atteindre des solutions approchées inaccessibles aux algorithmes multiplicatifs standards [Bertin *et al.*, 2009b]. Nous dressons ensuite au **chapitre V** un recensement des variantes du problème de NMF et en particulier, de l'emploi de contraintes en sus de la non-négativité des coefficients, telles que la parcimonie, la localisation, la continuité temporelle, qui permettent d'améliorer la forme et la sémantique de la factorisation résultante. Une nouvelle contrainte dite « d'harmonicité », propre à l'application musicale visée, est proposée ainsi qu'un algorithme multiplicatif permettant d'obtenir la factorisation harmonique (travail en collaboration, [Vincent *et al.*, 2007, Vincent *et al.*, 2008, Vincent *et al.*, 2010]).

Troisième partie : *Approche probabiliste* (p.93)

Dans un deuxième axe, on s'intéresse à une autre formalisation, probabiliste, du problème. Au **chapitre VI**, un état de l'art des liens entre modèles statistiques et NMF est présenté. On établit en particulier une équivalence formelle entre l'estimation du maximum de vraisemblance dans un modèle de somme de gaussiennes, d'une part, et la NMF des observations par minimisation de la distance d'Itakura-Saito d'autre part (travail en collaboration, [Févotte *et al.*, 2009]). Le travail précédent sur les contraintes est ensuite exploité, au **chapitre VII**, pour intégrer ces apports dans le modèle choisi, *via* un cadre bayésien. La régularité temporelle et l'harmonicité sont les deux contraintes que nous avons retenues dans cette partie, pour leur pertinence dans l'application visée. La contrainte harmonique est intégrée par l'adaptation du modèle proposé pour les observations, et la régularité temporelle *via* un a priori sur les paramètres à estimer. Plusieurs algorithmes originaux de type EM sont dérivés, implantés et étudiés [Bertin *et al.*, 2010, Bertin *et al.*, 2009a] du point de vue théorique (convergence) et pratique. Le **chapitre VIII** est consacré à une étude expérimentale de ces algorithmes sur données de simulation, visant à établir leur consistance et évaluer l'impact de la mauvaise estimation de certains paramètres.

Quatrième partie : *Application à la transcription* (p.127)

Enfin, la dernière partie de la thèse est consacrée à l'application de transcription musicale. Au **chapitre IX**, un système complet de transcription (du signal à la représentation MIDI), dont le cœur utilise les algorithmes de NMF précédemment exposés, est proposé, incluant une réflexion sur le choix de la représentation temps-fréquence, le choix de l'ordre du modèle et les post-traitements de la factorisation. Les différents algorithmes développés pendant la thèse sont confrontés à d'autres algorithmes de l'état de l'art en transcription, sur des signaux de musique en conditions réelles (pièces du répertoire classique de piano, polyphonie et vélocité élevée, conditions d'enregistrement réelles ou par synthèse logicielle de très haute qualité). Les performances sont évaluées quantitativement et qualitativement au **chapitre X.**

À l'issue de cette étude, nous concluons sur un bilan de nos travaux et envisageons les perspectives sur lesquelles ils pourraient s'ouvrir.

Première partie

Transcription automatique de la musique

Chapitre I

Présentation du problème et approches historiques

Résumé

Où l'on introduit la tâche de transcription musicale et ses enjeux, avant de présenter les premières approches historiques de résolution de ce problème.

I.1 Introduction

Par le mot *transcription*, on entend généralement une opération de conversion, de transformation d'une représentation de certaines données d'un domaine vers un autre, et en particulier, d'un domaine de bas-niveau, peu descriptif, vers un domaine de haut niveau, sémantiquement plus riche. Ainsi, dans le domaine du traitement de la parole, la transcription (également appelée « reconnaissance de la parole ») consiste à transformer le signal enregistré (forme d'onde) en la suite de mots qui ont été prononcés, sous une forme écrite. Pour prendre un exemple plus éloigné, en biologie moléculaire, la conversion au sein de la cellule de séquence d'ADN (longue suite de nucléotides non segmentée), en ARN (polymère court capable de produire une protéine fonctionnelle pour l'organisme), est également appelée transcription. En ce sens, elle se distingue de la *traduction*, qui consiste en une conversion vers une représentation de même niveau sémantique.

Dans le domaine du son musical, la transcription est une opération bien connue des musiciens, dont la formation inclut la fameuse et parfois traumatisante « dictée ». Au cours de cet exercice, l'élève doit noter sur papier à musique les notes jouées par l'enseignant, c'est-à-dire convertir en une représentation symbolique les variations de pression qui parviennent à son oreille. Cette tâche difficile demande apprentissage et expérience.

La transcription automatique de la musique, qui consiste à faire effectuer cette opération de dictée par un ordinateur, constitue un domaine de recherche très actif dans la communauté du traitement du signal, et dans celle dite de la « recherche d'information musicale » (*Music Information Retrieval*, ou MIR), car elle réunit et cristallise de très nombreux problèmes dont se préoccupent ces communautés : représentations temps-fréquences et questions de résolution, séparation de sources, utilisation de connaissances, apprentissage, classification, psycho-acoustique... la transcription musicale a même été décrite comme la « Quête du Graal » de ces communautés scientifiques [1]. En particulier, la littérature scientifique qui s'y rapporte est pléthorique.

Ce chapitre présente les problèmes recouverts par la notion de transcription musicale au sens le plus large (section I.2), et propose une restriction de cette tâche, qui sera l'objectif applicatif de cette thèse. La section I.3 expose ensuite les premières approches développées pour résoudre le problème de transcription.

I.2 La tâche de transcription

Dans cette section, nous envisageons la transcription musicale sous son acception la plus complète, c'est-à-dire comme l'opération visant à convertir la forme d'onde en une partition. La section I.2.1 examine la partition et ce qu'implique sa production. Cela nous amène à nous intéresser à un objet sonore en particulier, la note de musique, dans la section I.2.2, et à poser certaines restrictions à la tâche de transcription (section I.2.3) pour centrer notre problématique autour de cet objet. Notons à ce propos que nous nous limiterons à des musiques dont la représentation de choix est la partition musicale traditionnelle et dans lesquelles la note de musique est un objet pertinent, telles que la musique tonale occidentale ou le jazz [2].

1. "*Polyphonic transcription is the Holy Grail of the MIR community.*", Juan Bello, séminaire du 16.03.2006 à l'ENST.

2. De fait, nous excluerons donc de notre objet d'études d'autres types de musique telles que les musiques timbrale ou concrète.

I.2.1 Transcription au sens large

Observons la partition de la figure I.1. Limpide pour le musicien, elle paraîtra cabalistique à l'œil profane. D'une manière concise et très codifiée, elle contient en effet énormément d'information sur la musique qu'elle décrit :

- Des méta-informations de très haut niveau sur l'œuvre en elle-même : noms de l'auteur et du compositeur, leurs dates de naissance et de mort, opus et titre de la pièce, dédicace ;
- Des informations de moyen et haut niveau extractibles du contenu : la *tonalité principale (*armure), le *tempo, la *mesure (chiffrage et barres), les instruments en présence (voix chantée, piano) ;
- Des informations de bas niveau sur le contenu : notes (*hauteur, *valeur rythmique), paroles ;
- Des informations de nuance, d'intention, de jeu (*pianissimo*, *mässig*, *crescendo*).

Avec l'œil « traitement du signal », nous voyons que ces informations se rattachent à des domaines variés : notions temporelles (rythmes, durées) et fréquentielles (hauteur, tonalité), nécessitant ou pas le recours à des bases d'information extérieures (le nom du compositeur n'est évidemment pas extractible de la forme d'onde), problématiques de reconnaissance (timbre, paroles), de séparation (notes simultanées). Idéalement et *stricto sensu*, un système automatique de transcription devrait se préoccuper de tous ces aspects.

Remarquons toutefois que le symbole le plus présent sur cette page, et sans doute un des plus universellement connus même parmi les non-musiciens, est celui de la note : ♩. Unité élémentaire de la musique tonale occidentale, « brique » permettant de la construire, elle constitue un objet d'étude à la fois bien délimité mais complexe, qui va structurer notre approche de la transcription.

I.2.2 La note de musique

En théorie musicale, on dit généralement que la note de musique est un événement sonore défini par quatre paramètres ou « dimensions », chacune étant liée à la fois à une grandeur physique et à un corrélat perçu :

1. La **hauteur**, qui place la note sur l'échelle grave-aigu. Elle est physiquement reliée à la fréquence fondamentale de la partie stationnaire du son.
2. La **durée**, qui décrit la temporalité de la note, c'est-à-dire à la fois l'instant où elle démarre et sa valeur (durée quantifiée).
3. L'intensité, ou **sonie**, qui est la sensation appelée en langage courant « volume sonore ». Sa valeur est liée à l'amplitude du signal.
4. Le **timbre**, notion parfois définie de manière floue comme « tout ce qui n'est pas décrit par les trois autres dimensions », ou encore, « ce qui différencie la même note jouée par deux instruments différents ». Il est caractérisé entre autres par les enveloppes temporelle et spectrale de la note.

Un système de transcription complet devrait prendre en compte l'ensemble de ces dimensions. Toutefois, chacune constitue en elle-même un problème déjà complexe, et on trouve dans la littérature un ensemble de sous-tâches s'intéressant plus spécifiquement à l'une ou à l'autre :

- L'estimation de la ou des hauteurs. On rencontrera souvent les termes anglais *monopitch*, *multipitch*. Cette étude peut être qualifiée de « verticale » par analogie avec les dimensions de la partition, ou des représentations temps-fréquences. Parmi les disciplines liées, citons la définition et l'estimation de modèles sinusoïdaux, le suivi de *partiels...

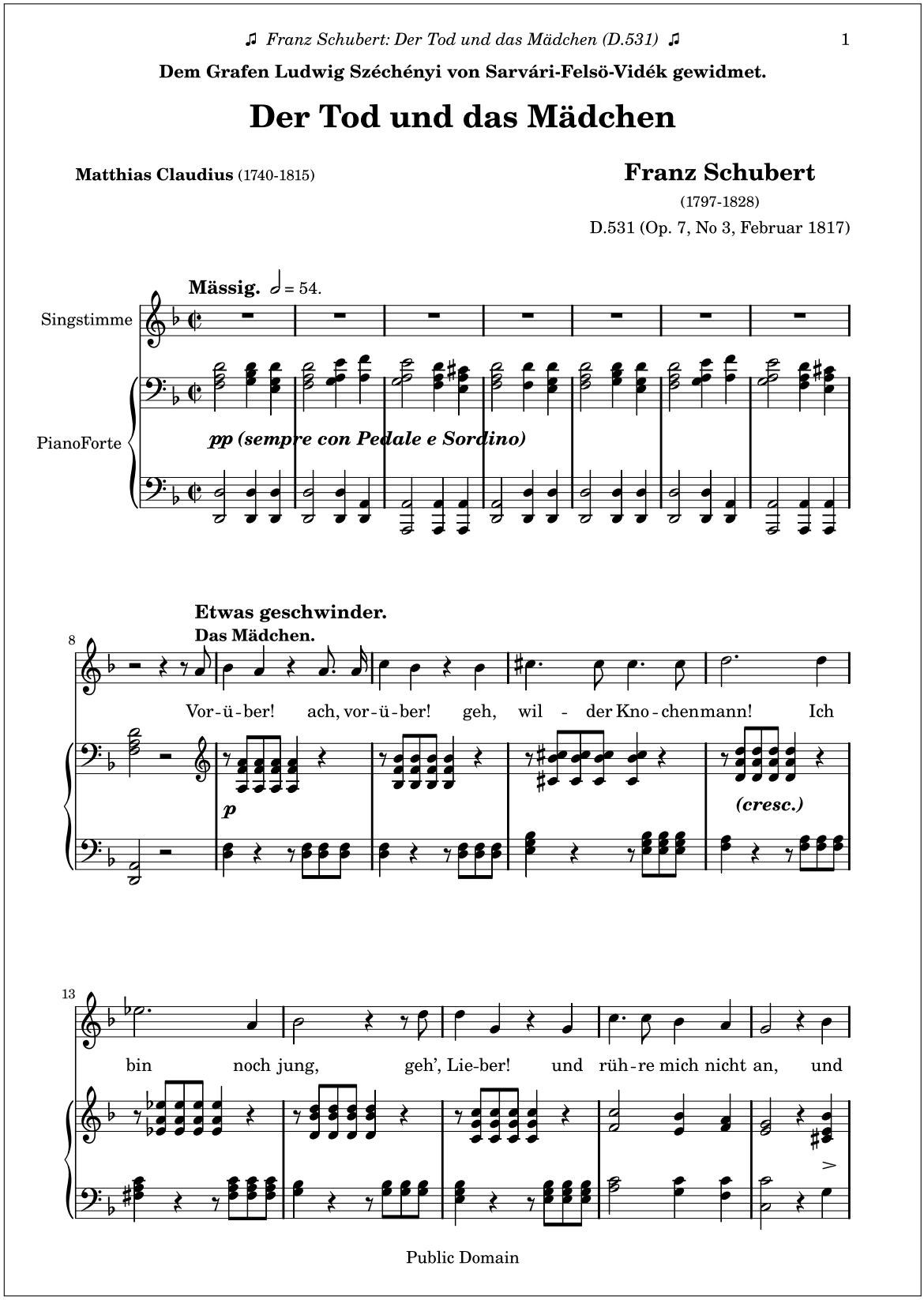


FIGURE I.1 – Extrait de partition (domaine public, `http://www.mutopiaproject.org`).

– L'étude de la dimension temporelle de la musique (étude « horizontale »), à différentes échelles de temps : détection d'attaque (en anglais *onset*), mais aussi à l'échelle du morceau, détection et suivi du *tempo (*beat-tracking*), étude hiérarchique de la *métrique.
– L'étude de la sensation d'intensité. De nombreux travaux de psycho-acoustique s'y rapportent (voir par exemple [Hartmann, 1998]).
– Le timbre est lié à la notion d'instrument (reconnaissance d'instruments) mais aussi à des notions de texture (similarité musicale).

Ces tâches sont loin d'être travail inutile : les systèmes de transcription plus génériques, on le verra dans le chapitre suivant, s'inspirent souvent de ces travaux plus ciblés.

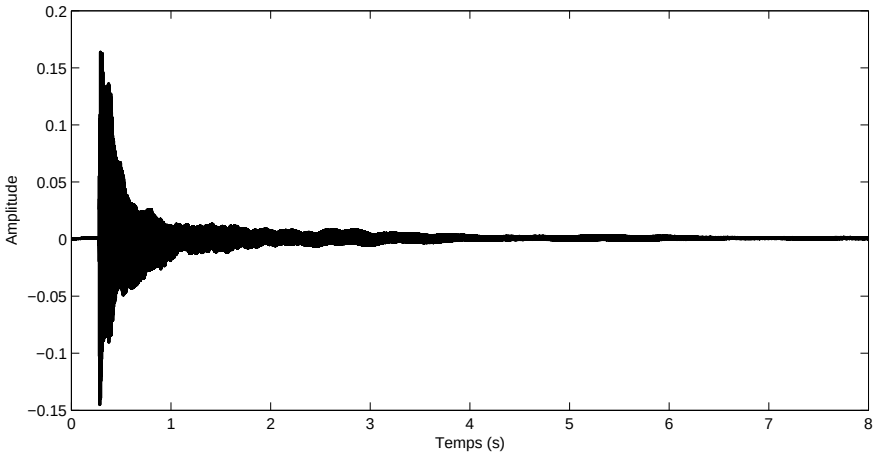
I.2.2.1 Représentations et visualisations

Les figures I.2, I.3 et I.4 présentent des représentations usuelles de trois notes de musique, respectivement jouées au piano, au violon et sur une caisse claire. Ces représentations classiques sont complémentaires et permettent de cerner différentes particularités des notes de musique, qui seront autant de problématiques au moment de les analyser.

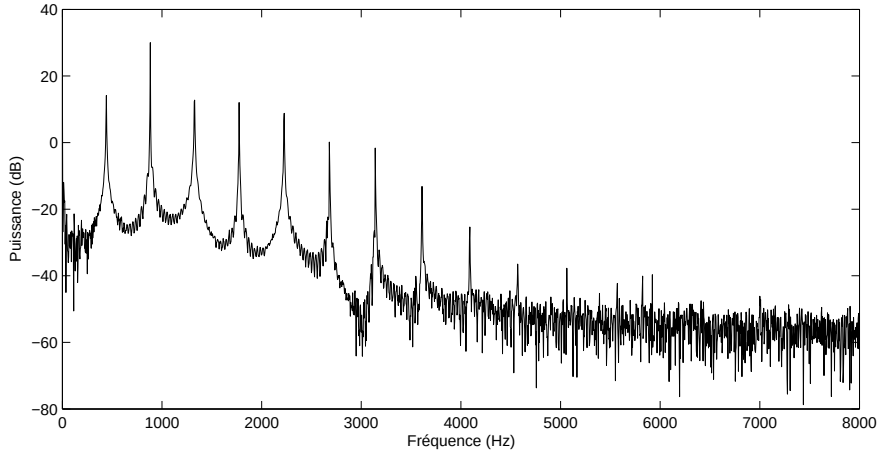
Dans le domaine temporel, on représente le signal sous sa forme d'onde, échantillonnée à une certaine fréquence f_s. Cette représentation permet de voir l'évolution temporelle de la note et sa délimitation, mais reste assez peu informative sur le contenu. Dans le domaine fréquentiel, le spectre décrit le contenu d'un segment stationnaire ou quasi-stationnaire. Il est calculé comme la valeur absolue de la transformée de Fourier discrète $TFD[x](f)$ du signal x, et nous permet d'observer une éventuelle structure. Ici, nous avons choisi manuellement la durée sur laquelle le spectre est calculé, de manière à analyser la plus longue durée possible où la note est quasi-stationnaire, et donc à bénéficier d'une bonne résolution fréquentielle. Malheureusement pour nous, la musique réelle ne nous offrira pas souvent ce genre de cas d'école (note bien isolée) et il n'est pas question de découper à la main les segments intéressants. Cela nous amène à la représentation temps-fréquence bien connue, le spectrogramme (issu de la Transformée de Fourier à Court Terme), qui permet d'avoir une visualisation nettement plus parlante du phénomène se déroulant sous nos yeux, mais qui nous oblige à faire un compromis entre les résolutions temporelle et fréquentielle.

Plusieurs remarques se dégagent de l'observation de ces figures :

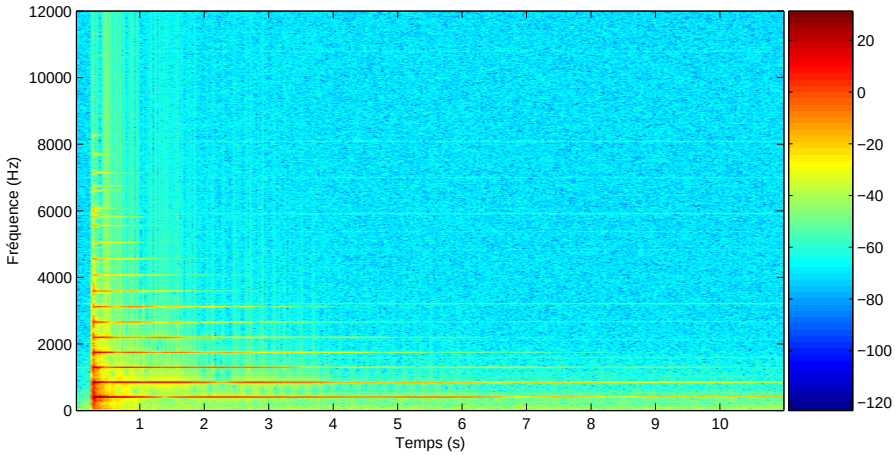
– La forme d'onde semble peu informative à cette échelle, et plutôt reliée à l'instrument qu'à la note jouée. Piano et percussion suivent, en toute logique, un schéma similaire, qu'on peut rapprocher de la physique (mode d'excitation : choc du marteau contre la corde, du maillet contre la membrane de la caisse claire). La forme d'onde du violon met clairement en évidence un phénomène de modulation d'amplitude, lui aussi à rapprocher de la physique (mouvement oscillant du doigt pressant la corde contre la touche).
– Les spectres des sons de violon et de piano exhibent une structure en peigne, qui met en évidence la pseudo-périodicité du son. Le son de percussion de la figure I.4 ne possède pas cette structure.
– Le spectrogramme met en rapport les dimensions temporelle et spectrale. C'est notamment la seule des trois représentations qui permet de réellement visualiser le *vibrato du violon. Il illustre également la forte dynamique énergétique, à un instant donné, entre les fréquences de faible et de forte énergie (la dynamique se compte en plusieurs dizaines de décibels).



(a) Forme d'onde.

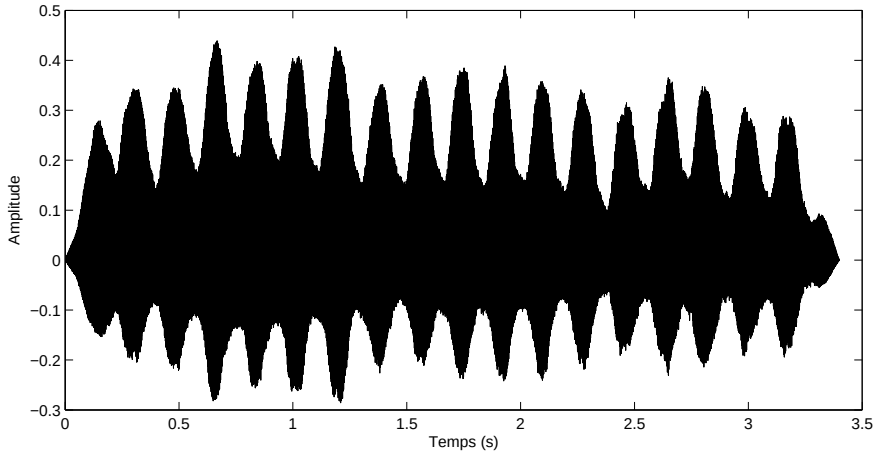


(b) Transformée de Fourier.

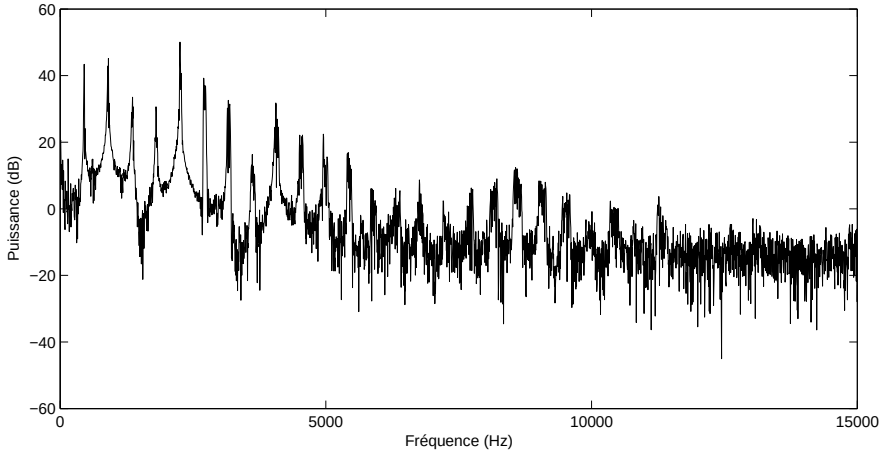


(c) Transformée de Fourier à Court-Terme.

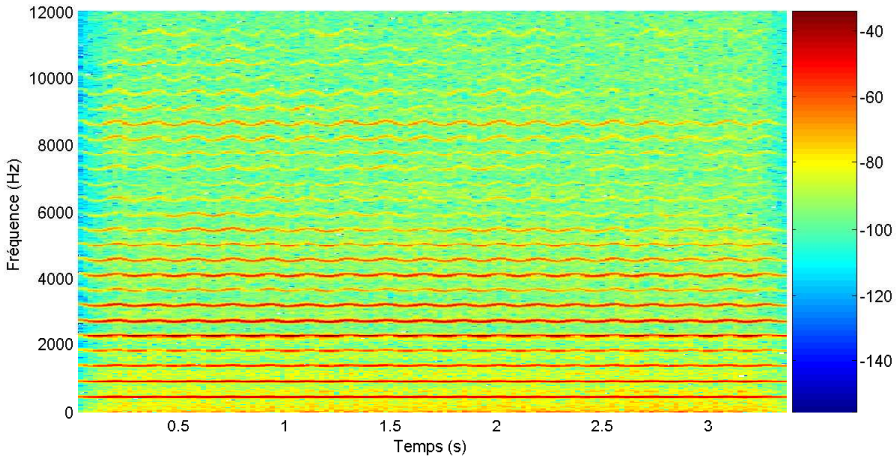
FIGURE I.2 – Représentations d'une note de piano.



(a) Forme d'onde.

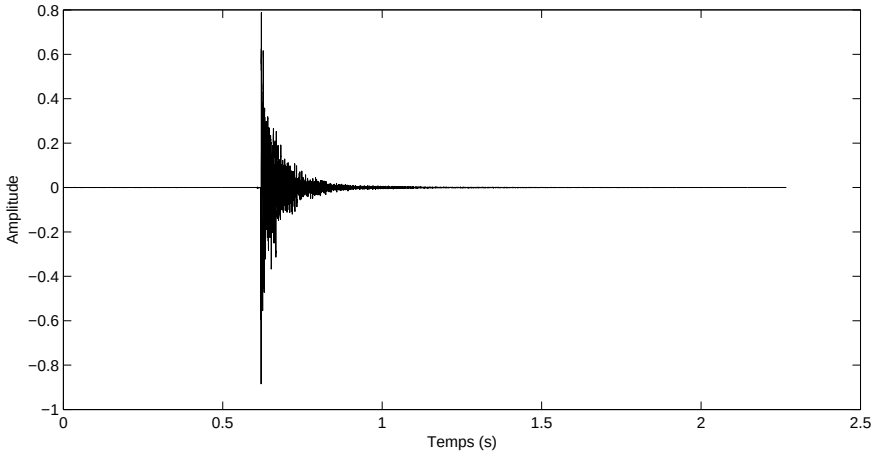


(b) Transformée de Fourier.

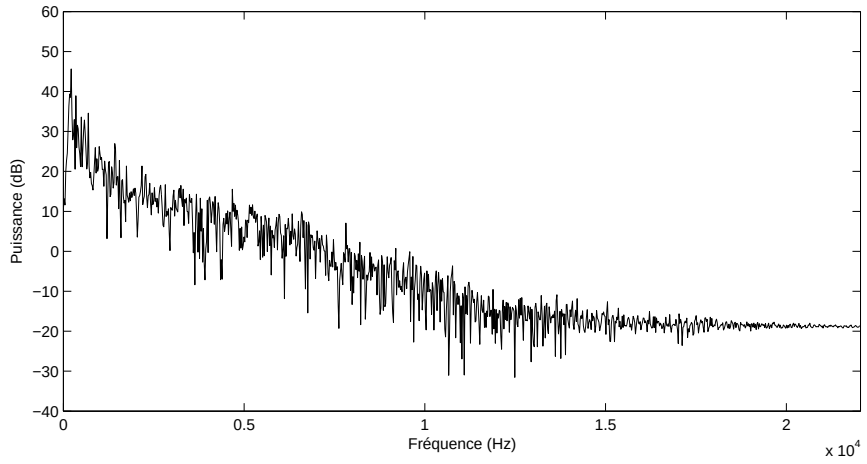


(c) Transformée de Fourier à Court-Terme.

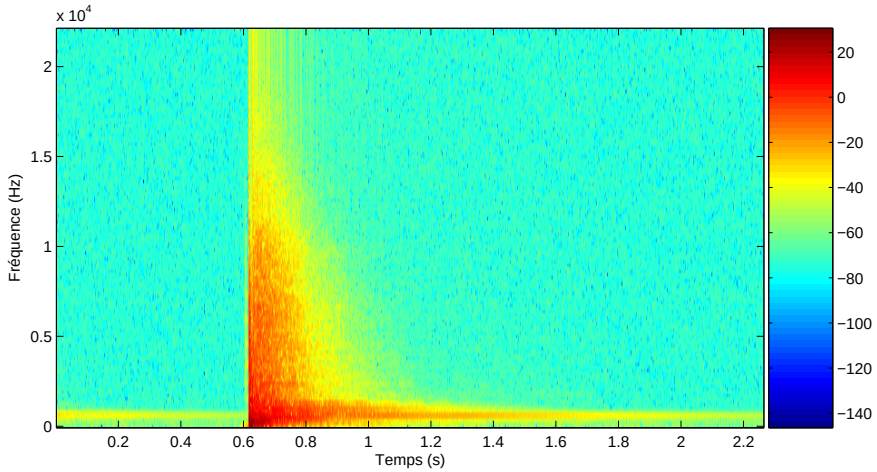
FIGURE I.3 – Représentations d'une note de violon avec *vibrato.



(a) Forme d'onde.



(b) Transformée de Fourier.



(c) Transformée de Fourier à Court-Terme.

FIGURE I.4 – Représentations d'un événement percussif (coup de caisse claire).

I.2.2.2 Hauteur musicale et fréquence fondamentale

On distingue deux types de sensations de *hauteur : la hauteur tonale et la hauteur spectrale. La hauteur dite « spectrale » désigne une *sensation* de hauteur difficilement quantifiable, généralement perçue par comparaison entre deux sons, même non-périodiques. Le sujet peut classer les sons entendus comme « plus graves », « plus aigus » les uns par rapport aux autres, comme par exemple pour la suite des *toms* de batterie. Les études psycho-acoustiques montrent que cette notion est reliée au « centroïde spectral » (centre de gravité du spectre de puissance).

La hauteur tonale, elle, est associée aux sons périodiques ou quasi-périodiques, et est caractérisée par sa période, ou sa fréquence fondamentale (f_0), inverse de la période. Un agrandissement (figure I.5) sur la forme d'onde du son de violon précédemment illustré permet de se convaincre de la quasi-périodicité du signal. La physique explique cette quasi-périodicité, partagée par de nombreux instruments de l'orchestre (notamment les instruments à sons entretenus) ; nous ne rentrerons pas dans les détails, voir par exemple [Leipp, 1971, Chaigne et Kergomard, 2008].

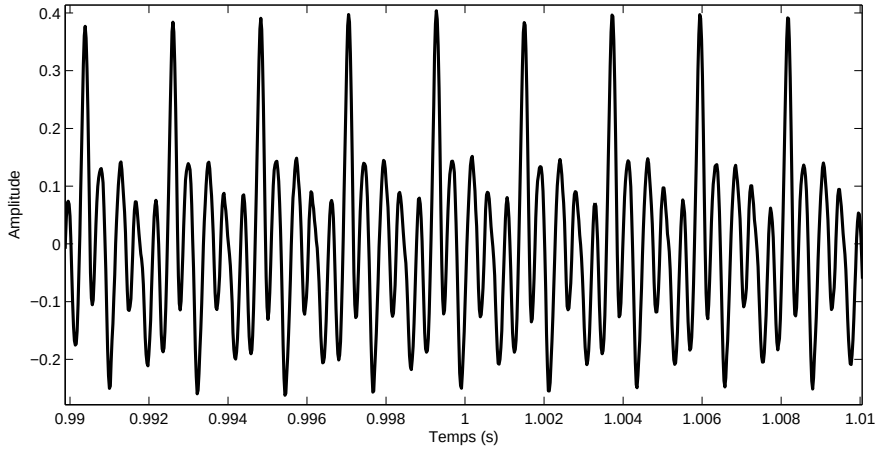


FIGURE I.5 – Zoom sur la forme d'onde.

Il existe donc une correspondance précise entre cette sensation subjective de hauteur et la physique : quantifiée en Hertz pour les scientifiques et sur l'échelle de la gamme (*do, ré, mi...*) pour les musiciens, elle sera indifféremment désignée par les vocables hauteur, fréquence fondamentale, ou par l'anglicisme *pitch* qui recouvre les deux notions. Phénomène le plus répandu et exploité dans la musique occidentale, c'est cette hauteur qui nous intéressera dans le contexte de transcription automatique.

Les théories classiques de la musique, comme par exemple [Chailley et Challan, 1951] font souvent mention du lien entre hauteur musicale et fréquence fondamentale. En effet, ce lien permet d'expliquer simplement la constitution de la gamme *diatonique[3] et les notions de *consonance et *dissonance, en s'appuyant sur la suite des harmoniques. Étant donné une note de fréquence fondamentale f_0, si l'on forme de nouvelles notes dont les fondamentales sont distribuées suivant la suite des partiels de cette note, nous obtenons :

– $f_1 = 2f_0$: l'octave

3. Ici, la gamme dite de « Zarlino ». Il existe plusieurs constructions théoriques possibles de la gamme : cycle des quintes de Pythagore, gamme tempérée chère à Bach... Le lecteur intéressé pourra se reporter avec profit à [Chailley et Challan, 1951].

– $f_2 = 3f_0$: la douzième, qui, ramenée dans une seule octave (intervalle $[f_0, 2f_0]$), devient la quinte
– $f_3 = 4f_0$: la double octave
– $f_4 = 5f_0$: la tierce majeure lorsque ramenée à l'octave

et ainsi de suite. Ainsi, la première suite de partiels de la note *do* produit l'accord parfait majeur « *do-mi-sol-do* », pilier de la musique tonale occidentale. La notion musicale, particulièrement subjective bien sûr, de « consonance » se fonde sur ce lien, et a des conséquences cruciales pour notre étude. En effet, les ensembles de sons considérés comme consonants, donc les plus fréquemment utilisés dans les musiques que nous serons appelés à analyser [4], présentent un très fort taux de recouvrement fréquentiel. Nous pouvons illustrer ce recouvrement simplement en visualisant le spectrogramme (figure I.7) de la séquence de la figure I.6, qui fait entendre l'accord parfait majeur de *do*, arpégé puis plaqué.

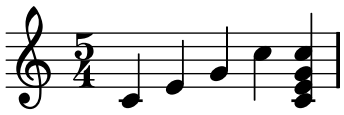


FIGURE I.6 – Accord parfait majeur arpégé et plaqué.

Sur ce son de piano synthétique, le recouvrement fréquentiel est bien visible, notamment entre la première note (do_4) et son octave (do_5) jouée en quatrième position, dont les fréquences sont toutes incluses dans celles de la fondamentale. L'accord final met bien en lumière la difficulté d'estimer les quatre notes présentes sans information complémentaire.

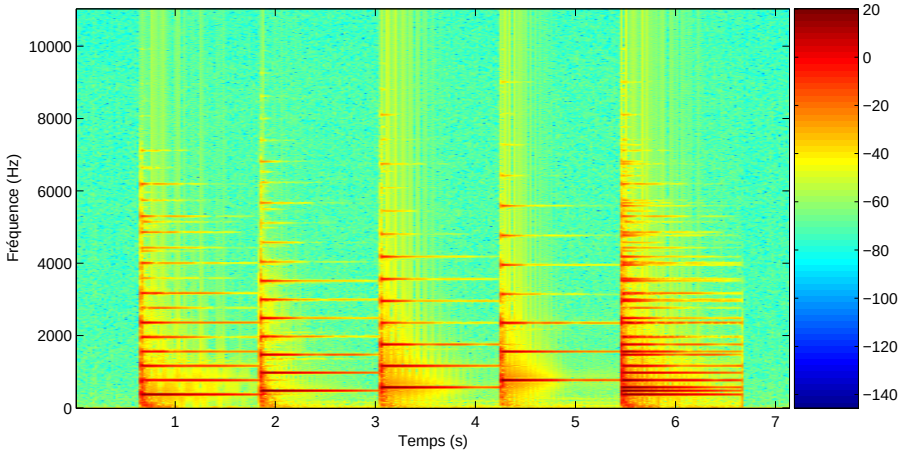


FIGURE I.7 – Spectrogramme de l'accord parfait majeur.

Un autre point mérite d'être discuté pour éclairer la suite de ce mémoire : la physiologie de la perception humaine de la hauteur tonale, et sa modélisation. Ces aspects étant largement traités dans nombre d'ouvrages de référence tels que [Leipp, 1971, Hartmann, 1998], nous nous contenterons d'en donner les grandes lignes, dans la perspective de faire le lien avec les traitements automatiques de la littérature.

4. Ils sont considérés dans l'esthétique classique occidentale comme « agréables à l'oreille », notion très culturelle et discutable, mais leur omniprésence dans la production musicale est indéniable.

On considère généralement que la perception de la hauteur tonale résulte de la contribution de deux phénomènes : le codage dit « tonotopique », en lien avec le contenu fréquentiel, et le codage temporel. Le codage tonotopique est une véritable analyse spectrale du signal : une membrane de l'oreille interne, la membrane basilaire, de forme trapézoïdale, a des propriétés mécaniques lui permettant de vibrer suivant un certain « patron » d'excitation (*pattern*) dont la localisation le long de la membrane dépend directement des fréquences du son [5]. Le codage temporel, pour sa part, correspond à la cadence et la forme des signaux nerveux (potentiels d'action) transmis depuis la membrane basilaire (*via* les cellules ciliées) vers le cerveau. Il existe nombre de preuves physiologiques en faveur d'une double exploitation par le cerveau de ces informations temporelles et spatiales pour générer la sensation de hauteur tonale. On retrouvera ainsi dans la littérature cette dualité entre traitement temporel et spectral du son.

L'étude plus poussée de la perception humaine et de l'ensemble du système auditif (oreille externe, moyenne, interne, nerf auditif et traitement central) et sa modélisation en traitement du signal font partie des guides de conception des systèmes cherchant à reproduire leurs performances. [Meddis et Hewitt, 1991] propose par exemple une chaîne de traitement complète, sous forme de blocs associés à la fois à un étage physiologique et à une mise en œuvre possible en traitement du signal. Cette chaîne associe des blocs de filtrage (modélisant le pavillon, l'oreille externe et moyenne), des bancs de filtres simulant l'action tonotopique de la membrane basilaire, et une modélisation de la transduction réalisée par les cellules ciliées (conversion de l'excitation mécanique de la membrane en excitation électrique des fibres nerveuses), prenant en compte divers phénomènes physiologiques tels que la période réfractaire [6] des cellules nerveuses. On retrouve nombre de ces idées et modèles dans la littérature sur l'estimation de hauteur.

I.2.2.3 Enveloppe temporelle

Si l'on examine les formes d'onde de la note de piano et du coup de caisse claire, elles suivent un schéma similaire dans la forme. Ce schéma peut être très grossièrement modélisé par le modèle d'enveloppe dit « ADSR », modèle particulièrement utilisé pour la synthèse de sons dans les années 1980. Ce modèle est représenté sur la figure I.8. Il décrit le déroulement de la note comme la succession de quatre phases :

1. **A**ttack (*attaque*) : intervalle de temps débutant au moment de l'excitation de l'instrument (frappe de la touche, établissement du contact corde/archet...) et se terminant lorsque le son a atteint son intensité maximale. On parle également de « transitoire ».
2. **D**ecay (*relâchement*) : intervalle de temps débutant à la fin de l'intervalle d'attaque et correspondant à un amortissement du signal jusqu'à sa stabilisation en intensité.
3. **S**ustain (*entretien*) : phase durant laquelle l'amplitude de l'enveloppe reste plus ou moins constante après la fin de l'intervalle de Decay, jusqu'à ce que le contact excitateur-résonateur soit rompu.
4. **R**elease (*décroissance*) : délai d'extinction de la note, entre le moment du relâchement et le moment où la note devient inaudible (amplitude nulle).

Ce modèle schématique contient de nombreuses limites, mais permet de mettre en lumière de possibles difficultés pour la détection et le traitement de cette enveloppe, au moment de délimiter le début et la fin d'une note analysée. La définition et la durée de la phase d'attaque n'est pas triviale :

5. D'où le terme « tonotopique » : du grec $\tau\acute{o}\nu o\varsigma$, tension, $\tau\acute{o}\pi o\varsigma$, lieu, pour désigner le fait qu'une région spatiale lui est associée. En effet, les Grecs, excellents physiciens, associaient par le biais du même terme le ton, hauteur de la note, et la tension de la corde qui le produisait.

6. Période suivant immédiatement un potentiel d'action nerveux, au cours de laquelle la fibre n'est plus excitable.

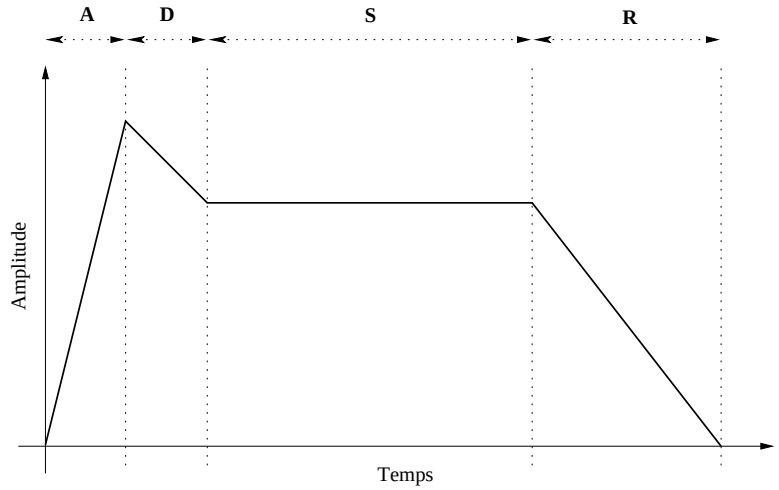


FIGURE I.8 – Modèle ADSR de l'enveloppe temporelle d'une note.

elle peut être longue et progressive pour des instruments à cordes frottées ou à vents, très brève pour les percussions ou pour des instruments joués *staccato. Le relâchement n'est pas toujours observable en pratique. L'amplitude peut varier de manière importante au cours de la phase d'entretien (par exemple en cas de *vibrato ou de *tremolo). La phase d'entretien peut même être réduite à néant (dans le cas d'un instrument percussif). La décroissance peut être écourtée ou inachevée (notes jouées *legato, notes répétées, phénomènes de coarticulation). C'est pourquoi la détection d'*onset* est un problème en soi, largement traité dans la littérature, pour lui-même ou en préambule d'une étude plus poussée du rythme (voir par exemple [Alonso-Arevalo, 2006]).

I.2.3 Restriction du problème

Nous voyons que les seules questions d'étude de la note *via* l'estimation de sa hauteur, de son début et de sa fin sont d'ores et déjà des questions très riches. Nous décidons donc de nous limiter dans cette étude à ce problème, dans un cadre polyphonique c'est-à-dire lorsque, à un instant donné, plusieurs notes se font entendre. Nous ne nous occuperons donc pas d'attribuer les notes détectées à différents instruments, voix ou lignes mélodiques ; nous ne chercherons ni à placer des barres de mesure, ni à détecter la métrique ou le tempo, ni à quantifier les durées des notes en valeur. Enfin, nous ne chercherons pas à estimer la nuance de jeu.

La tâche que nous nous proposerons de traiter peut donc s'exprimer de manière concise comme une tâche de « WAV vers MIDI ». En effet, le format descriptif MIDI est un bon choix pour décrire une musique sous la forme d'une succession de notes caractérisées par leur hauteur et leurs instants de début et de fin. Dans cette représentation, les hauteurs sont quantifiées avec un pas d'un demi-ton, chaque numéro étant associé à une note (par exemple : le numéro 69 est associé au la_4 c'est à dire le *la* du *diapason à 440 Hz, *cf.* table B.1 page 188), sans chercher à résoudre les ambiguïtés dues aux *enharmonies. Partant d'un fichier MIDI (SMF, pour *Standard MIDI File*), on peut obtenir une représentation temps-hauteur appelée « pianoroll » qui schématise notre objectif. Le pianoroll du début du morceau décrit par la partition de la figure I.1 est présenté sur la figure I.9. Notons que le format MIDI permet également, de manière optionnelle, de stocker d'autres informations (tempo, instruments...) dont nous ne tiendrons pas compte ici.

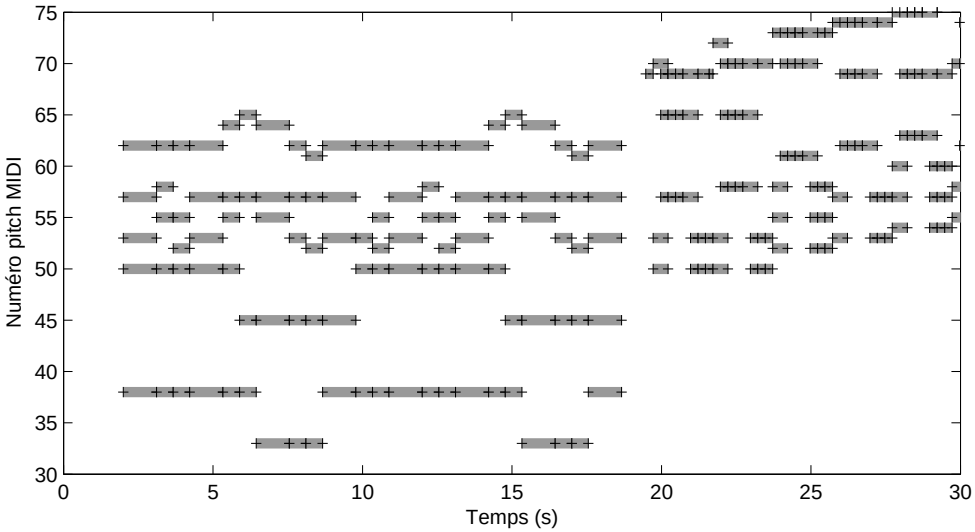


FIGURE I.9 – Pianoroll du début de *La jeune fille et la mort* (F. Schubert). Les débuts et fins des notes sont signalés par le symbole « + ».

I.3 Approches historiques

L'estimation de fréquence fondamentale dans les signaux audio-fréquences remonte au milieu des années 1960, en particulier dans le domaine du traitement de la parole [Rabiner *et al.*, 1976], les sons de parole dits « voisés » possédant en effet une répartition harmonique de fréquences (due à la vibration des cordes vocales). De nombreux travaux dans ce domaine sont donc issus de cette communauté. Toutefois, les problématiques propres à la musique, et en particulier la *polyphonie, conduisent à des travaux originaux dès la même période [Moorer, 1975]. Ces approches historiques sont présentées ici.

I.3.1 La transcription monodique

Tant pour des raisons historiques (travaux antérieurs en traitement de la parole) que pour des raisons de simplification du problème, les premières transcriptions s'intéressent aux sons dits **monodiques*, c'est-à-dire tels qu'une seule note est jouée à la fois[7]. Historiquement, la transcription monodique se confond quasiment avec l'estimation monopitch (estimation d'une seule fréquence fondamentale, ou hauteur simple), que nous présentons ici. Les techniques d'estimation se répartissent principalement en deux catégories, inspirées de la perception humaine : les approches dans le domaine spectral et les approches dans le domaine temporel.

I.3.1.1 Méthodes temporelles

La première catégorie de méthodes se place dans le domaine temporel, partant du constat que l'on examine des séries chronologiques périodiques ou quasi-périodiques, et que la pseudo-période P_0 doit pouvoir être obtenue en cherchant une ressemblance maximale entre le signal observé $\mathbf{x}$ et des versions

7. Nous préférons ce terme à *mélodique*, qui contient une connotation sémantique.

décalées de lui-même dans le temps.

En vertu du théorème de Fourier, et considérant des signaux échantillonnés — donc ne contenant pas de fréquences au-delà de la fréquence de Nyquist $f_s/2$ — le signal quasi-périodique x_n peut s'exprimer suivant un modèle sinusoïdal simple :

$$x_n = \sum_{h=1}^{H} 2A_h \cos(2\pi hn/P_0 + \varphi_h) + b_n, \tag{I.1}$$

où : P_0 : période fondamentale réduite
H : nombre d'harmoniques du signal (choisi tel que $Hf_0 < 1/2$)
A_h : amplitude du partiel h
φ_h : phase du partiel h, variable aléatoire de loi uniforme sur $[0, 2\pi]$
b : bruit blanc centré de variance σ^2, indépendant des phases

Sous ce modèle, on peut voir le signal $\mathbf{x}$ comme un processus centré et stationnaire au sens large (SSL). On peut donc définir sa fonction d'autocovariance R_{xx} :

$$R_{xx}(m) \stackrel{\text{def}}{=} \mathbb{E}[x_n x_{n+m}] = \sum_{h=1}^{H} 2A_h^2 \cos(2\pi hm/P_0) + \sigma^2\delta(m) \tag{I.2}$$

L'autocovariance mesure effectivement la ressemblance du signal avec des versions décalées de lui-même. Son intérêt est de faire disparaître les déphasages entre les partiels et l'influence du bruit hors de $m = n$. Ainsi, R_{xx} atteint son maximum lorsque m est un multiple de P_0.

Cette idée est à la base de la méthode de référence de [Rabiner, 1977], basée sur l'autocorrélation (ACF), une version normalisée de l'autocovariance, et déclinée sous de nombreuses variantes ou techniques reliées, telles que l'usage du cepstre (transformée de Fourier inverse du logarithme du module de la transformée de Fourier [Noll, 1967]) ou de la fonction de différences d'amplitudes moyennes (ou AMDF, pour *Average Difference Magnitude Function* [Ross *et al.*, 1974]) ; l'enjeu de ces méthodes revient à estimer cette fonction d'autocovariance, évidemment inconnue. Signalons également qu'une modélisation probabiliste du problème permet de faire le lien entre l'autocorrélation et une estimation de type « maximum de vraisemblance » (MV) de P_0 [Wise *et al.*, 1976].

Une fois l'ACF estimée, le moyen le plus simple d'estimer P_0 est de choisir la valeur qui réalise son maximum. Cependant, compte-tenu de la périodicité, des pics peuvent apparaître à $P = 2P_0, 4P_0...$ rendant la méthode susceptible de produire des erreurs de sous-octave. De plus, cette méthode est sensible à l'enveloppe spectrale du signal.

I.3.1.2 Méthodes fréquentielles

Dans le domaine fréquentiel, nous avons vu sur les figures I.2 et I.3 que le spectre de la note est constitué de pics régulièrement espacés (on parle de distribution *harmonique, ou quasi-harmonique). Ceci se retrouve évidemment si l'on écrit la transformée de Fourier du modèle (I.2). Les méthodes fréquentielles d'estimation de hauteur consistent à estimer ce « peigne » pour obtenir la fréquence fondamentale $f_0 = 1/P_0$, visualisée comme « l'écart entre les dents du peigne ». La méthode « brutale » qui consisterait à choisir, par exemple, f_0 au pic maximum du spectre, serait susceptible de produire des erreurs d'octave et manquerait de robustesse.

Dans ce domaine, [Schroeder, 1968] est le travail de référence. Après calcul du spectre et détection des pics, l'auteur propose de former un histogramme en fonction de la fréquence ; pour chaque pic

du spectre, on dénombre les pics dont la fréquence est un multiple de la fréquence du pic considéré. Le maximum de l'histogramme doit alors correspondre à la fréquence fondamentale, puisque c'est la fréquence qui rassemble le plus grand nombre de multiples. Ce principe étant posé, l'auteur propose ensuite de pondérer les contributions des pics dans le dénombrement par leur amplitude, puis de supprimer l'étape de détection de pics et de construire le même histogramme pour chaque point fréquentiel. Ceci aboutit à un estimateur de fréquence fondamentale aujourd'hui extrêmement répandu : la **somme spectrale**, définie par :

$$S(f) \stackrel{\text{def}}{=} \sum_{h=1}^{H} |X(hf)|^2 \tag{I.3}$$

où H désigne un nombre de partiels fixé, et choisi tel qu'on ne dépasse pas la fréquence de Nyquist. Cette fonction devrait atteindre son maximum en f_0.

Une autre variante peut être obtenue en pondérant les contributions non pas par l'amplitude des partiels, mais par leur logarithme, ce qui conduit au **produit spectral** :

$$P(f) \stackrel{\text{def}}{=} \prod_{h=1}^{H} |X(hf)|^2 \tag{I.4}$$

où X est la transformée de Fourier discrète du signal. Ces méthodes sont particulièrement efficaces lorsque l'on peut fixer le nombre H de partiels, c'est-à-dire lorsque l'on recherche f_0 dans un intervalle $[f_{min}, f_{max}]$ relativement restreint (puisqu'on impose $H f_{max} < 1/2$). Ceci concernera donc des instruments de ⋆tessiture limitée, ou des pièces d'⋆ambitus raisonnable.

Comme le remarque [Emiya, 2008], de très nombreuses méthodes d'estimation de hauteur aboutissent à des variantes de la somme ou du produit spectraux, bien que leur point de départ puisse être sensiblement différent de celui de [Schroeder, 1968]. Par exemple, [Doval et Rodet, 1991] et [Brown, 1992] proposent une estimation fondée sur des produits scalaires entre le spectre à estimer et des spectres de référence (méthode dite de *pattern matching*), dont la formulation revient à une somme pondérée ; [Klapuri, 2005] définit une fonction dite de « saillance », fonction de la variable temporelle et liée à l'autocorrélation, mais qui peut également être interprétée comme une variante de la somme spectrale. Enfin, signalons que la somme spectrale calculée sur le module au carré du spectre est également l'estimateur du maximum de vraisemblance du paramètre f_0 dans le modèle de l'équation (I.1) lorsque le bruit est gaussien.

Un des intérêts de la somme et du produit spectraux est de ne faire aucune hypothèse sur la présence ou l'absence d'énergie à la fréquence fondamentale, ni sur le fait que cette énergie devrait être supérieure à celle des partiels, par exemple. Ceci rend la méthode robuste aux fondamentales absentes et à l'enveloppe spectrale.

Pour conclure cette partie, signalons enfin que certaines méthodes comme [Hess, 1983, Peeters, 2006] associent des approches temporelle et fréquentielle afin de compenser leurs défauts réciproques et combiner leurs avantages.

I.3.2 Premiers systèmes de transcription polyphonique

Parallèlement à ces travaux pionniers sur la fréquence fondamentale, d'autres auteurs décident dans la même période de s'atteler à la question de la transcription musicale avec un regard de plus haut niveau. Leurs approches sont présentées ici.

I.3.2.1 Premiers systèmes

La transcription automatique de la musique en tant que problématique en soi a été introduite par les travaux de [Moorer, 1975], dans lesquels l'auteur définit en détail le problème, et propose un premier système de transcription. L'objectif théorique est de générer une partition, par un ordinateur, à partir d'un enregistrement musical. Un certain nombre d'hypothèses simplificatrices permettent alors de restreindre ce vaste cadre et de concevoir un système de transcription élémentaire : limitation de la polyphonie à deux voix, absence de notes simultanées en rapport harmonique et notamment d'*unisson, absence de *vibrato ou de *glissando.

Ces travaux soulèvent les principales questions à résoudre dans le cadre de la transcription :

- le choix d'une représentation adaptée des données, c'est-à-dire la conversion de la forme d'onde vers un domaine mettant en valeur les caractéristiques du signal, en particulier son contenu fréquentiel. Cette représentation doit servir de base à l'estimation ultérieure des notes ;
- le problème des fréquences fondamentales en rapport harmonique et du recouvrement spectral entre notes ;
- l'intégration temporelle de cette information, c'est-à-dire le passage d'une succession de fréquences fondamentales à la construction de l'objet « note » ;
- la définition de l'ensemble des éléments à transcrire, à différents niveaux sémantiques : les instruments, les voix (mélodie, accompagnement...), le tempérament, le rythme, etc.

Dans la même période, [Piszczalski et Galler, 1979] présente un algorithme d'estimation multipitch fondé sur l'examen des rapports arithmétiques de fréquences, avec un fonctionnement de type hypothèses/validation. Dans ce papier, l'auteur développe des considérations similaires à celles de [Moorer, 1975] sur la tâche de transcription musicale.

Ces approches historiques adoptent un paradigme similaire qui imprégnera nombre de systèmes ultérieurs que nous évoquerons au chapitre suivant : une démarche dite *bottom-up*, c'est-à-dire partant du plus bas niveau (la forme d'onde) pour s'élever progressivement sur l'échelle sémantique, d'abord par un changement de représentation, puis par l'estimation des fréquences fondamentales sur des segments, pour arriver au niveau de la note et éventuellement à des niveaux d'information plus élevés.

I.3.2.2 Tableau noir

Dans les années 1990, un autre type d'approches voit le jour. Contrairement aux précédentes, celles-ci cherchent à intégrer des informations extérieures au signal lui-même et envisagent une approche *top-down*, c'est-à-dire l'utilisation précoce de connaissances de haut-niveau pour informer et aider l'analyse du signal. Cette approche est dite « tableau noir » (*blackboard*), par analogie avec un groupe d'experts cherchant à résoudre un problème devant un tableau noir, chacun intervenant seulement lorsque son domaine d'expertise est requis. De nombreux systèmes de la littérature utilisent et/ou considèrent ce paradigme comme une approche historique majeure de la transcription musicale [Martin, 1996b, Hainsworth, 2001, Plumbley *et al.*, 2002].

La méthode du tableau noir est née à l'Université du Massachussetts en 1993 via le schéma IPUS (*Integrated Processing and Understanding of Signals*) [Lesser *et al.*, 1993] mais s'inspire des principes de conception des systèmes-experts de l'intelligence artificielle des années 1970. Son principe consiste à extraire de l'information du signal original trame par trame, puis à trouver les paramètres qui décrivent ce signal en utilisant un ensemble de sources de connaissances mises en compétition et d'hypothèses à plusieurs niveaux de description. Ainsi, le système IPUS combine un traitement *bottom-up* du signal et

l'utilisation *top-down* d'a priori ou d'informations globales pour choisir entre différentes hypothèses inscrites au tableau noir. Le système est dynamique, en ce que les paramètres d'analyse bas-niveau peuvent être ajustés en fonction des niveaux supérieurs d'information, avec un système d'« allers-retours » entre les hypothèses et l'analyse. Plusieurs auteurs arguent que ce fonctionnement pourrait être similaire à celui de l'écoute humaine, qui traiterait en permanence l'information de manière complémentaire à la fois sur les plans auditifs (oreille, nerf) et cognitif (cerveau).

Le système est principalement constitué de trois parties : le tableau, où les hypothèses sont proposées et examinées ; les sources de connaissances, qui interviennent pour aider l'analyse et apporter l'information nécessaire à chaque étape ; le contrôleur (*planner* ou *controller*), qui gère les interactions entre ces deux parties, en sollicitant le niveau de connaissances approprié.

Le schéma IPUS en tant que tel n'a jamais été appliqué au problème de la transcription audio, mais il a largement inspiré de nombreux travaux dans ce domaine [Martin, 1996b, Martin, 1996a, Godsmark et Brown, 1999, Bello *et al.*, 2000, Plumbley *et al.*, 2002].

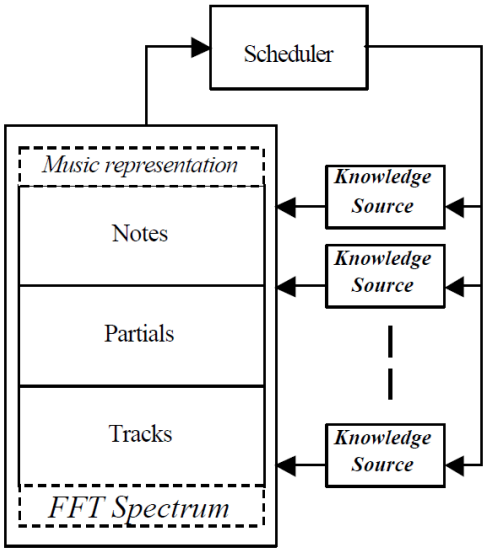


FIGURE I.10 – Système en tableau noir pour la transcription musicale. D'après [Bello *et al.*, 2000].

[Bello *et al.*, 2000] décrit en particulier le fonctionnement chronologique et les sous-tâches d'un système de transcription de type tableau noir. Le contrôleur établit des priorités entre les sources de connaissance et détermine l'ordre dans lequel leur action sur le tableau va s'effectuer. Chaque source de connaissance est représentée sous la forme d'une paire condition/action (*if/then*). Lorsque la condition d'une certaine source de connaissance est satisfaite, le contrôleur déclenche l'action et place son résultat sur le tableau. Les actions réalisées peuvent consister soit à supprimer du tableau des hypothèses infirmées, soit à déclencher une action d'analyse, comme la ré-estimation des partiels en fonction d'une hypothèse sur la note. La figure I.10 illustre la mise en œuvre globale du système proposé.

En ce qui concerne les sources de connaissance et leurs interactions, la figure I.11 [Martin, 1996a] en propose une illustration globale. Elle montre la hiérarchie des hypothèses dans un cas de transcription musicale. Chaque source de connaissance est représentée par un graphe, dont chaque sommet est une hypothèse du tableau noir, et dont les arcs symbolisent une relation de type « causer un changement sur ». Les sommets d'où partent les flèches représentent les hypothèses satisfaisant les préconditions

de la source de connaissance ; les sommets où aboutissent ces flèches sont les hypothèses modifiées par l'action de la source de connaissance. Les sommets représentés en blanc sont des hypothèses dites « concurrentielles », c'est-à-dire qui mettent en compétition plusieurs sous-hypothèses et tranchent parmi elles.

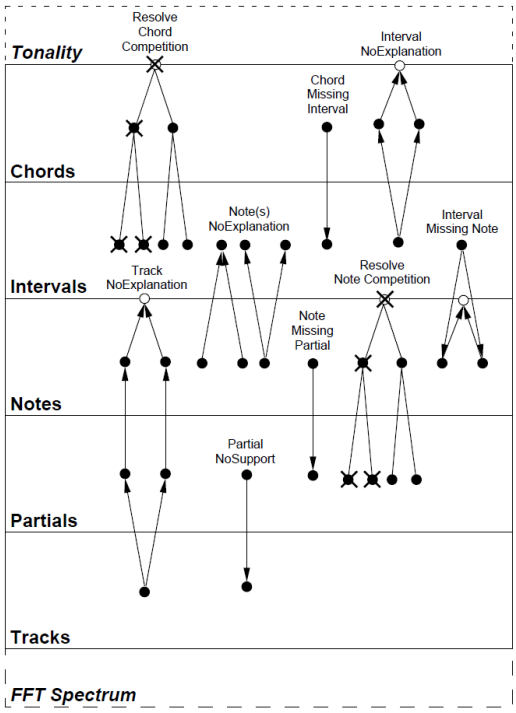


FIGURE I.11 – Exemple de sources de connaissance. D'après [Martin, 1996a].

Au premier abord, ces systèmes en tableau noir semblent être un bon moyen de combiner des informations extérieures de haut niveau avec un flux d'analyse *bottom-up*. Cependant, ils sont heuristiques par nature, et leur performance dépend considérablement des caractéristiques précises de leur mise en œuvre. En effet, s'ils proposent une architecture générale pour aborder le problème, ils laissent une grande latitude de choix quant à l'implantation des sources de connaissances et de leurs actions, les niveaux des hypothèses considérées et la manière de trancher entre elles. Au chapitre suivant, nous étudierons plus précisément les techniques mises en œuvre et implantées concrètement dans les systèmes de l'état de l'art.

Dès ces premiers systèmes, il se dégage des tendances qui vont imprégner durablement les développements futurs de systèmes plus sophistiqués :

- Le rôle central de l'estimation de hauteur fondamentale à l'échelle de la trame. Il constitue un problème en soi, et soulève la question de l'intégration temporelle de l'information de fréquence.
- L'emploi de connaissances extérieures aux données analysées elles-mêmes.

L'état de l'art des systèmes de transcription musicale qui est présenté au chapitre suivant s'intéressera particulièrement à ces deux angles d'éclairage.

Chapitre II

État de l'art

Résumé

Où l'on dresse l'état de l'art des approches et systèmes de transcription musicale, sous l'angle qui motive notre démarche.

II.1 Introduction

La transcription automatique de la musique a déjà été l'objet d'une abondante littérature ; les références sont nombreuses et complètes et il n'est pas dans notre ambition d'être ici exhaustive. Des bibliographies très complètes pourront être consultées par exemple dans [Cemgil, 2004, Hainsworth, 2004, Emiya, 2008]. Comme introduit précédemment, nous choisissons ici deux angles d'approche pour présenter brièvement l'état de l'art et motiver notre propre travail :

- Suivant la manière d'utiliser l'ensemble du signal et sa redondance intrinsèque ;
- Suivant la quantité et la nature de connaissances extérieures éventuellement apportées au système.

Toute entreprise de catégorisation grossière de la littérature est évidemment vouée à l'échec, chaque système méritant sa propre catégorie, et les frontières étant souvent trop floues pour classer l'un ou l'autre des systèmes dans une catégorie donnée. Consciente de ce constat, nous proposons malgré tout des délimitations que nous espérons éclairantes. Dans ce chapitre, nous présentons en premier lieu (section II.2) les techniques d'estimation de fréquences fondamentales multiples, en considérant en particulier comment leur résultat est intégré temporellement pour produire une transcription. Nous présentons ensuite divers systèmes de transcription de l'état de l'art répartis en trois catégories : les systèmes informés, utilisant des connaissances extérieures au signal (section II.3), les méthodes bayésiennes à la frontière entre approche informée et approche aveugle (section II.4) et enfin les méthodes aveugles (section II.5). Ceci nous amène à introduire dans la section II.5.2 l'usage de la factorisation en matrices non-négatives (NMF) pour la transcription automatique de la musique, et à poser dans la section II.6 les problématiques de cette thèse.

II.2 Estimation de fréquences fondamentales multiples

La musique, et en particulier les musiques que nous cherchons à analyser ici, est par essence polyphonique : plusieurs notes sont jouées simultanément. Si l'oreille humaine est capable d'appréhender cette polyphonie avec une certaine performance, en combinant écoute analytique (focalisation sur un instrument en particulier) et écoute synthétique (appréciation du tout), l'estimation automatique de hauteurs multiples n'est pas sans poser un certain nombre de problèmes. En effet, il ne s'agit pas de la simple généralisation du problème d'estimation de hauteur simple : dans le domaine temporel, détecter « plusieurs périodicités » n'a pas de sens, tandis que dans le domaine fréquentiel, une somme de peignes harmoniques peut être associée à plusieurs ensembles de fréquences fondamentales. Ainsi, les difficultés liées aux rapports harmoniques entre fréquences fondamentales se posent de manière encore plus aiguë dans le cas polyphonique : outre la tendance, déjà observée dans le cas monodique, des estimateurs à surestimer ou sous-estimer la fréquence fondamentale d'une ou de plusieurs octaves, le cas polyphonique impose également de pouvoir déterminer si des notes dont les fréquences sont en rapport harmonique sont présentes simultanément. Cette difficulté est d'autant plus problématique que la musique tonale occidentale est justement friande de l'association de telles notes.

II.2.1 État de l'art de l'estimation de hauteurs multiples

Contrairement à l'estimation de fréquence fondamentale simple, le problème des hauteurs simultanées n'a été que peu abordé dans le domaine du traitement de la parole ; il intéresse en revanche beaucoup les spécialistes du signal musical. Deux grandes familles d'approches sont suivies dans la

littérature. Elles sont présentées ici.

II.2.1.1 Estimation itérative

Une première approche de l'estimation de f_0 multiples consiste à considérer que, puisqu'un son comportant plusieurs hauteurs est un mélange additif de sons à hauteurs simples, on doit pouvoir le « déconstruire » en estimant chaque hauteur simple l'une après l'autre et en soustrayant chaque note de l'ensemble lorsqu'elle est détectée. La méthode repose sur l'idée que dans un mélange polyphonique, l'une des notes est prédominante, par exemple suivant un critère énergétique. Si l'on parvient à estimer cette note et à la soustraire du mélange, le résiduel contient théoriquement une note de moins que l'original, et il ne reste plus qu'à itérer le procédé jusqu'à épuiser les hauteurs présentes. C'est le principe des méthodes iteratives [Karjalainen et Tolonen, 1999, Ortiz-Berenguer, 2002, Klapuri, 2003, Klapuri, 2008].

Entrée : Signal de musique polyphonique x
Sortie : Liste de fréquences fondamentales
Initialisation : Résiduel $r = x$
while r contient une note **do**
  Trouver la fréquence fondamentale f_0 prédominante dans r
  Estimer le signal s correspondant à f_0
  $r \leftarrow r - s$, ajouter f_0 à la liste de sortie.
end while

TABLE II.1 – Estimation itérative de fréquences fondamentales multiples.

La sélection de la note prédominante repose souvent sur un critère énergétique, qui correspond parfois à une méthode d'estimation conçue pour le cas monodique, par exemple le maximum du produit spectral. Une fois la fréquence fondamentale prédominante estimée, on doit estimer quelle partie de la note lui correspond et soustraire cette contribution, ce qui va souvent se faire de manière approchée et constitue un problème difficile, notamment en raison du recouvrement spectral entre les différentes notes simultanées. Une solution consiste à exploiter l'information portée par l'enveloppe spectrale de la note, qui, pour la plupart des instruments de musique, a une allure régulière. Ce principe baptisé *spectral smoothness* et introduit dans [Klapuri, 2001b] fait référence dans le domaine.

Dans ce modèle, une hauteur est caractérisée non seulement par l'énergie de chaque partiel pris individuellement mais aussi selon l'énergie relative entre les partiels. Ainsi, en cas de recouvrement spectral, l'amplitude des partiels est déterminée comme une valeur moyenne ou médiane des amplitudes des partiels d'ordre voisin. De cette manière, on évitera par exemple de soustraire l'octave en même temps que la note que l'on vient de détecter, le spectre de la note et de son octave jouées ensemble ayant des pics plus importants une fréquence sur deux. La définition d'un critère d'arrêt (de type énergétique, par exemple le rapport signal à bruit) n'est pas aisée, mais permet une estimation simultanée du degré de polyphonie.

II.2.1.2 Estimation jointe

Une autre approche consiste à chercher d'emblée à identifier toutes les hauteurs présentes en même temps. Cette estimation jointe est traitée par exemple dans [de Cheveigné, 2003, Tolonen et Karjalainen,

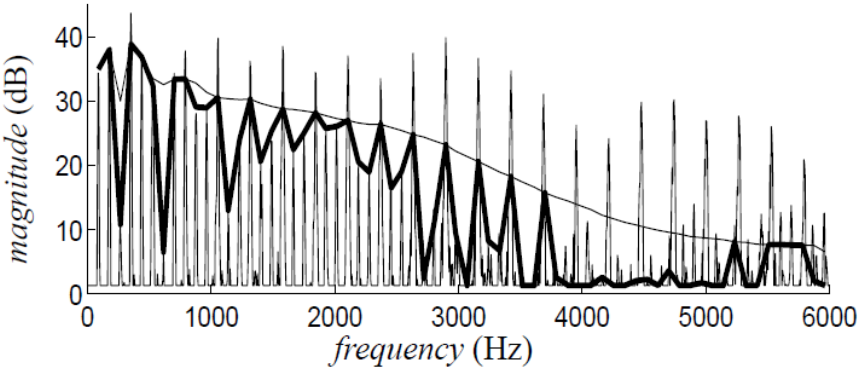


FIGURE II.1 – Principe du *spectral smoothness*. D'après [Klapuri, 2001b].

2000]. La méthode consiste à passer le signal dans des filtres en peigne qui suppriment les partiels répartis suivant la distribution harmonique correspondant à des fréquences fondamentales supposées puis à calculer l'énergie du résiduel ; cette énergie doit être minimale lorsque ces fréquences ont été correctement choisies.

L'estimation conjointe résout le problème de la soustraction imparfaite des notes successives dans la méthode itérative. Cependant, elle pose un problème combinatoire important. Si on se limite simplement à un degré de polyphonie maximal de 4 (jamais plus de 4 notes simultanées), le nombre total de filtres en peigne nécessaires pour estimer les hauteurs jouées sur un piano est $\sum_{p=1}^{4} 88^p$, c'est-à-dire plus de 60 millions ! Cette méthode n'est donc raisonnable que pour des *tessitures et des polyphonies faibles. De plus, elle impose de fixer *a priori* un degré de polyphonie maximum, donc d'estimer par une autre méthode la polyphonie de la pièce analysée. Un alternative à cette explosion combinatoire est l'approche probabiliste [Walmsley *et al.*, 1999, Godsill, 2002, Davy *et al.*, 2006, Emiya *et al.*, 2007], que nous discuterons dans la section suivante.

II.2.2 Du multipitch à la transcription

Même en supposant que nous disposons d'un moyen d'estimer efficacement et précisément un ensemble de fréquences fondamentales dans un extrait quasi-stationnaire de signal, cela ne fait pas une transcription. Deux approches sont possibles pour intégrer l'information temporelle :

1. L'approche par segmentation, qui consiste à pré-découper le signal en segments susceptibles de correspondre à une note, en détectant d'abord les débuts de note (*onsets*), pour analyser chaque segment séparément.
2. L'approche par fusion, qui consiste à analyser le signal trame à trame, puis à apparier ces informations d'une trame à l'autre pour reconstituer une information temporelle.

[Klapuri, 2003] propose une étude de la métrique préalable à l'estimation des fréquences fondamentales. Les attaques sont détectées et le signal segmenté. On considère ensuite que le signal à l'intérieur d'un segment est quasi-stationnaire et que son contenu en terme de hauteurs est constant sur le segment, ce qui permet d'utiliser toute sa durée pour l'estimation des fréquences fondamentales. Cette approche peut convenir à une musique *homorythmique, mais risque de poser problème dans le cas général, la question des recollements entre segments se posant.

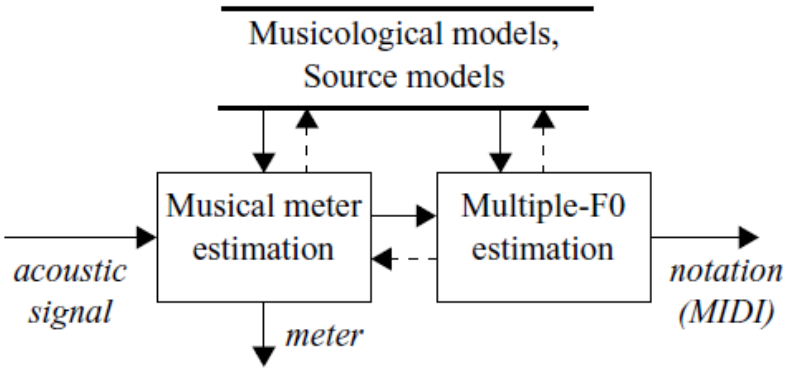


FIGURE II.2 – Modèle de système de transcription par segmentation. D'après [Klapuri, 2003].

À l'inverse, [Ryynänen et Klapuri, 2005] propose un système qui illustre bien l'approche par fusion (voir figure II.3). Il utilise l'estimateur de fréquences fondamentales multiples de [Klapuri, 2005], qui fournit une saillance des fréquences fondamentales multiples estimées dans une trame. Pour chaque trame, un vecteur de caractéristiques (*features*) est calculé à partir des cinq fréquences fondamentales les plus saillantes, des valeurs de leur saillance, et des cinq fréquences fondamentales pour lesquelles la dérivée temporelle de la saillance est maximale (« fréquences prédominantes pour les attaques »). Ce vecteur de caractéristiques alimente ensuite un modèle de Markov caché (HMM) à trois états (attaque, entretien et silence) qui réalise l'intégration temporelle. Les modèles de Markov cachés, introduits par Leonard Baum dans les années 1960, sont sans aucun doute le moyen le plus largement utilisé dans les méthodes par fusion [Raphael, 2002, Ryynänen et Klapuri, 2005, Poliner et Ellis, 2007], avec cependant des choix variables en ce qui concerne le paramétrage, la définition des états et la détermination des probabilités de transition.

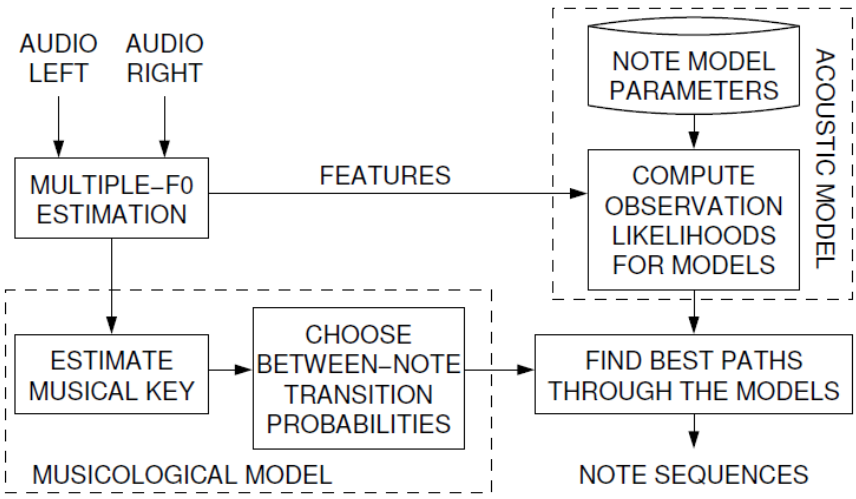


FIGURE II.3 – Modèle de système de transcription par fusion. D'après [Ryynänen et Klapuri, 2005].

Il est possible de combiner ces deux approches, ce que nous appellerons l'approche mixte. Elle consiste à pré-segmenter le signal, les frontières entre segments étant les *onsets* détectés, puis à analyser trame par trame le signal à l'intérieur de ce segment, pour finalement re-fusionner les informations de ces trames. Cette approche, développée par exemple dans [Emiya, 2008], permet d'être plus robuste aux erreurs de détection d'*onsets* ou à certains cas particuliers tels que les notes répétées, tout en réduisant la complexité du modèle HMM, en interdisant un grand nombre de transitions à l'intérieur d'un segment.

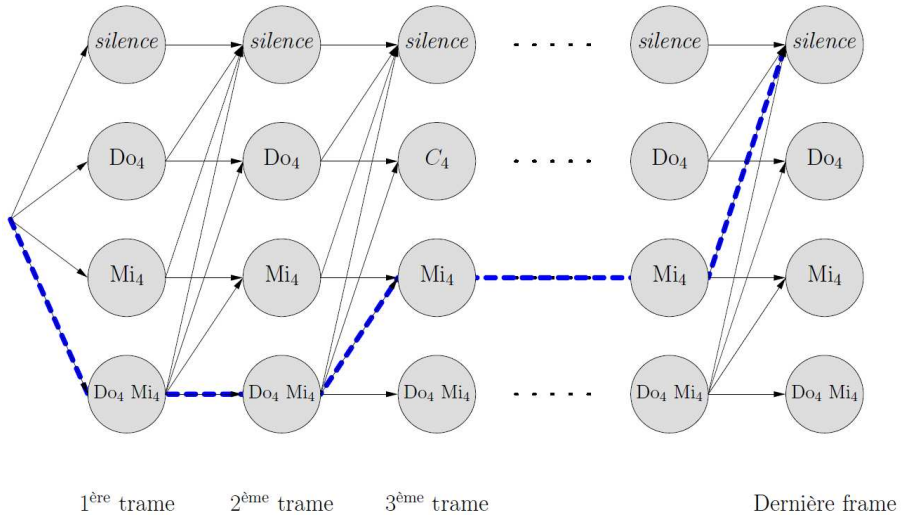


FIGURE II.4 – Modèle de système de transcription mixte. D'après [Emiya, 2008].

II.3 Transcription fondée sur des connaissances

De nombreux systèmes de transcription utilisent des connaissances extérieures au signal analysé, à des degrés et avec des natures diverses. Nous les recensons ici.

II.3.1 Apprentissage hors ligne

Les systèmes de transcription les plus « informés » sont ceux qui font appel à un apprentissage statistique. Des informations issues de bases de données de référence, et extraites préalablement à l'analyse des données à transcrire, sont utilisées pour résoudre le problème.

II.3.1.1 Approches par dictionnaire

Un premier principe d'apprentissage consiste à construire un dictionnaire dont les éléments correspondent en général à des notes, et à chercher ensuite à minimiser une distance donnée entre une combinaison d'éléments du dictionnaire et le morceau à transcrire. C'est la démarche adoptée par exemple par [Rossi, 1998, Ortiz-Berenguer, 2002, Plumbley *et al.*, 2006, Cont *et al.*, 2007]. Dans ces travaux, un dictionnaire de spectres de notes est appris sur des sons de notes isolées. Il est également possible de construire des dictionnaires dans d'autres domaines que le domaine fréquentiel, par

exemples des dictionnaires de formes d'onde [Leveau *et al.*, 2008]. Le signal à analyser est ensuite décomposé suivant ce dictionnaire, à l'aide d'algorithmes dits de « poursuite » minimisant une erreur entre le signal analysé et sa reconstruction.

L'emploi de données extérieures a l'avantage d'apporter une information importante et complexe, qu'on ne pourrait pas atteindre avec un simple modèle, mais présente aussi le risque d'un sur-apprentissage, ou de performances médiocres si les données analysées sont trop différentes des données d'apprentissage. Elle suppose l'additivité de la représentation choisie, ce qui est vérifié par un dictionnaire de formes d'ondes, mais n'est qu'une approximation dans le cas de dictionnaires formés de spectres d'amplitude ou de puissance [1]. De plus cette approche considère que le spectre d'une note tout au long de sa durée se résume à un spectre de référence multiplié par un gain global, ce qui est une approximation (les variations d'amplitude d'un spectre en fonction de la nuance, par exemple, sont non linéaires, et au sein d'une même note les partiels de hautes fréquences s'éteignent souvent plus tôt que ceux proches de la fondamentale).

II.3.1.2 Approches par classification automatique

Pour aller plus loin, plusieurs auteurs proposent d'employer des techniques de classification automatique en sus de l'apprentissage de données. Le système apprend ainsi des informations sur la variabilité des données, en construisant des classes à partir de plusieurs échantillons.

Parmi ces systèmes, l'un des plus performants à notre connaissance est celui exposé dans [Marolt, 2004], qui repose sur un ensemble de réseaux de neurones spécialisés dans des sous-tâches : la détection d'attaques (« *Integrate-and-Fire Neural Network* »), la reconnaissance de notes à partir de fréquences de partiels (« *Time Delay Neural Network* ») et la détection de notes répétées (« *Multilayer Perceptron Neural Network* »). Le schéma de la figure II.5 résume le fonctionnement de ce système.

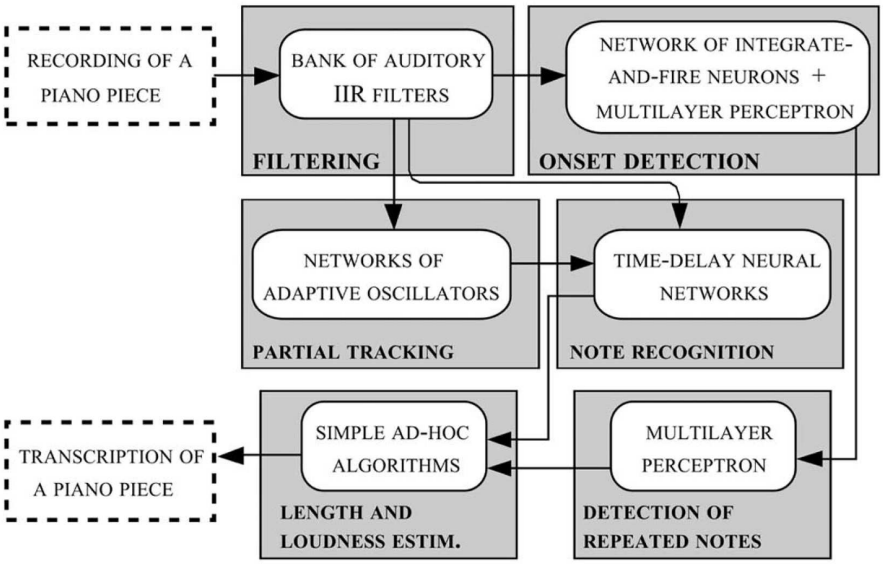


FIGURE II.5 – Modèle de système de transcription par réseaux de neurones. D'après [Marolt, 2004].

1. Ce sera le cas dans le modèle de factorisation en matrices non-négatives que nous aborderons dans la suite de cette thèse.

En utilisant les machines à vecteurs supports (SVM), [Poliner et Ellis, 2007] effectue un apprentissage non plus sur des notes isolées mais sur des mélanges de notes. Ainsi, le système n'apprend pas seulement les caractéristiques de chaque note, mais aussi les conséquences de leur simultanéité telles que le recouvrement spectral. Le système aura par exemple appris à reconnaître une note à la fois lorsqu'elle est isolée et lorsqu'elle se trouve en présence de son octave.

La capacité de généralisation du système dépend du choix de la base d'apprentissage et du paramétrage du système, étapes délicates qui peuvent mener à une dégradation des performances dans le cas d'un sous-apprentissage (base trop petite, manquant de diversité) ou, à l'opposé, dans le cas d'un sur-apprentissage.

Les systèmes de transcription utilisant des HMM pour l'intégration temporelle des informations de fréquences fondamentales et/ou la détection d'attaque et d'extinction ont également recours à des bases de données extérieures pour l'apprentissage des probabilités de transition entre états. Suivant la nature du vecteur d'observation, cet apprentissage peut être réalisé sur des références MIDI [Emiya, 2008] ou sur des bases de notes isolées ou d'accords [Ryynänen et Klapuri, 2005]. Lorsque les états décrivent des mélanges de notes, la combinatoire peut devenir rapidement rédhibitoire. C'est la raison pour laquelle [Ryynänen et Klapuri, 2008], par exemple, propose de transcrire les accords d'accompagnement non pas *in extenso* mais en considérant les chroma (pitch modulo 12), et en transcrivant une tablature plutôt qu'un fichier MIDI.

Dans [Ryynänen et Klapuri, 2005], le vecteur d'observation en entrée du HMM est composé de trois paramètres : l'écart entre la fréquence fondamentale trouvée et la fréquence théorique, la saillance associée et l'intensité éventuelle d'une attaque dont la fréquence fondamentale est proche de celle de la note. Le HMM est illustré sur la figure II.3. Les trois états utilisés sont associés à l'attaque, à l'entretien et au bruit de fond, avec des contraintes sur les transitions possibles entre états. La vraisemblance des observations est modélisée par un GMM (*Gaussian Mixture Model*, modèle de mélange de gaussiennes). L'avantage de cette approche réside dans la simplification qu'elle propose. Grâce à cette division en deux tâches distinctes, la décision s'appuie sur un nombre restreint de paramètres, et sur un modèle de note à trois états : l'attaque caractérisée par une forte variation de saillance, le sustain pour lequel la saillance est élevée, stable et l'écart de fréquence fondamentale faible, et le silence dans les autres cas.

II.3.2 Utilisation de connaissances musicologiques

Des informations de plus haut-niveau que celles concernant la note de musique peuvent être prises en compte, soit pour être transcrites en tant que telles, soit pour améliorer les performances en apportant des connaissances guidant les choix et l'analyse. Par exemple, l'estimation et l'utilisation de la *tonalité du morceau, ou d'éléments de « grammaire » musicale (théorie de l'harmonie classique), peuvent aider l'estimation de fréquences fondamentales multiples. Par exemple, des notes consonantes sont davantage susceptibles de se produire simultanément, tandis que certains mouvements mélodiques sont prescrits par l'harmonie (« la sensible monte à la tonique »). Par exemple, [Kashino *et al.*, 1998] propose d'utiliser une connaissance a priori sur les accords et les transitions entre eux. [Ryynänen et Klapuri, 2005] utilise également un modèle musicologique, issu de [Viitaniemi *et al.*, 2003], qui consiste à pondérer les probabilités de transition entre notes sachant la tonalité du morceau.

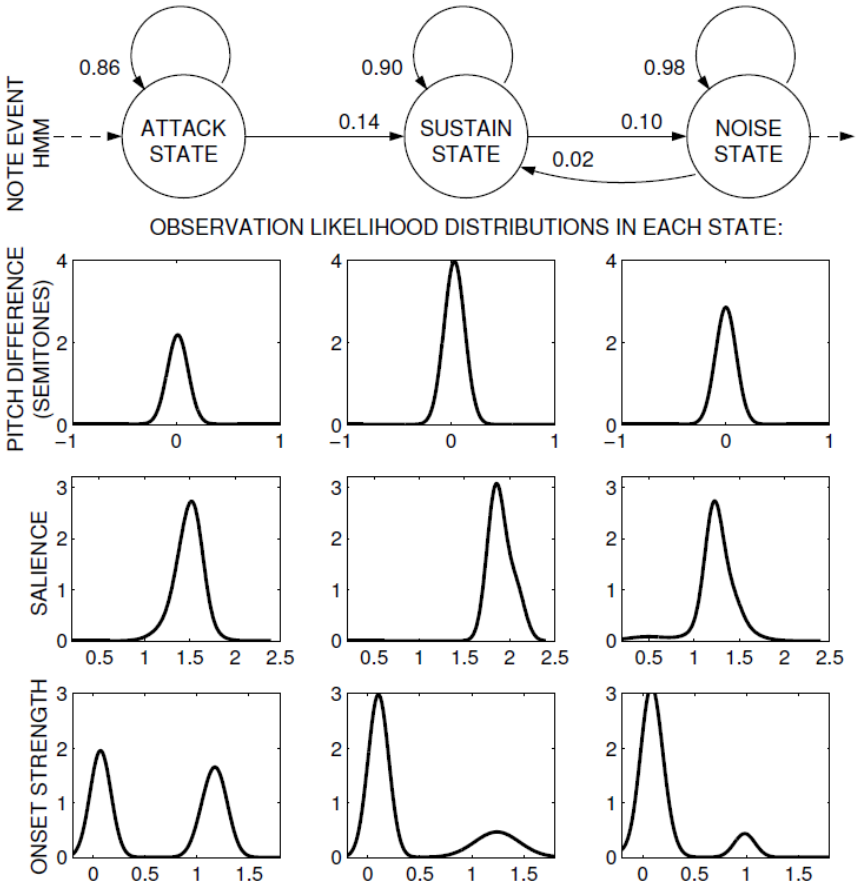


FIGURE II.6 – Modèle de Markov caché pour le suivi des f_0. D'après [Ryynänen et Klapuri, 2005].

II.4 Méthodes bayésiennes

À la charnière entre méthodes informées et aveugles, nous trouvons une autre grande famille d'approches, les méthodes probabilistes bayésiennes. Il s'agit d'un autre moyen d'inclure une forme de connaissance dans le système, et ce sans nécessairement lui apporter de données extérieures. Il consiste à poser un modèle de signal, plus ou moins perfectionné, sous forme d'une loi de probabilité a priori, et d'en estimer les paramètres. Évidemment, si les données ne suivent pas le modèle suffisamment bien, cette approche risque d'échouer. Adaptatives (les paramètres sont appris sur le signal à analyser), elles sont plus ou moins informées suivant le raffinement des modèles.

Les méthodes bayésiennes offrent au problème un cadre statistique théorique et permettent d'élargir les capacités de modélisation des systèmes. Dans ces approches, le signal est décrit de manière unifiée et systématique par des variables aléatoires représentant chaque élément à modéliser : distribution des fréquences, amplitudes, bruit, mais aussi polyphonie, évolution des partiels, notes, instruments, tonalité, etc. Ces variables aléatoires interdépendantes forment ainsi un réseau, caractérisé par le choix de leurs distributions probabilistes. Le problème est alors résolu *via* la théorie de l'inférence bayésienne, en estimant les modèles et paramètres optimaux à partir du signal observé. Deux quantités probabilistes interviennent : les lois a priori, qui traduisent la connaissance sur les paramètres que l'on choisit d'injecter dans le système, et la vraisemblance, qui établit le rapport entre les paramètres et le signal. L'approche bayésienne offre l'avantage de proposer un cadre statistique rigoureux pour modéliser l'ensemble des paramètres et des variables interdépendantes en jeu dans le problème de transcription automatique ; des méthodes d'estimation extrêmement puissantes telles que les méthodes de Monte-Carlo (MCMC) sont disponibles pour réaliser l'inférence des modèles.

Les premiers modèles bayésiens pour le traitement de la musique sont développés par Kashino dès 1995. Le système proposé traite l'information à trois niveaux différents (partiels simples, notes, accords) dans une architecture en tableau noir, et les sources de connaissances modifiant les hypothèses sont des classificateurs de Bayes [Kashino *et al.*, 1998]. [Sterian *et al.*, 1999] propose d'effectuer le groupage des partiels en une note par des méthodes bayésiennes s'inspirant des règles de groupage perceptuel de [Bregman, 1994] (début commun, harmonicité). Les partiels sont détectés par un filtrage de Kalman sur une transformée modale issue de [Pielemeier *et al.*, 1996].

Ces exemples sont les prémices d'une utilisation bien plus poussée de modèles bayésiens. Le véritable précurseur de l'utilisation de ces techniques en transcription automatique de la musique est plutôt [Walmsley *et al.*, 1999]. Dans ces travaux, le signal est considéré, trame par trame, comme la somme de signaux de notes et d'un bruit blanc. Chaque signal de note est lui-même une somme de sinusoïdes dont les fréquences suivent une distribution harmonique. La fréquence fondamentale, les amplitudes, le nombre de composantes d'une note sont considérées comme des variables aléatoires. Les amplitudes sont supposées iid (indépendantes identiquement distribuées) et de loi uniforme. Une contrainte de continuité des fréquences d'une trame à l'autre est introduite dans la loi suivie par la fréquence fondamentale, et l'ensemble des paramètres est estimé par une méthode de type MCMC.

Ces travaux sont repris et prolongés dans [Godsill, 2002], avec une quantité notable de raffinements. Des contraintes sont introduites sur les enveloppes spectrales des notes et du bruit, et l'inharmonicité est autorisée par la modélisation de l'écart entre la fréquence réelle et la fréquence théorique d'un partiel (en principe multiple de la fondamentale) sous la forme d'une variable gaussienne centrée. [Davy *et al.*, 2006] reprend également ces idées dans un modèle extrêmement complet, fondé sur une décomposition en atomes de Gabor harmoniques, avec une structure hiérarchique de connaissances a priori (certains paramètres des lois a priori des observations sont eux-même considérés comme des variables aléatoires sur lesquels on pose des lois a priori).

[Cemgil, 2004, Cemgil *et al.*, 2006] sont d'autres exemples de modélisation bayésienne. La tâche est présentée comme le suivi d'un ensemble d'oscillateurs. Les notes sont supposées être parfaitement harmoniques, et subir un amortissement temporel multiplicatif qui modélise la décroissance exponentielle de l'enveloppe spectrale.

Un intérêt des approches bayésiennes décrites ci-dessus réside dans le modèle additif de notes qu'elles proposent pour les mélanges polyphoniques. Contrairement à d'autres méthodes faisant des approximations d'additivité des spectres d'amplitude ou de puissance, ou qui simplement ne tiennent pas compte de l'information de phase, les approches bayésiennes, temporelles pour la plupart, permettent une estimation des amplitudes et des phases des composantes sans faire d'approximation.

II.5 Transcription aveugle ou semi-aveugle

II.5.1 Panorama

En général, les méthodes avec apprentissage préalable ont une capacité de généralisation limitée : les résultats se dégradent lorsque le signal à traiter présente des différences importantes avec les données apprises. La variabilité des timbres, la réverbération ou les effets de mixage peuvent par exemple être à l'origine de cette dégradation puisque la base d'apprentissage ne peut refléter tous les cas de signaux rencontrés ensuite. Plusieurs approches récentes proposent d'effectuer un apprentissage directement sur le signal à analyser. Nous qualifions ces approches de « méthodes aveugles », par analogie avec la séparation de sources (BSS pour *Blind Source Separation*) car elles ne font appel ni à des données extérieures, ni à des modèles sophistiqués contraignants.

C'est le cas par exemple de [Bello *et al.*, 2006], qui propose l'apprentissage adaptatif d'un dictionnaire de formes d'ondes de notes de piano. Dans une première phase d'analyse du signal, les notes isolées sont repérées à l'aide d'une méthode spectrale d'estimation. Les formes d'onde des notes sélectionnées servent alors à construire un dictionnaire de formes d'onde. Les signaux des notes manquantes sont obtenues par interpolation de ceux des notes existantes. La transcription est finalement obtenue en appliquant la méthode des moindres carrés afin d'identifier l'apparition des formes d'onde du dictionnaire dans le signal.

Les méthodes inspirées de la BSS entrent également dans le cadre des systèmes aveugles. Nous pouvons par exemple citer l'analyse en composantes indépendantes et les décompositions parcimonieuses [Plumbley *et al.*, 2006]. Enfin, [Duan *et al.*, 2008] propose une méthode d'apprentissage adaptatif des enveloppes spectrales, dans une approche s'appuyant sur une distance entre peignes harmoniques et modèles gaussiens associés.

Parmi les techniques aveugles de décomposition de signaux, la factorisation en matrices nonnégatives (NMF), introduite par [Lee et Seung, 1999], a pris ces dernières années un envol spectaculaire dans le monde du traitement du signal audio. Totalement aveugle (la seule hypothèse étant la nonnégativité de la représentation temps-fréquence), elle offre un cadre simple pour calculer simultanément un dictionnaire de notes et les enveloppes temporelles associées, et estimer dans le même temps le degré de polyphonie et les notes en présence sans recourir à une estimation multipitch. La NMF est l'objet central de ce mémoire. Nous motivons ce choix au paragraphe suivant.

II.5.2 Méthodes basées sur la NMF

Soit $\mathbf{V}$ une matrice de dimensions $F \times N$ à coefficients réels positifs ou nuls. La NMF est la détermination d'une factorisation approchée :

$$\mathbf{V} \approx \mathbf{WH} = \hat{\mathbf{V}} \tag{II.1}$$

où $\mathbf{W}$ et $\mathbf{H}$ sont des matrices de dimensions respectives $F \times K$ et $K \times N$ dont tous les coefficients sont des réels positifs ou nuls, et où l'opérateur $\approx$ désigne une « approximation » à définir. L'ordre du modèle, K, est habituellement choisi tel que $FK + KN \ll FN$, ce qui fait de la NMF une technique de réduction de la dimensionnalité. Dans la suite de ce document, f désigne l'indice de fréquence et n l'indice de trame temporelle ; la matrice des observations $\mathbf{V}$ étant, par le fait, une représentation temps-fréquence du signal.

Pour motiver l'usage de la NMF comme outil de production d'une représentation sémantiquement pertinente de données à valeurs positives, [Lee et Seung, 1999] recourt à des arguments physiologiques et cognitifs : le cerveau humain représenterait des objets complexes par la somme de ses parties. La NMF, contrairement à d'autres représentations de même type mais n'exploitant pas la non-négativité (telles que la SVD, la PCA, l'ICA), a de bonnes chances de produire une telle représentation [2], grâce à cette contrainte : d'une part, les composantes extraites ($\mathbf{W}$) sont directement interprétables dans le même domaine que les données analysées ; d'autre part, la non-négativité des coefficients de la décomposition ($\mathbf{H}$) interdit toute « soustraction » d'une composante à une autre, rendant la représentation purement additive.

En transcription, où nous cherchons à extraire des notes de musique, l'objet à représenter est un ensemble de notes, et les « parties » qui composeraient la représentation souhaitée sont donc ces notes. Il semble en effet assez naturel de définir, par exemple, un accord comme la somme des notes qui le composent. Évidemment, on est ici dans le domaine symbolique, et la NMF ne pourra s'opérer que sur une représentation quantitative du signal. Le spectrogramme d'amplitude ou de puissance, obtenu à partir de la Transformée de Fourier à Court-Terme (TFCT), est le premier choix qui vient à l'esprit : il s'exprime comme une matrice à coefficients positifs ou nuls, et le spectre d'un accord musical n'est pas très différent de la somme des spectres des notes isolées qui le composent.

Pour visualiser l'intérêt de la NMF du spectrogramme d'amplitude, prenons un exemple très simple, à savoir une mélodie de trois notes bien connue, représentée figure II.7, jouée par un piano de synthèse, et supposons que nous disposons d'ores et déjà d'une méthode de calcul des facteurs $\mathbf{W}$ et $\mathbf{H}$. Nous cherchons trois parties dans notre objet musical, nous fixerons donc $K = 3$.

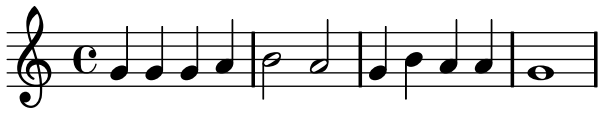


FIGURE II.7 – Un exemple monodique simple : *Au clair de la Lune.*

La factorisation en matrices non-négatives du spectrogramme de cette mélodie est illustrée sur la figure II.8, la disposition étant celle d'un produit matriciel.

2. On rencontre fréquemment dans la littérature l'expression « parts-based representations ».

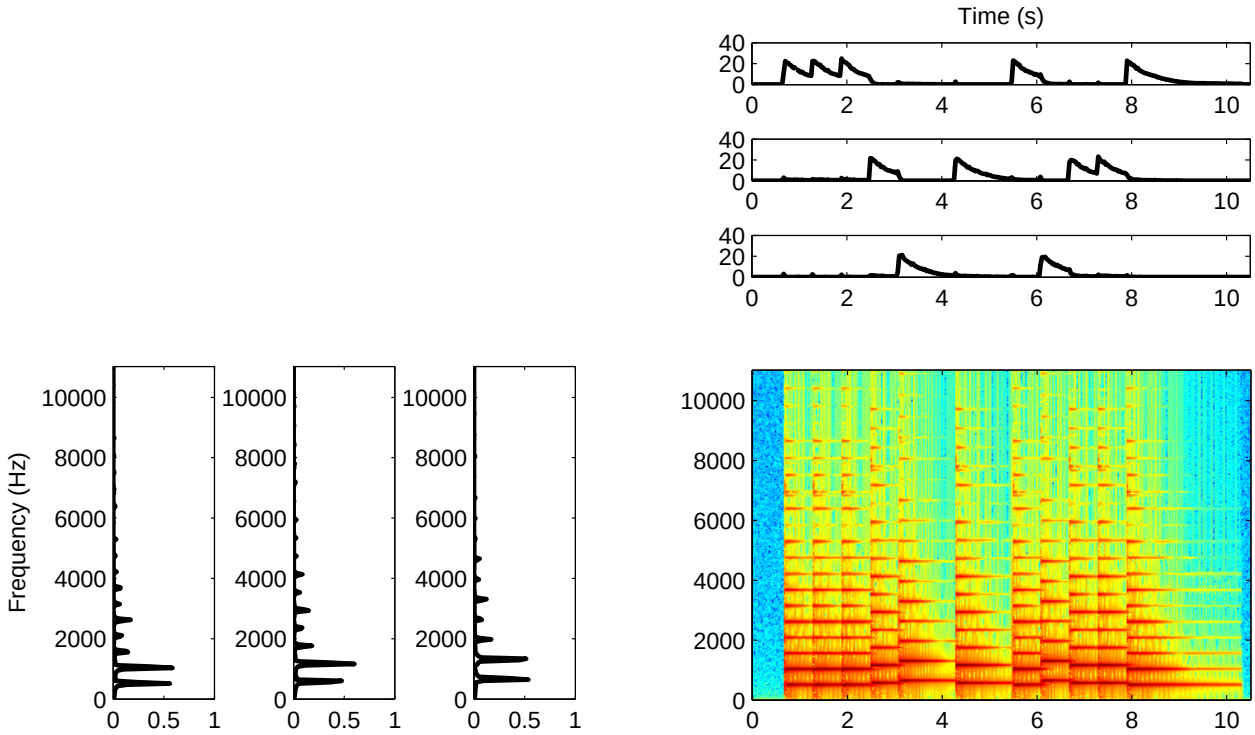


FIGURE II.8 – Factorisation du spectrogramme d'*Au clair de la Lune*.

Dans ce cas très simple mais illustratif, on remarque que la NMF standard a été capable d'identifier les trois événements élémentaires suffisants pour représenter l'ensemble du signal. Les colonnes de $\mathbf{W}$, représentées en bas à gauche du produit, sont les spectres des trois notes en présence (*sol*, *la*, *si*), tandis que les lignes de $\mathbf{H}$, en haut à droite, ont capturé les enveloppes ou « activations » de ces notes dans le temps. La représentation obtenue reflète donc une certaine sémantique de la musique originale, ce qu'on appellera une « représentation mi-niveau ».

Ce cas est évidemment trivial. Prenons un exemple à peine plus compliqué, mais cette fois polyphonique, dont la partition est présentée figure II.9. Il comporte quatre notes, nous choisissons donc de fixer $K = 4$.

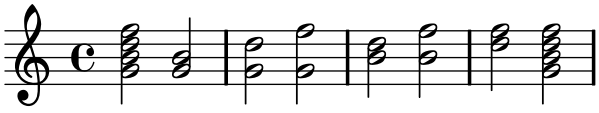


FIGURE II.9 – Un exemple polyphonique simple.

La NMF va ici prendre tout son sens en terme de méthode de réduction de rang. En effet huit événements sonores se produisent, mais on ne dispose que de quatre mots pour représenter la séquence. Il est donc à espérer que la NMF va être capable d'identifier ces quatre « mots » élémentaires et sémantiquement pertinents pour représenter l'enchaînement des huit événements qui les combinent. C'est effectivement ce qui se produit, comme on peut le voir sur la figure II.10, qui représente $\mathbf{W}^T$ et $\mathbf{H}$ respectivement à gauche et à droite de la figure.

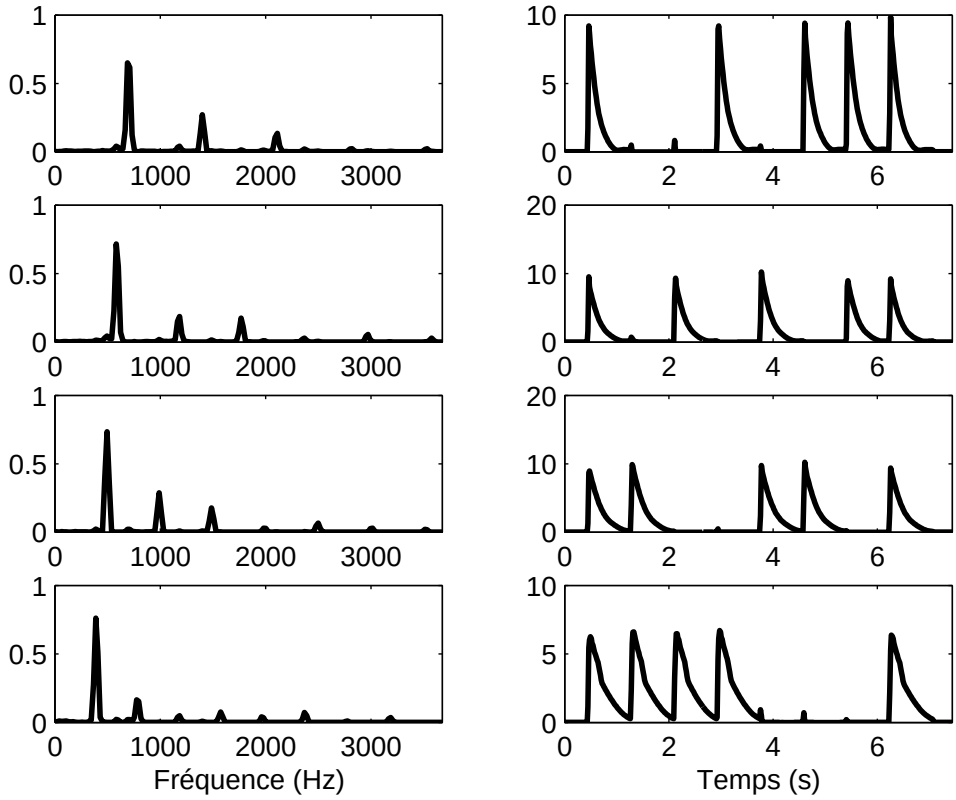


FIGURE II.10 – Factorisation de la séquence de la figure II.9

Il est à noter que la séquence ne contient aucune note isolée, c'est-à-dire que la NMF se révèle capable d'extraire les notes sans les avoir « entendues seules » à un moment ou à un autre de l'extrait analysé.

Évidemment, si l'on pratique une NMF seulement sur les quatre accords centraux (en rouge sur la figure II.11), on se retrouve alors dans une situation où le nombre d'événements et la taille du dictionnaire sont égaux. On peut alors voir sur la figure II.12 que dans ce cas, les « parties » extraites sont directement les accords. En effet, la redondance du morceau analysé est cruciale pour que la réduction de rang soit opérante ; faute d'une redondance suffisante, il est illusoire d'espérer séparer les notes qui composent le morceau.

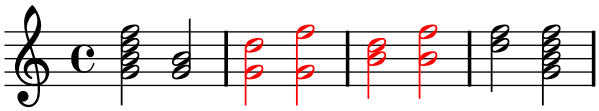


FIGURE II.11 – Sous-séquence polyphonique.

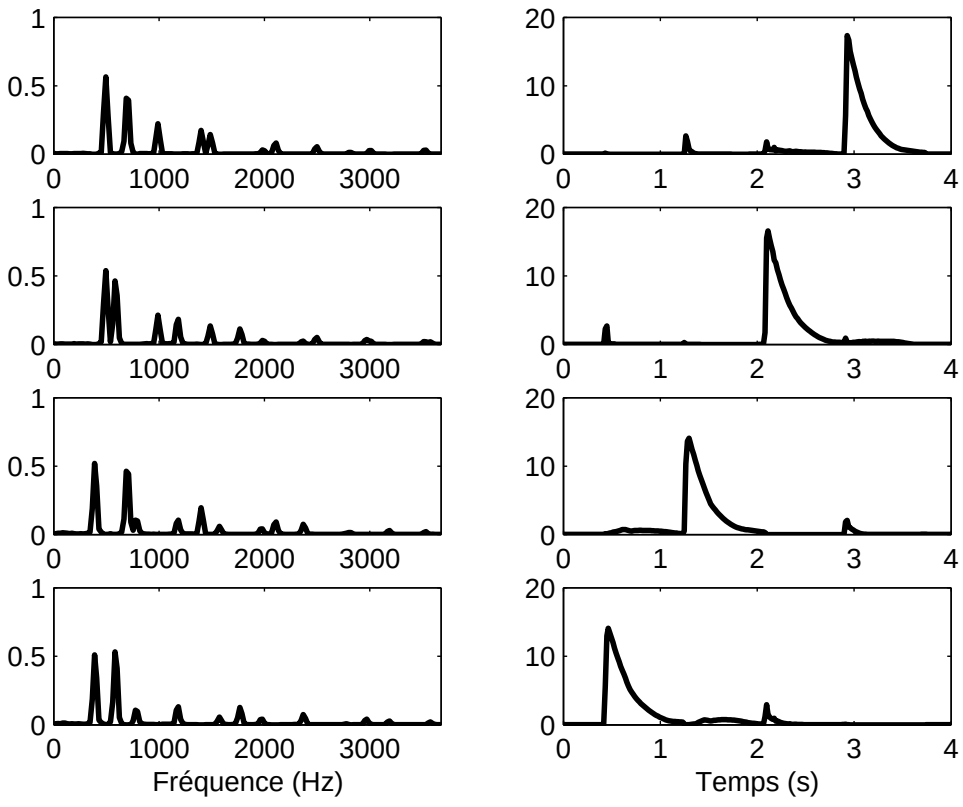


FIGURE II.12 – Factorisation de la séquence de la figure II.11.

Un autre aspect est à prendre en compte : bien que la grande partie de l'objet qui nous intéresse, à savoir la note de musique, possède un spectre quasi-stationnaire de nature harmonique, d'autres événements liés à la note et à l'enregistrement sont susceptibles de perturber cette représentation :

le transitoire (ou attaque), et l'existence de bruits résiduels (erreur de représentation). Il pourra être nécessaire de « garder de la place » dans le dictionnaire pour représenter ces parties, et donc de choisir un ordre K supérieur au nombre de notes attendues.

II.6 Problématique de la thèse

En comparaison d'autres systèmes de transcription plus complexes et plus informés, la NMF apparaît comme un moyen simple d'obtenir une représentation mi-niveau sémantique de la musique, prometteuse pour la transcription, sans information extérieure, exploitant la redondance du signal et traitant simultanément les dimensions temporelle et fréquentielle du signal. Son fonctionnement revêt cependant un caractère de « boîte noire » et son efficacité pour produire une réelle transcription reste à établir. À la lumière de cet état de l'art introductif, nous proposons dans cette thèse d'examiner les questions suivantes.

Comment fonctionne la NMF, et pourquoi fonctionne-t-elle ? Quelles sont ses propriétés ?

Nous examinerons en particulier ses propriétés théoriques, ainsi que les fonctions de coût sous-jacentes à sa résolution et leur pertinence pour la représentation des signaux audio.

Les algorithmes de NMF sont-ils efficaces ? Peut-on les améliorer ?

De nombreux algorithmes de résolution du problème de NMF sont disponibles dans la littérature. Leurs propriétés, de convergence en particulier, sont diverses. L'initialisation et l'évitement des minima locaux font partie des questions à élucider.

L'approche totalement aveugle est-elle suffisante ? Quel est le degré minimum de connaissances à injecter dans le système ? Quelles connaissances et sous quelle forme ?

La plupart des systèmes de transcription de l'état de l'art sont, d'une manière ou d'une autre, informés, c'est-à-dire qu'ils utilisent des connaissances extérieures au signal pour l'analyser. Ce n'est pas le cas de la NMF standard, qui n'impose que la non-négativité des coefficients. Nous verrons dans quelle mesure il est possible et souhaitable de lui apporter des connaissances supplémentaires, notamment sous la forme de contraintes.

Comment convertir la factorisation en réelle transcription ? S'agit-il d'une approche efficace ?

Si la pertinence de la représentation NMF saute aux yeux sur un exemple simple, son efficacité pour produire une réelle transcription de la musique une fois intégrée à un système complet reste à établir. Nous proposerons une architecture de système de transcription basée sur les NMF et confronterons les mises en œuvre qui en découlent à des algorithmes de l'état de l'art en conditions réelles de transcription.

Deuxième partie

Approche déterministe

Chapitre III

État de l'art de la NMF non contrainte

Résumé

Où l'on introduit le problème standard de factorisation en matrices à coefficients positifs (NMF), en abordant la question de l'existence et de l'unicité des solutions à ce problème, avant de dresser l'état de l'art des fonctions de coût sous-jacentes et des algorithmes usuels de la littérature s'y intéressant, en gardant à l'esprit les questions pratiques de convergence, d'initialisation et de minima locaux.

III.1 Introduction

Le problème standard de factorisation en matrices à coefficients positifs, sous la forme la plus générique possible, s'exprime comme suit : étant donnée une matrice $\mathbf{V}$ de dimensions $F \times N$ à coefficients réels positifs ou nuls, la NMF est la détermination d'une factorisation approchée :

$$\mathbf{V} \approx \mathbf{WH} = \hat{\mathbf{V}} \tag{III.1}$$

où $\mathbf{W}$ et $\mathbf{H}$ sont des matrices de dimensions respectives $F \times K$ et $K \times N$ dont tous les coefficients sont des réels positifs ou nuls, et où l'opérateur $\approx$ désigne une « approximation » à définir. L'ordre du modèle, K, est habituellement choisi tel que $FK + KN \ll FN$, ce qui fait de la NMF une technique de réduction de la dimensionnalité.

Si l'on attribue plus ou moins de fait la « paternité » de la NMF à [Lee et Seung, 1999], il faut cependant remarquer que, cinq ans plus tôt, [Paatero et Tapper, 1994] pose et résout le même problème. La gloire lui échappera pour une dénomination malheureuse : la « factorisation en matrices positives » (PMF, ou *Positive Matrix Factorization*) contient une ambiguïté sur la notion de matrice positive (est-elle *à coefficients positifs*, ou *semi-définie positive* c'est-à-dire à valeurs propres réelles positives ?). Malgré cette maladresse, on trouve de nombreuses applications de la PMF entre 1994 et 1999, principalement dans des domaines liés à la chimie, l'étude des pollutions et les sciences de l'environnement [Junnto et Paatero, 1994, Anttila *et al.*, 1995, Polissar *et al.*, 1998, Paterson *et al.*, 1999].

Néanmoins, c'est véritablement [Lee et Seung, 1999] qui fait date et crée l'enthousiasme autour de la NMF, qui prend alors son nom définitif et commence à être appliquée dans de très nombreux domaines [1]. Nous illustrerons la variété des domaines d'applications par quelques exemples, sans avoir aucunement l'ambition d'être exhaustive :

- Traitement de l'image : représentation d'images de visages [Lee et Seung, 1999, Li *et al.*, 2001, Wang *et al.*, 2004], classification d'images [Guillamet *et al.*, 2001]
- Traitement du texte : surveillance de messages électroniques [Berry et Browne, 2005], classification de documents [Ding *et al.*, 2008], extraction de caractéristiques sémantiques dans des textes [Lee et Seung, 1999]
- Économie : diversification de portefeuilles d'actions [Drakakis *et al.*, 2008]
- Biologie : clustering [2] de gènes impliqués dans le cancer [Liu et Yuan, 2008], détection de l'activité neuronale pour les interfaces cerveau-machine [Kim *et al.*, 2005]
- Gastronomie : clustering de scotch whiskeys [Young *et al.*, 2006]

À notre connaissance, [Smaragdis et Brown, 2003] est le premier travail proposant d'appliquer la NMF à des signaux audionumériques et plus particulièrement à la transcription de musique polyphonique. Dans les applications audionumériques, f désigne la plupart du temps l'indice de fréquence et n l'indice de trame temporelle ; la matrice des observations $\mathbf{V}$ étant, par le fait, une représentation temps-fréquence du signal.

Ce chapitre dresse l'état de l'art de l'approche classique (déterministe) du problème de NMF. Nous commencerons par examiner dans la section III.2 la question de l'existence et de l'unicité des solutions

1. Pour l'anecdote, les sites bien connus `http://citeseerx.ist.psu.edu` et `http://scholar.google.com` reportent respectivement 415 et 1405 citations de cet article à la date où nous écrivons.

2. Bien que l'Office Québécois de la Langue Française recommande l'usage de la traduction « groupage », nous avons préféré dans ce document conserver l'usage communément répandu du terme anglais.

du problème (III.1). Dans l'approche classique de la NMF, l'approximation est définie à partir d'une fonction de coût, que l'on va chercher à minimiser. Les fonctions de coût usuelles et leurs propriétés sont présentées dans la section III.3. La section III.4 présente les algorithmes de la littérature minimisant (empiriquement ou formellement) ces fonctions de coût. Leurs propriétés formelles de convergence sont examinées dans la section III.5. Enfin, en lien avec les questions d'unicité et de convergence précédentes, la section III.6 dresse un bref état de l'art des solutions proposées pour l'initialisation des algorithmes de NMF.

III.2 Existence et unicité des solutions

III.2.1 Existence des solutions

Sous la forme générique (III.1), l'existence d'une solution va évidemment dépendre de la façon dont on va définir l'approximation. En pratique, nous verrons ultérieurement que la solution approchée du problème (III.1) est définie comme la minimisation d'une fonction de coût (voir section III.3). Il suffit que cette fonction, définie sur $\mathbb{R}_+^2$ soit continue et tende vers l'infini lorsque ses variables tendent vers l'infini pour que l'existence d'une solution soit garantie. La question de l'existence n'est par conséquent pas une préoccupation de la littérature. Cependant, on gardera à l'esprit que l'existence d'une solution au problème mathématique ne garantit aucunement la pertinence de cette solution par rapport à la sémantique des données et à l'application.

III.2.2 Degrés d'invariance

Compte-tenu de la formulation du problème de NMF, il est clairement à craindre l'existence de matrices carrées inversibles $\mathbf{T}$ telles que les couples $(\mathbf{WT}, \mathbf{T}^{-1}\mathbf{H})$ soient également solutions du problème, puisque $\mathbf{WH} = (\mathbf{WT})(\mathbf{T}^{-1}\mathbf{H})$ et que le coût de reconstruction ne dépend que du produit $\mathbf{WH}$. Étant donné une solution $(\mathbf{W}_0, \mathbf{H}_0)$, le couple $(\mathbf{W}_0\mathbf{T}, \mathbf{T}^{-1}\mathbf{H}_0)$ est solution du problème si et seulement si (ssi) les deux matrices $\mathbf{W}_0\mathbf{T}$ et $\mathbf{T}^{-1}\mathbf{H}_0$ sont à coefficients positifs ou nuls. On peut exhiber au moins deux types de cas où cela est possible.

III.2.2.1 Invariances triviales

Si l'on impose que $\mathbf{T}$ et son inverse $\mathbf{T}^{-1}$ soient à coefficients positifs ou nuls, les produits $\mathbf{WT}$ et $\mathbf{T}^{-1}\mathbf{H}$ le sont également et nous sommes bien en présence d'une nouvelle solution au problème initial. En l'occurrence, il est aisé de prouver qu'une telle matrice $\mathbf{T}$ est nécessairement le produit d'une matrice de permutation et d'une matrice diagonale à coefficients positifs [Minc, 1988]. La matrice de permutation introduit K degrés d'invariance, mais ne modifie pas réellement la solution ; du reste, l'unicité d'une quantité est souvent définie à une permutation près. En ce qui concerne les facteurs d'échelle (matrice diagonale), la question peut être résolue en imposant une normalisation sur l'un des facteurs $\mathbf{W}$ ou $\mathbf{H}$ (en pratique, on choisira souvent de normaliser les colonnes de $\mathbf{W}$ en norme L^2). Ces invariances ne sont donc pas un réel obstacle à une éventuelle unicité de la solution, que l'on définira à une permutation et un changement d'échelle près.

III.2.2.2 Invariances locales

Le produit $(\mathbf{W}_0\mathbf{H}_0)$ peut également rester invariant sur des points au voisinage du couple qui le réalise. Supposons que $\mathbf{W}_0$ et $\mathbf{H}_0$ sont à coefficients strictements positifs. Étant donné une matrice carrée $\mathbf{U}$, pas nécessairement à coefficients positifs, on peut trouver ε suffisamment petit tel que $(\mathbf{I}+\varepsilon\mathbf{U})$ est inversible. On peut faire le développement limité : $(\mathbf{I}+\varepsilon\mathbf{U})^{-1} \approx \mathbf{I}-\varepsilon\mathbf{U}$. Les matrices $(\mathbf{I}+\varepsilon\mathbf{U})$ et $(\mathbf{I}+\varepsilon\mathbf{U})^{-1}$ réalisent des transformations locales autour de $(\mathbf{W}_0\mathbf{H}_0)$; pourvu que ce point ne soit pas situé sur les bords du quadrant positif, et que ε soit choisi suffisamment petit, les points $(\mathbf{W}_0(\mathbf{I}+\varepsilon\mathbf{U}))$ et $((\mathbf{I}+\varepsilon\mathbf{U})^{-1}\mathbf{H}_0)$ restent dans ce quadrant. Ainsi, nous obtenons une nouvelle solution au problème de NMF, le couple $(\mathbf{W}_0(\mathbf{I}+\varepsilon\mathbf{U}), (\mathbf{I}+\varepsilon\mathbf{U})^{-1}\mathbf{H}_0)$.

Ces invariances introduisent un nombre maximum de K^2 degrés d'invariance (il peut être inférieur, si $\mathbf{W}_0$ et $\mathbf{H}_0$ contiennent des coefficients nuls). Notons par ailleurs que seule la non-négativité des produits $(\mathbf{WT}, \mathbf{T}^{-1}\mathbf{H})$ nous importe. Ainsi, dans ce cas d'invariances locales, la matrice $\mathbf{T} = (\mathbf{I}+\varepsilon\mathbf{U})$ peut comporter des coefficients négatifs, ce qui explique qu'il existe d'autres invariances que les invariances triviales précédemment évoquées.

III.2.3 Conditions d'unicité

L'existence de ces degrés d'invariance réduit d'emblée la possibilité de définir un cadre d'unicité de la solution au problème de NMF, notamment les invariances locales. Toutefois, on pourrait imaginer définir l'unicité à une permutation, un changement d'échelle ou une transformation locale près. Nous allons voir que ce n'est pas suffisant, mais qu'il est toutefois possible de considérer des cas où la solution sera unique.

III.2.3.1 Non-unicité du cas général

Nous n'avons considéré que des paires de solutions s'exprimant l'une par rapport à l'autre via une transformation linéaire : $(\mathbf{W}_0, \mathbf{H}_0)$ et $(\mathbf{W}_0\mathbf{T}, \mathbf{T}^{-1}\mathbf{H}_0)$. En réalité, rien n'impose que deux solutions réalisant le même produit $\mathbf{WH}$ soient reliées de la sorte. Il suffit d'un contre-exemple, tiré de [Jeter et Pye, 1982] pour s'en convaincre :

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \mathbf{V}\,\mathbf{I} = \mathbf{I}\,\mathbf{V} \tag{III.2}$$

où $\mathbf{I}$ désigne la matrice identité. La matrice $\mathbf{V}$ est de rang 3. On peut choisir comme factorisations $\mathbf{W}_0 = \mathbf{V}$ et $\mathbf{H}_0 = \mathbf{I}$, ou $\mathbf{W}_1 = \mathbf{I}$ et $\mathbf{H}_1 = \mathbf{V}$. Il n'existe aucune matrice inversible $\mathbf{T}$ telle que $\mathbf{W}_0\mathbf{T} = \mathbf{W}_1$, pour des raisons de rang. Cependant, si l'on choisit $K = rg(V)$, on peut montrer que de tels contre-exemples sont impossibles. Dans ce cas, toutes les solutions sont reliées les unes aux autres par des transformations linéaires [Thomas, 1974].

III.2.3.2 Solution exacte

[Donoho et Stodden, 2003] aborde la question de l'unicité. Il exhibe trois conditions suffisantes à l'unicité de la solution et montre un exemple expérimental dans lequel la solution rendue par l'algorithme de NMF de [Lee et Seung, 2001] est bien la solution unique, dans le cas où il existe une

solution exacte $\mathbf{V} = \mathbf{WH}$. Sa démarche est la suivante : il se donne une famille générative (en l'occurrence d'images), génère des mélanges positifs de ces familles, et applique l'algorithme de NMF sur ces mélanges pour les comparer avec la solution recherchée, c'est-à-dire la famille générative de départ.

Étant donnée la famille $(\psi_k)_{1 \leq k \leq K}$, avec $\psi_k \in \mathbb{R}^F$, et $\mathbf{V}$ la matrice sur laquelle on va appliquer la NMF, les trois conditions d'unicité de la solution du problème de NMF sont :

1. **Modèle génératif** : chaque colonne de $\mathbf{V}$ peut s'exprimer comme une combinaison linéaire positive des vecteurs ψ_k.
2. **Représentativité** : chaque ψ_k apparaît au moins une fois dans la base d'exemples $\mathbf{V}$.
3. **Séparabilité** : pour chaque k, il existe une composante ψ_{fk} du vecteur ψ_k qui le caractérise, c'est-à-dire qui est nulle pour tout autre $\psi_{k'}, k' \neq k$.

Les deux premières conditions ne nous posent pas de problème. En effet, la première condition est l'hypothèse de base de toute l'approche NMF, à savoir que le spectrogramme peut être considéré comme additif en première approche (le spectrogramme d'un accord est à peu près la somme des spectrogrammes des notes qui le composent). Cette condition signifie qu'il existe une solution exacte au problème posé, et que la matrice $\mathbf{V}$ est, au plus, de rang K. La condition 2. est une tautologie : on ne peut espérer recouvrer une note que si elle est effectivement jouée au cours du morceau.

C'est la troisième condition qui est particulièrement intéressante (et problématique en transcription musicale) ici. Si on considère les spectres des notes comme une famille générative de musique, cette famille est hautement non séparable en musique tonale, qui repose justement sur les recouvrements en fréquence des notes consonantes. Prenons un exemple très simple :

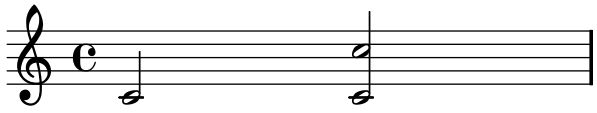


FIGURE III.1 – Le problème de l'octave.

Si les spectres des notes sont parfaitement harmoniques, celui du do_4 est entièrement inclus dans celui du do_3, la famille que constituent ces deux notes n'est pas séparable. Il y a plus grave ! On peut construire une famille séparable qui va générer l'exemple, et qui sera donc la solution unique renvoyée par l'algorithme de factorisation.

Sur la représentation très schématique de la figure III.2, à gauche, la décomposition qu'on souhaiterait obtenir n'est pas « séparable » : aucune composante fréquentielle ne caractérise le do le plus aigu. En revanche, la famille de droite est séparable et permet de générer l'exemple. [Donoho et Stodden, 2003] prouve que c'est cette dernière famille qui est la solution unique du problème, et donne à penser que c'est elle qui sera rendue par l'algorithme. Ce problème risque de se poser pour tous les accords consonants.

Il semblerait donc que les cas où la solution n'est pas unique soient finalement plus favorables à l'application de transcription. On pourrait alors choisir des critères supplémentaires pour sélectionner parmi les solutions multiples. Si la solution unique indésirable existe, il faut trouver un moyen de l'éviter ; par exemple en acceptant des solutions pour lesquelles l'erreur de reconstruction est moins bonne mais qui ont une forme plus favorable, ou par l'usage de contraintes.

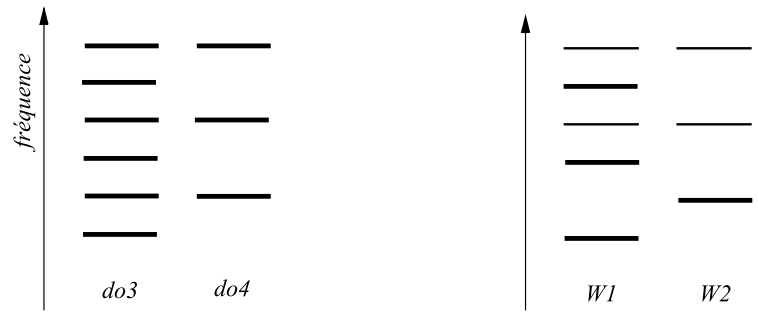


FIGURE III.2 – Deux factorisations possibles pour la séquence de la figure III.1.

III.2.3.3 Théorèmes d'unicité

[Laurberg *et al.*, 2008] établit également des résultats d'unicité sous la forme d'une condition nécessaire, et d'une condition suffisante plus restrictive, dans le cas où $\mathbf{V}$ est exactement de rang K. L'approche est géométrique et s'appuie sur les notions de dualité convexe, et de cône simpliciel (enveloppe convexe du sous-espace positivement engendré par un ensemble de vecteurs). Sans rentrer dans les détails techniques de la propriété et de sa démonstration, une interprétation géométrique intuitive (mais partielle) est de dire que l'enveloppe convexe du sous-espace engendré positivement par les colonnes de $\mathbf{W}$ est « suffisamment proche » du quadrant positif[3], ce qui peut-être vu comme une forme affaiblie de parcimonie (*cf.* section V.2.1, page 80). L'auteur signale d'ailleurs que des preuves plus fortes d'unicité peuvent être établies dans un cadre plus strictement parcimonieux [Theis *et al.*, 2005].

III.2.3.4 Commentaires

Les conditions d'unicité précédemment posées sont relativement restrictives et peu utilisables en pratique. De plus, ces travaux, à l'élégance formelle indéniable, ne préjugent pour autant ni de la capacité des algorithmes à atteindre ces solutions uniques, ni bien sûr de la pertinence sémantique de la représentation produite. Dans la suite, ces deux questions nous intéresseront plus particulièrement.

III.3 Fonctions de coût

La factorisation (III.1) est obtenue en général par la minimisation d'une fonction de coût définie par

$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \tag{III.3}$$

où $d(a|b)$ est une fonction de deux variables scalaires, en général à valeurs positives ou nulles, et s'annulant ssi $a = b$. Cette section présente les choix les plus usités dans la littérature.

3. Cette condition est à rapprocher de la condition probabiliste de [Plumbley, 2002], qui impose aux sources à séparer d'avoir une probabilité non-nulle d'être arbitrairement proche de zéro.

III.3.1 Fonctions usuelles

Les fonctions de coût les plus populaires de la littérature sont la distance euclidienne (EUC) et la divergence généralisée de Kullback-Leibler (KL), qui ont été particulièrement mises en vogue (tout comme la NMF elle-même) par [Lee et Seung, 1999,Lee et Seung, 2001]. La distance euclidienne (EUC) s'exprime :

$$d_{EUC}(a|b) = \frac{1}{2}(a-b)^2, \tag{III.4}$$

et la divergence généralisée de Kullback-Leibler (KL) :

$$d_{KL}(a|b) = a\log\left(\frac{a}{b}\right) - a + b. \tag{III.5}$$

Le choix de la distance EUC est évidemment naturel. Les auteurs de [Lee et Seung, 1999] motivent le choix de la divergence KL par des considérations probabilistes (modèle génératif des données) que nous discuterons au chapitre VI.

Dans cette thèse, nous nous intéresserons également particulièrement à la divergence d'Itakura-Saito (IS), qui s'exprime ainsi :

$$d_{IS}(a|b) = \frac{a}{b} - \log\left(\frac{a}{b}\right) - 1. \tag{III.6}$$

Cette divergence a été obtenue dans [Itakura et Saito, 1968] à partir de l'estimation du maximum de vraisemblance (MV) de spectres à court-terme de signaux de parole, dans un modèle autorégressif. Elle a été présentée comme « une mesure de correspondance entre deux spectres » [4] et est devenue populaire dans la communauté du traitement de la parole des années 1970. En particulier, elle a été appréciée pour les bonnes propriétés perceptives des signaux reconstruits auxquels elle conduit [Gray *et al.*, 1980].

Ces trois coûts sont illustrés sur la figure III.3 (page 58).

Citons enfin la possibilité de pondérer ces distances (en général, la distance euclidienne) afin d'inclure des considérations psycho-acoustiques à leur sémantique (sensibilité variable de la perception auditive humaine en fonction de la fréquence). La distance globale s'écrira alors :

$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{n=1}^{N} g_{vfn}\, d(v_{fn}|\hat{v}_{fn}) \tag{III.7}$$

où g_{vfn} est un poids dépendant de l'énergie portée par le coefficient v_{fn}, déterminé sur des bases psycho-acoustiques, ceci afin de ne pas négliger, dans la mesure de l'erreur de reconstruction, des composantes de faible énergie [Vincent et Plumbley, 2007]. L'utilisation d'une mesure pondérée dans la NMF a été suggérée pour la première fois dans un contexte de séparation de sources [Virtanen, 2004]. Elle associe un poids g_{vfn} plus élevé aux points temps-fréquence de faible énergie, et prend en compte des règles simples de masquage fréquentiel [Zwicker et Fastl, 1999].

III.3.2 Propriétés

Nous examinons ici les propriétés mathématiques des fonctions de coût précédemment présentées.

4. "*A measure of the goodness of fit between two spectra*", sic [Itakura et Saito, 1968].

III.3.2.1 Distance vs. divergence

Il faut noter en premier lieu que la fonction d n'est pas nécessairement une *distance* au sens mathématique. La seule condition nécessaire à son usage comme évaluation quantitative de la notion d'approximation est la propriété de séparation. Rappelons les trois propriétés définissant une distance :

1. Séparation : $\forall(a,b) \quad d(a|b) = 0 \Leftrightarrow a = b$
2. Symétrie : $\forall(a,b) \quad d(a|b) = d(b|a)$
3. Inégalité triangulaire : $\forall(a,b,c) \quad d(a|c) \leq d(a|b) + d(b|c)$

Notons que ces propriétés induisent la positivité de d (*i.e.* $\forall(a,b) \quad d(a|b) \geq 0$).

Si d_{EUC} remplit évidemment ces trois critères, il n'en est pas de même pour d_{KL} et d_{IS}, qui ne sont pas symétriques (cependant, elles sont aisément symétrisables). En ce qui concerne l'inégalité triangulaire, on peut vérifier les égalités :

$$d_{KL}(a|b) + d_{KL}(b|c) = d_{KL}(a|c) + (b-a)\log\left(\frac{b}{c}\right) \tag{III.8}$$

$$d_{IS}(a|b) + d_{IS}(b|c) = d_{IS}(a|c) + \frac{a}{b} + \frac{b}{c} - \frac{a}{c} - 1 \tag{III.9}$$

qui permettent de se convaincre que ces fonctions ne vérifient pas l'inégalité triangulaire. Pour cette raison, on préférera le vocable *divergence* (et non distance) pour les désigner.

Bien que n'étant pas des distances, les divergences se justifient comme mesure de l'erreur de reconstruction pourvu qu'elles soient croissantes quand $|a-b|$ croît, ce qui est le cas des divergences KL et IS.

III.3.2.2 Convexité

Une propriété qui nous intéressera particulièrement dans la suite est la propriété de convexité, qui peut s'écrire dans le cas d'une fonction d d'une variable vectorielle $\mathbf{x}$:

$$\forall\lambda \in [0;1] \quad d(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda\, d(\mathbf{x}) + (1-\lambda)d(\mathbf{y}) \tag{III.10}$$

On vérifiera ultérieurement (section III.3.3) dans un cas plus général que d_{EUC} et d_{KL} respectent cette propriété de convexité, tandis que ce n'est pas le cas de d_{IS}. Par conséquent, les coûts matriciels correspondants D_{EUC} et D_{KL} sont au moins convexes en chacune des variables prises séparément (l'autre étant fixée). En revanche, rien ne peut être dit de D_{IS} à ce sujet. Nous verrons section III.5 que la convexité permet de démontrer la décroissance des fonctions de coût sous les règles de mise-à-jour usuelles. D'autre part, on peut s'attendre à ce que la non-convexité de D_{IS} implique qu'elle possède de plus nombreux minima locaux, susceptibles de perturber les algorithmes visant à la minimiser. Cette sensibilité aux minima locaux sera examinée expérimentalement au chapitre IV.

III.3.2.3 Invariance par homothétie

Une autre propriété intéressante de la divergence IS est son invariance par homothétie, *i.e.* :

$$d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y) \tag{III.11}$$

Cette propriété n'est pas partagée par les deux autres fonctions présentées. En effet :

$$d_{EUC}(\lambda x|\lambda y) = \lambda^2\, d_{EUC}(x|y), \tag{III.12}$$
$$d_{KL}(\lambda x|\lambda y) = \lambda\, d_{KL}(x|y). \tag{III.13}$$

L'invariance par homothétie signifie que la même importance est donnée aux coefficients de $\mathbf{V}$ de faible ou forte valeur dans le calcul de l'erreur (III.3). Ainsi, une mauvaise représentation d'un coefficient v_{fn} de faible énergie sera tout aussi coûteuse qu'une mauvaise représentation d'un coefficient $v_{f'n'}$ de plus forte énergie. Cette invariance est pertinente pour la représentation de signaux audio, qui possèdent une forte dynamique, typiquement exponentiellement décroissante en fréquence. Les signaux musicaux comportent de plus des composantes transitoires de faible énergie et des parties quasi-sinusoïdales de plus forte énergie. En ce sens, la distance IS offre nativement la propriété recherchée dans l'ajout de pondérations perceptuelles à la distance euclidienne (cf. *supra*).

III.3.2.4 Modélisation d'un coefficient par zéro

Avec le critère EUC, modéliser une quantité d'énergie observée v_{fn} par son double ($\hat{v}_{fn} = 2v_{fn}$) ou par zéro ($\hat{v}_{fn} = 0$) a le même coût ; il n'est pas « très » coûteux de représenter un coefficient par zéro, et on peut craindre une représentation appauvrie, tendant à ne pas représenter du tout certains coefficients. En revanche, avec le critère IS, la modélisation par zéro est beaucoup plus pénalisée. Ceci permet de s'assurer que le modèle est proche des observations sur une échelle logarithmique, ce qui a également un sens du point de vue de la perception (loi de Weber-Fechner).

III.3.3 Fonctions de coût généralisées

En dépit de la popularité de la distance euclidienne et de la divergence de Kullback-Leibler, la littérature offre de nombreux choix alternatifs, en particulier sous forme de distances généralisées, que nous proposons de recenser et d'examiner ici.

III.3.3.1 Catalogue

[Cichocki *et al.*, 2006] propose par exemple une étude de la NMF par minimisation des divergences de Csiszár :

$$d_{CSI}(x|y) = y\,\varphi\left(\frac{y}{x}\right) \tag{III.14}$$

où φ est une application de $\mathbb{R}_+$ dans $\mathbb{R}$, convexe, continue en zéro et vérifiant $\varphi(1) = 0$.

Sous cette forme très générale, on peut exprimer différents cas particuliers (*e.g.* distances de Hellinger, χ_2 de Pearson et de Neyman) et d'autres divergences généralisées plus restrictives, comme l'α-divergence d'Amari :

$$d_{AMA}(x|y) = y\frac{(y/z)^{\beta-1}-1}{\beta(\beta-1)} + \frac{x-y}{\beta} \tag{III.15}$$

avec $\beta = (1+\alpha)/2$, qui est bien un cas de divergence de Csiszár pour $\varphi(u) = u^{\beta-1} - 1/(\beta^2-\beta) + (1-u)/\beta$. Lorsque $\beta \to 1$, l'α-divergence est égale à la divergence KL.

[Kompass, 2007] propose une autre généralisation, présentée comme une interpolation entre EUC et KL avec un degré d'asymétrie contrôlable via un paramètre $\alpha \in [0,1]$:

$$d_{KOM}(x|y) = \begin{cases} x\frac{x^\alpha - y^\alpha}{\alpha} + y^\alpha(y-x) & \alpha \in\,]0,1] \\ x(\log x - \log y) + (y-x) & \alpha = 0 \end{cases} \tag{III.16}$$

On vérifie que cette divergence est bien égale à d_{EUC} lorsque $\alpha = 1$ et à d_{KL} lorsque $\alpha = 0$.

[Dhillon et Sra, 2005] décrit, lui, la famille des divergences de Bregman, sous la forme générique :

$$d_\varphi(x|y) = \varphi(x) - \varphi(y) - \nabla\varphi(y)\,(x - y) \tag{III.17}$$

où φ est une fonction strictement convexe à valeurs dans $\mathbb{R}$, de classe C^1 (sa dérivée $\nabla\varphi$ est continue). Les trois distances usuelles peuvent être représentées sous cette forme : d_{EUC} pour $\varphi(u) = \frac{1}{2}u^2$, d_{KL} pour $\varphi(u) = u \log u$ et d_{IS} pour $\varphi(u) = -\log(u)$.

L'intérêt de ces approches généralisées est multiple : dériver des algorithmes de minimisation plus ou moins génériques et possiblement plus rapides et efficaces que les règles de [Lee et Seung, 1999], et des preuves de décroissance du critère lors de l'exécution de ces algorithmes, là encore sous une forme relativement générale.

Cependant, le choix d'une fonction de coût pour la NMF devrait en premier lieu être conduit par le type de données à analyser ; si une bonne partie de la littérature citée s'intéresse à améliorer les performances d'algorithmes pour une fonction de coût donnée, peu d'articles sont consacrés au choix d'un coût adapté au type particulier de données et d'application [Guillamet et Vitri, 2002]. C'est ce qui nous conduit notamment dans la suite à nous intéresser plus particulièrement à la divergence d'Itakura-Saito et aux distances qui la généralisent.

III.3.3.2 β-divergences

Les β-divergences, introduites par [Eguchi et Kano, 2001], sont définies comme suit :

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log\frac{x}{y} + (y - x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{III.18}$$

Comme observé dans [Cichocki *et al.*, 2006], la divergence d'Itakura-Saito est un cas limite de la β-divergence, pour $\beta \to 0$. [Eguchi et Kano, 2001] se limite au cas $\beta > 1$, mais le domaine de définition peut être étendu à $\beta \in \mathbb{R}$ sans difficulté. On montre la continuité de la β-divergence comme fonction de β en recourant à l'identité usuelle $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$. La divergence KL est obtenue de même par passage à la limite ($\beta \to 1$) et la distance euclidienne pour $\beta = 2$.

Étudions $d_\beta(x|y)$ comme une fonction de la variable y (gardons en mémoire que la variable x reçoit, en pratique, les valeurs des observations). Les dérivées du premier et second ordre de cette fonction s'écrivent :

$$\nabla_y d_\beta(x|y) = y^{\beta-2}(y - x), \tag{III.19}$$

$$\nabla_y^2 d_\beta(x|y) = y^{\beta-3}\left((\beta-1)y + (2-\beta)x\right). \tag{III.20}$$

On en déduit directement les propriétés suivantes :

- $d_\beta(x|y)$ possède un unique minimum global au point $y = x$, et croît lorsque la distance $|y - x|$ croît. Ceci justifie son utilisation comme mesure d'erreur de reconstruction, quelle que soit la valeur de β.
- $d_\beta(x|0)$ a une valeur finie si et seulement si (ssi) $\beta > 1$.
- $d_\beta(x|y)$ est convexe sur $\mathbb{R}_+$ ssi $1 \leq \beta \leq 2$.

La figure III.3 illustre les comportements typiques de la β-divergence pour des valeurs bien choisies de β.

La fonction de coût matricielle $D_\beta(\mathbf{V}|\mathbf{WH})$ n'est en général pas convexe par rapport au couple de variables $(\mathbf{W}, \mathbf{H})$, même si le coût scalaire $d_\beta(x|y)$ l'est. Cependant, lorsque $d_\beta(x|y)$ est convexe, D_β est au moins convexe en tant que fonction de $\mathbf{W}$ (resp. $\mathbf{H}$) lorsque $\mathbf{V}$ et $\mathbf{H}$ (resp. $\mathbf{W}$) sont fixées. En effet, elle s'exprime comme la somme de fonctions convexes composées avec des fonctions linéaires (conséquence de la relation (III.3)).

Remarquons également que $\forall \beta > 0$, on a la limite $\lim_{y \to +\infty} d_\beta(x, y) = +\infty$, ce qui, ajouté à la continuité de la fonction, garantit l'existence d'une solution (*cf.* section III.2).

III.4 Algorithmes usuels

Nous présentons ici les principaux algorithmes de l'état de l'art visant à minimiser les fonctions de coût précédentes.

III.4.1 Algorithmes multiplicatifs

[Lee et Seung, 1999] décrit des règles de mise à jour des facteurs $\mathbf{W}$ et $\mathbf{H}$ s'exprimant sous forme multiplicative, qui assurent à la fois la non-négativité des coefficients à chaque itération et la décroissance monotone du critère $D(\mathbf{V}|\mathbf{WH})$. Pour le problème d'EUC-NMF :

$$\begin{cases} \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T(\mathbf{WH})} \\ \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\mathbf{V}\mathbf{H}^T}{(\mathbf{WH})\mathbf{H}^T} \end{cases} \tag{III.21}$$

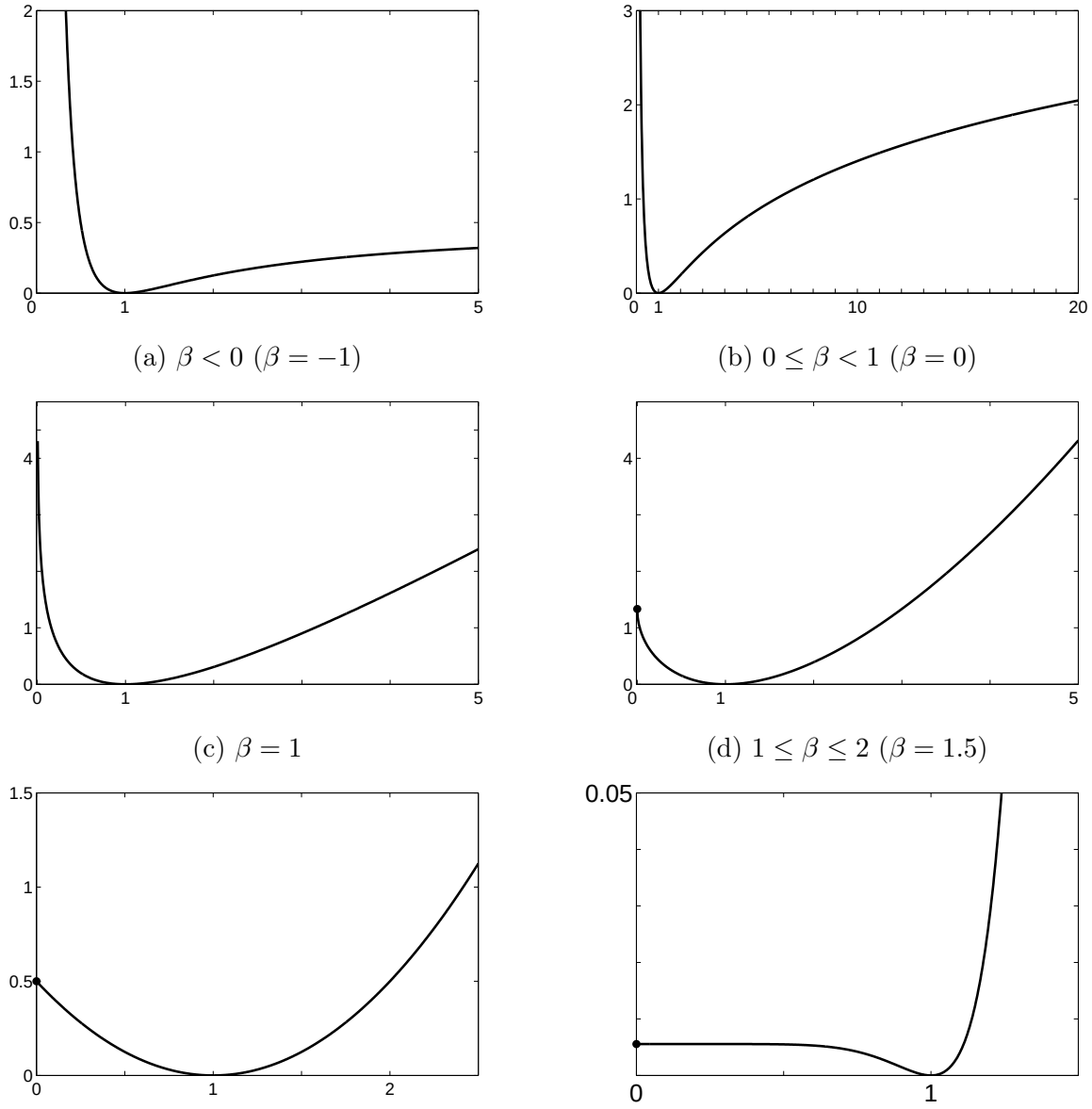
où $\otimes$ et la barre de fraction désignent respectivement le produit et le quotient de Hadamard (*i.e.* terme à terme). Pour le problème de KL-NMF ($\oslash$ désigne également le quotient de Hadamard) :

$$\begin{cases} \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\mathbf{W}^T(\mathbf{V} \oslash (\mathbf{WH}))}{\mathbf{W}^T\mathbf{1}} \\ \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{(\mathbf{V} \oslash (\mathbf{WH}))\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \end{cases} \tag{III.22}$$

Dans la suite, nous désignerons ces algorithmes multiplicatifs par les acronymes **EUC-NMF** et **KL-NMF**. [Lee et Seung, 2001] prouvent que sous ces règles de mise à jour, les distances associées ont une décroissance monotone au fil des itérations. La preuve sera discutée en section III.5.

Toutefois, si [Lee et Seung, 2001] prouve la décroissance des critères suivant ces règles, aucune allusion n'est faite à la manière d'obtenir ces règles ; elles sont d'abord posées, puis les auteurs prouvent qu'elles font effectivement décroître le critère. En réalité, ces mises à jour peuvent également être obtenues de deux manières différentes :

– Par une approche de descente de gradient, pour un choix approprié du pas de descente.



(a) $\beta < 0$ ($\beta = -1$)

(b) $0 \leq \beta < 1$ ($\beta = 0$)

(c) $\beta = 1$

(d) $1 \leq \beta \leq 2$ ($\beta = 1.5$)

(e) $\beta = 2$

(f) $\beta > 2$ ($\beta = 10$)

FIGURE III.3 – La β-divergence comme fonction de la seule variable y (avec $x = 1$). Les sous-figures (e), (c) et (b) représentent les coûts EUC, KL et IS respectivement.

– Par une heuristique simple.

En effet, un algorithme de descente de gradient minimisant D_{EUC} s'écrirait :

$$\begin{cases} \mathbf{H} = \mathbf{H} + \eta_{\mathbf{H}} \otimes ((\mathbf{W}^T\mathbf{V}) - (\mathbf{W}^T\mathbf{W}\mathbf{H})) \\ \mathbf{W} = \mathbf{W} + \eta_{\mathbf{W}} \otimes ((\mathbf{V}\mathbf{H}^T) - (\mathbf{W}\mathbf{H}\mathbf{H}^T)) \end{cases} \tag{III.23}$$

S'il on choisit des pas de descente η variables :

$$\begin{cases} \eta_{\mathbf{H}} = \dfrac{\mathbf{H}}{\mathbf{W}^T\mathbf{W}\mathbf{H}} \\ \eta_{\mathbf{W}} = \dfrac{\mathbf{W}}{\mathbf{W}\mathbf{H}\mathbf{H}^T}, \end{cases} \tag{III.24}$$

il est aisé de vérifier que l'on retrouve les règles de mises à jour (III.21). On peut faire de même pour la divergence KL.

Le second moyen d'obtention de ces règles est d'exprimer le gradient de la fonction de coût ∇D comme la différence de deux termes positifs $\nabla^+ D$ et $\nabla^- D$. On montre alors (dans des cas particuliers), ou bien on observe expérimentalement, la décroissance monotone du critère suivant les règles de mise à jour :

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\nabla^-_{\mathbf{W}} D(\mathbf{V}|\mathbf{W}\mathbf{H})}{\nabla^+_{\mathbf{W}} D(\mathbf{V}|\mathbf{W}\mathbf{H})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\nabla^-_{\mathbf{H}} D(\mathbf{V}|\mathbf{W}\mathbf{H})}{\nabla^+_{\mathbf{H}} D(\mathbf{V}|\mathbf{W}\mathbf{H})} \end{cases} \tag{III.25}$$

Ce choix repose sur l'intuition : on cherche un point fixe pour ces mises à jour, pour lequel le quotient $\nabla^- D / \nabla^+ D$ vaut 1. Cela correspond également à approcher l'égalité $\nabla^-_{\mathbf{W}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \nabla^+_{\mathbf{W}} D(\mathbf{V}|\mathbf{W}\mathbf{H})$, c'est-à-dire un gradient nul, ou encore un point critique (stationnaire) de la fonction de coût, qui sera un minimum sous de bonnes conditions.

Pour certains choix appropriés de la distance scalaire d, dont font partie la distance euclidienne et la divergence de Kullback, la décroissance du critère au cours de l'application de ces mises à jour peut être formellement prouvée [Lee et Seung, 2001], mais dans le cas général, aucune forme de convergence n'est garantie par la théorie.

On peut vérifier que la complexité de ces algorithmes est de $\mathcal{O}(KFN)$ par itération [Lee et Seung, 2001].

III.4.2 Autres algorithmes

Outre les algorithmes multiplicatifs, très populaires, la littérature offre une grande variété d'algorithmes d'autre nature, visant à « mieux » résoudre le problème de NMF (plus rapidement, en atteignant de meilleures valeurs de l'objectif) que les règles de Lee et Seung. Citons parmi ceux-ci :

– Moindres carrés alternés [Finesso et Spreij, 2004, Lin, 2008]
– Gradient conjugué [Zdunek et Cichocki, 2007]
– Gradient projeté [Lin, 2007b, Zdunek et Cichocki, 2008, Wang et Zou, 2008]
– Méthodes de Quasi Newton [Zdunek et Cichocki, 2007, Cichocki *et al.*, 2008]

Un inventaire exhaustif de ces algorithmes ainsi qu'une discussion critique est disponible par exemple dans [Cichocki *et al.*, 2009]. Contrairement aux algorithmes multiplicatifs, ils ne préservent pas naturellement la non-négativité, ce qui impose des étapes supplémentaires pour garantir cette contrainte. Par ailleurs, certains d'entre eux font intervenir des quantités de second-ordre (matrices hessiennes) dont l'estimation peut être délicate, et dont l'éventuel mauvais conditionnement peut poser des problèmes numériques. Leur technicité les rend moins souples et plus difficilement adaptables à des contraintes supplémentaires. Pour ces raisons, nous avons choisi de nous limiter dans ce mémoire à l'étude et l'utilisation d'algorithmes de type multiplicatif.

III.5 Convergence des algorithmes multiplicatifs

Les algorithmes multiplicatifs étant construits sur une heuristique, la question de leur capacité à atteindre un minimum (au moins local) de la fonction de coût se pose.

III.5.1 Décroissance du critère

Deux preuves pour des cas particuliers d'algorithmes multiplicatifs sont disponibles dans la littérature : [Lee et Seung, 2001] prouve que les algorithmes de la section III.4 font décroître le critère à chaque itération (dans les cas de la distance EUC et de la divergence KL) ; [Kompass, 2007] étend cette preuve à toutes les β-divergences pour $\beta \in [1;2]$, en suivant le même schéma de démonstration.

La preuve repose sur la définition d'une fonction $G_{\mathbf{W}} : \mathbb{R}^{K\times N} \times \mathbb{R}^{K\times N} \to \mathbb{R}^+$, dite « auxiliaire » pour D, vérifiant :

$$\begin{cases} \forall \mathbf{H}, \mathbf{H}'\ G_{\mathbf{W}}(\mathbf{H}, \mathbf{H}') \geq D(\mathbf{V}|\mathbf{WH}) \\ \qquad \forall \mathbf{H}\ G_{\mathbf{W}}(\mathbf{H}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) \end{cases} \tag{III.26}$$

On peut alors montrer que :

$$D(\mathbf{V}|\mathbf{WH}^{(\ell+1)}) \leq G_{\mathbf{W}}(\mathbf{H}^{(\ell+1)}, \mathbf{H}^{(\ell)}) \leq G_{\mathbf{W}}(\mathbf{H}^{(\ell)}, \mathbf{H}^{(\ell)}) = D(\mathbf{V}|\mathbf{WH}^{(\ell)}), \tag{III.27}$$

à chaque itération ℓ, ce qui implique alors que la règle de mise à jour de $\mathbf{H}$ ne fait pas croître $D(\mathbf{V}|\mathbf{WH})$. En inversant les rôles des deux facteurs on montre la même chose pour la règle de mise à jour de $\mathbf{W}$.

On peut noter que ce schéma est inspiré de la preuve de convergence des algorithmes de type Espérance-Maximisation (EM, [Dempster *et al.*, 1977]). D'autre part, le point important de cette preuve est que la convexité de la fonction de coût scalaire d est utilisée, et que la preuve proposée ne tient plus sans cet argument.

III.5.2 Convergence vers un point stationnaire

[Lee et Seung, 2001] prouve que la valeur du critère, nécessairement positive, décroît à chaque itération. Cela n'établit cependant pas la convergence de la suite des valeurs de $\mathbf{W}$ et $\mathbf{H}$ et ne permet pas d'affirmer que la valeur finale $(\mathbf{W}, \mathbf{H})$ est un point stationnaire de la fonction de coût (c'est-à-dire un point où le gradient s'annule), ni un minimum.

Dans le cas d'une optimisation non contrainte, les minima locaux d'une fonction sont nécessairement des points stationnaires, c'est-à-dire des points où le gradient s'annule. Dans le cas de la NMF, ceci

n'est plus vrai puisque le domaine de définition possède des bords. Les mises à jour multiplicatives (section III.4) imposent qu'aucune ligne de $\mathbf{H}$ ni aucune colonne de $\mathbf{W}$ ne soit entièrement nulle (ce qui est souhaitable, puisqu'il s'agirait de composante dégénérées), mais n'imposent pas la non nullité des coefficients pris individuellement. On peut d'ailleurs s'attendre à ce que ce soit le cas pour certains d'entre eux. Si la solution atteinte ne contient aucun coefficient nul (c'est-à-dire si elle appartient à l'intérieur du domaine de définition des paramètres), on peut alors affirmer immédiatement qu'il s'agit d'un point stationnaire, sans pour autant pouvoir en déduire qu'il s'agit d'un minimum, même local (il peut s'agir par exemple d'un point selle). Ce fait est souligné par [Berry *et al.*, 2007] dans le cas de EUC-NMF mais est tout aussi vrai lorsque les divergences KL ou IS sont utilisées.

Le cas où les algorithmes de [Lee et Seung, 2001] atteignent un point aux bords du domaine de définition est examiné dans [Lin, 2007a]. La discussion s'appuie sur les conditions d'optimalité de Karush, Kuhn et Tucker (KKT) [Kuhn et Tucker, 1951], qui stipulent qu'un point fixe $(\mathbf{W}, \mathbf{H})$ de la fonction à optimiser ne peut être un minimum local que si le gradient est nul en les coefficients w_{fk} et h_{kn} non nuls, et positif ou nul en les coefficients nuls (points où la contrainte est active) [5]. [Gonzales et Zhang, 2005] observe numériquement que ces conditions ne sont pas vérifiées après un nombre très grand (mais arbitrairement fixé) d'itérations ; il en « déduit » que les algorithmes multiplicatifs ne convergent pas nécessairement vers un minimum local, ce qui est hâtif, mais suggère *a minima* la lenteur des algorithmes et leur éventuelle sensibilité à des erreurs numériques lorsque les dénominateurs s'approchent de zéro. Plusieurs auteurs proposent d'éviter cet écueil en ajoutant au dénominateur des mises à jour une petite quantité strictement positive [Lin, 2007b, Virtanen, 2006]. [Lin, 2007b] prouve que cela permet d'atteindre un point stationnaire. Sauf cas pathologiques, il est à espérer que ce point est un minimum local.

III.6 Initialisation

Comme pour tout algorithme de type itératif, la question de l'initialisation se pose évidemment. Elle est ici d'autant plus aiguë qu'il n'y a, d'une part, pas d'unicité de la solution, et que d'autre part la fonction optimisée possède très probablement des minima locaux. Pourtant, beaucoup d'algorithmes de la littérature se contentent d'initialiser les facteurs au hasard. La question de l'initialisation a été cependant abordée par plusieurs auteurs, principalement pour des applications à l'image [Wild *et al.*, 2004, Kim et Choi, 2007, Zheng *et al.*, 2007, Xue *et al.*, 2008]. Notons que ces articles se préoccupent uniquement de l'initialisation du facteur $\mathbf{W}$, ce qui rapproche ces techniques de la problématique de « design de dictionnaire ».

Ces initialisations structurées reposent toutes sur un pré-clustering non supervisé opéré sur $\mathbf{V}$, les centroïdes des classes apprises devenant les colonnes du facteur $\mathbf{W}$ initial. Nous nous intéressons ici à deux de ces méthodes ayant montré de bons résultats pour l'initialisation de la NMF appliquée à la représentation d'images, l'une opérant sur les colonnes de $\mathbf{V}$ (regroupant donc les trames temporelles), l'autre sur les lignes de $\mathbf{V}$ (regroupant donc les points fréquentiels). Nous présentons ici ces deux méthodes ; les résultats expérimentaux correspondants sont présentés en section IV.2.

5. Ceci est une condition nécessaire, mais non suffisante. Si l'on ajoute une contrainte supplémentaire : que la matrice hessienne réduite aux points où la contrainte n'est pas active soit *définie positive*, on obtient une condition suffisante — mais non nécessaire, et non vérifiée en pratique.

III.6.1 K-moyennes sphériques

L'algorithme des k-moyennes (également appelé algorithme des nuées dynamiques, en anglais *k-means*) est un algorithme couramment utilisé en analyse de données [MacQueen, 1967]. Il permet de partitionner une collection d'objets en K classes $\mathcal{C}_1 \dots \mathcal{C}_K$, K étant un nombre fixé par l'utilisateur. L'algorithme des k-moyennes se déroule de la façon suivante :

1. Choisir K objets au hasard parmi les objets de la collection : $\{\mathbf{r}_k\}_{k=1:K} = \{\mathbf{v}_{i_k}\}_{k=1:K}$. Chaque $\mathbf{r}_k$ est appelé représentant de la classe $\mathcal{C}_k$, vide à l'initialisation.
2. Affecter chaque objet de la collection $\mathbf{v}_k$ à l'une des classes $\mathcal{C}_j$, en fonction du représentant le plus proche $\mathbf{r}_j = \text{argmin}_{\mathbf{r}_1 \dots \mathbf{r}_k} d(\mathbf{v}_k, \mathbf{r}_j)$, où d est une distance de similarité entre vecteurs.
3. Calculer de nouveaux représentants pour les classes. Ces nouveaux représentants sont calculés comme la moyenne arithmétique des membres de la classe : $\mathbf{r}_k = (\#\mathcal{C}_k)^{-1} \sum_{j/\mathbf{v}_k \in \mathcal{C}_k} \mathbf{v}_j$.
4. Retourner en 2. tant que la différence entre les anciens et les nouveaux représentants est supérieure à un seuil fixé et arbitrairement petit.

Historiquement, on a souvent choisi $d = d_{EUC}$. [Dhillon et Modha, 2001] a cependant introduit un autre choix, en proposant de normaliser les vecteurs initiaux, de manière à ne tenir compte que de leur direction, et non de leur longueur (c'est-à-dire, dans notre cas, de leur énergie). Pour ce faire, il suffit de normaliser chaque $\mathbf{v}_k$ en norme euclidienne, ou, de manière équivalente, d'utiliser la distance :

$$d(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\mathbf{y}^T \mathbf{x}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \cos \theta_{\mathbf{xy}} \tag{III.28}$$

où $\theta_{\mathbf{xy}}$ désigne l'angle entre les deux vecteurs $\mathbf{x}$ et $\mathbf{y}$, correctement défini puisque les données sont non-négatives (on évolue dans le quadrant positif). Les vecteurs et les centroïdes étant normalisés, l'inégalité de Cauchy-Schwartz entre produits scalaires garantit la cohérence des classes, c'est-à-dire le fait qu'un vecteur $\mathbf{v}_k$ appartenant à la classe $\mathcal{C}_j$ est plus proche, au sens du produit scalaire, du centroïde $\mathbf{r}_j$ que de tous les autres centroïdes $\{\mathbf{r}_i\}_{i \neq j}$.

La table III.1 résume l'algorithme des K-moyennes sphériques.

Entrée : Matrice à coefficients positifs ou nuls $\mathbf{V}$
Sortie : Clusters $\mathcal{C}_1 \dots \mathcal{C}_K$
Initialiser les clusters et leurs centroïdes.
while $\exists j \quad \left\| \mathbf{c}_j^{(\ell+1)} - \mathbf{c}_j^{(\ell)} \right\|_2 > \varepsilon$ **do**
  Calculer $\forall i, j \quad d_{ij} = \mathbf{v}_i^T \mathbf{c}_j$
  Mettre à jour les clusters : $\mathcal{C}_k = \left\{ \mathbf{v}_i / k = \text{argmax}_j \, d_{jk} \right\}$
  Recalculer les centroïdes : $\mathbf{c}_k = \frac{\sum_{\mathbf{v}_i \in \mathcal{C}_k} \mathbf{v}_i}{\left\| \sum_{\mathbf{v}_i \in \mathcal{C}_k} \mathbf{v}_i \right\|_2}$
end while

TABLE III.1 – Clustering par K-moyennes sphériques.

[Wild *et al.*, 2004] propose l'usage de cet algorithme pour l'initialisation de la NMF, en choisissant, pour tout k, $\mathbf{w}_k = \mathbf{c}_k$. Son argumentation repose à la fois sur des considérations calculatoires (accélération de la convergence des algorithmes) et sémantique. En effet, dans le modèle de NMF, et

particulièrement dans le cas de signaux audio, on cherche à représenter avec la même composante k des parties du signal de faible ou de forte énergie possédant une enveloppe spectrale similaire, les gains $\mathbf{H}$ s'occupant de représenter cette dynamique. Typiquement, en musique, nous faisons l'approximation qu'une même note jouée pianissimo ou fortissimo possède un spectre similaire une fois normalisé. Il convient donc de regrouper les spectres indépendamment de leur norme, ce qui est fait lorsqu'on remplace la norme euclidienne par le cosinus dans l'algorithme de K-moyennes.

III.6.2 Clustering des lignes

[Kim et Choi, 2007] propose une approche sensiblement différente, puisqu'il ne s'agit plus de regrouper les colonnes de $\mathbf{V}$ (spectres de la même note jouée à différents instants), mais ses lignes, l'idée étant que des lignes similaires correspondent à des points fréquentiels qui sont souvent « simultanément actifs » [6]. Autrement dit, si dans le spectrogramme, plusieurs lignes sont colinéaires, nous pouvons penser qu'il s'agit des partiels d'une même note, gagnant à être représentés dans un seul et même atome.

La technique proposée dans [Kim et Choi, 2007] repose sur un clustering hiérarchique, construit à partir de la mesure de « proximité au rang 1 » (*closeness to rank one*) définie par :

$$CRO(\mathbf{M}) \stackrel{\text{def}}{=} \frac{\sigma_1^2}{\sum_{r=1}^{R} \sigma_r^2} \tag{III.29}$$

où R est le rang de la matrice $\mathbf{M}$ et $\sigma_1 \dots \sigma_R$ sont ses valeurs propres, classées par ordre décroissant. Cette quantité, comprise entre 0 et 1, mesure la proximité de la matrice $\mathbf{M}$ au rang 1 ($CRO(\mathbf{M}) = 1 \Leftrightarrow rg(\mathbf{M})$). Par extension, on définit la même mesure entre deux clusters $\mathcal{C}_1, \mathcal{C}_2$ comme le CRO de la matrice dont les lignes sont les vecteurs v_f faisant partie de ces deux clusters. Si deux clusters contiennent des lignes proches en terme de colinéarité, le rang de la matrice associée est proche de 1, ce qui motive cette technique. La table III.2 résume cet algorithme de clustering.

Entrée : Matrice à coefficients positifs ou nuls $\mathbf{V}$
Sortie : Clusters $\mathcal{C}_1 \dots \mathcal{C}_K$
Initialisation : $\forall f \mathcal{C}_f = v_f$.
Calculer les CRO entre chaque paire de cluster $(\mathcal{C}_i, \mathcal{C}_j)$
while Nombre de clusters $> K$ **do**
  Trouver la paire de clusters $(\mathcal{C}_i, \mathcal{C}_j)$ ayant le CRO le plus élevé
  Les fusionner en un seul cluster
  Mettre à jour les CRO entre ce nouveau cluster et les autres.
end while

TABLE III.2 – Clustering par proximité au rang 1.

Une fois ce clustering effectué, chaque colonne $\mathbf{w_k}$, restreinte aux indices des lignes appartenant au cluster associé, est initialisée par le vecteur singulier dominant approchant le cluster ; les autres coefficients de $\mathbf{w_k}$ sont fixés à une petite valeur ε.

6. [Kim et Choi, 2007] traite un problème de factorisation d'ensemble d'images, et propose de regrouper les pixels souvent actifs simultanément. Ici, nous présentons son algorithme directement dans le cas des signaux audio.

L'auteur souligne que cette approche met l'accent sur le côté « représentation en sous-parties », tandis que la méthode de clustering des colonnes de [Wild *et al.*, 2004] exploite plutôt l'aspect de « redondance du signal ». De plus, cette méthode est purement déterministe, contrairement à la précédente et à d'autres approches par clustering, qui nécessitent elles-même d'être initialisées, éventuellement au hasard.

Une évaluation expérimentale de ces méthodes est proposée au chapitre suivant.

Chapitre IV

Éviter les minima locaux

Résumé

Où l'on propose et expérimente deux moyens d'éviter les minima locaux de la fonction de coût : l'initialisation structurée, et un algorithme original utilisant les fonctions de coût généralisées.

IV.1 Introduction

Les raisons théoriques évoquées au chapitre précédent laissent à craindre l'existence de minima locaux des fonctions de coût, et l'incapacité des algorithmes multiplicatifs à les éviter. Dans ce chapitre, nous nous interrogeons sur l'existence pratique de ces minima et le comportement des algorithmes. Pour ce faire, nous proposons d'étudier dans un cadre expérimental la convergence des algorithmes multiplicatifs, la valeur finale du coût qu'ils atteignent et le lien entre cette valeur et la pertinence de la factorisation produite en termes de représentation du signal musical. Dans la section IV.2, nous examinons ces questions en comparant différentes initialisations de ces algorithmes, y compris l'initialisation standard, c'est-à-dire au hasard. Incidemment, nous établissons l'intérêt de la divergence d'Itakura-Saito par rapport aux autres fonctions de coût, mais aussi ses faiblesses. La confirmation de l'existence de nombreux minima locaux et l'échec de ces initialisations structurées à les éviter nous amène à proposer à la section IV.3 un nouveau schéma de mise à jour pour la minimisation de la divergence d'Itakura-Saito.

IV.2 Initialisation

Nous examinons ici l'existence de minima locaux des trois fonctions de coût usuelles de la NMF sur des exemples musicaux simples, et l'influence de l'initialisation sur la capacité des algorithmes multiplicatifs à éviter ces minima. Ce travail a été présenté sous forme de poster en congrès international [Bertin et Badeau, 2008].

IV.2.1 Sur un exemple simple

IV.2.1.1 Expérience

Nous analysons l'exemple simple déjà évoqué au chapitre I (page 41) : contenant 4 notes différentes.

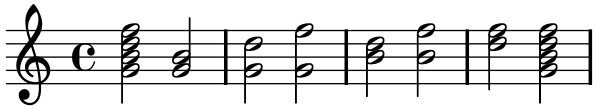


Figure IV.1 – Un exemple déjà vu.

Nous l'analysons cette fois dans une version audio réelle, enregistrée sur un piano acoustique[1]. En préambule, on détermine expérimentalement l'ordre optimal pour ce cas : $K = 6$ (une pour chaque note, une pour le transitoire et une pour le résiduel, *cf.* section IX.2.2 page 130). Après calcul du spectrogramme, on réalise les NMF avec coûts EUC, KL et IS et différentes conditions d'initialisation :

- Tirage de chaque coefficient au hasard, comme la valeur absolue d'un variable $\mathcal{N}(1, 1)$ (RAND)
- Tirage au hasard (suivant une loi uniforme) de K colonnes de $\mathbf{V}$ (VCOL)
- K-moyennes sphériques (*cf.* section III.6.1) à partir d'un tirage RAND (RAND+SKM)
- K-moyennes sphériques à partir d'un tirage VCOL (VCOL+SKM)
- Clustering par proximité au rang 1 (CRO, *cf.* section III.6.2).

1. Avec l'aimable collaboration de Mélanie Desportes, pianiste semi-professionnelle.

Les algorithmes d'initialisation sont décrits dans la section III.6 (page 61). Cette expérience est répétée pour 50 tirages RAND et 50 tirages VCOL. On examine le nombre de minima locaux observés, le taux de convergence vers le minimum global observé, la valeur finale de la fonction de coût et la vitesse de convergence de l'algorithme NMF.

IV.2.1.2 Résultats

Pour repérer l'existence éventuelle de minima locaux vers lesquels les algorithmes multiplicatifs convergeraient, nous visualisons, pour chaque couple distance/méthode d'initialisation, l'histogramme des valeurs finales du coût atteintes au cours des 50 réalisations. Ces histogrammes sont présentés sur la figure IV.2.

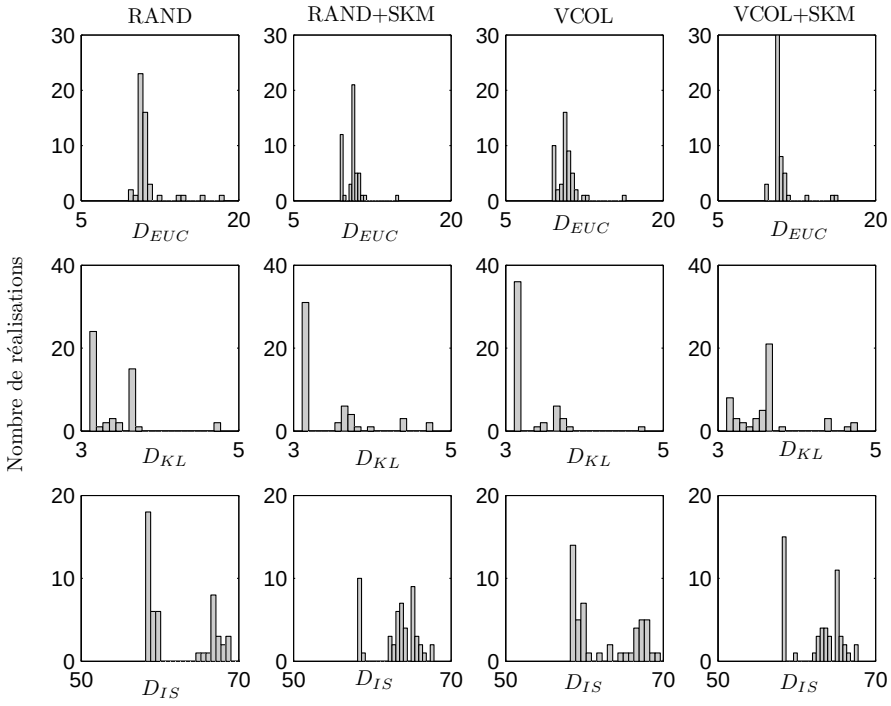


FIGURE IV.2 – Valeurs finales du coût.

Nous pouvons constater l'existence de minima locaux pour les trois distances, et ce, quel que soit le mode d'initialisation choisi. Suivant le coût et le mode d'initialisation, une proportion plus ou moins grande de réalisations conduisent à la plus faible valeur observée sur l'ensemble de l'expérience : tandis que D_{EUC} semble relativement robuste à ses minima locaux (la proportion de réalisations s'achevant sur un minimum local est faible), D_{IS} en revanche y est plus sensible (plus de la moitié des expériences sont dans ce cas). Le mode d'initialisation choisi ne semble pas influer significativement sur la répartition des valeurs finales du coût.

Par ailleurs, nous pouvons examiner, sur un exemple, l'influence de la valeur du critère sur la pertinence de la représentation. Nous représentons sur les figures IV.3 et IV.4 deux réalisations de l'algorithme multiplicatif de minimisation de D_{IS} initialisé au hasard, pour lesquelles la valeur finale du coût est significativement différente (correspondant aux deux pics observables sur la dernière ligne

de la figure IV.2). Sur la figure IV.3, nous pouvons constater que les quatre notes de l'exemple IV.1 sont correctement séparées et identifiables (ce que nous appellerons une « solution admissible »). En revanche, la réalisation représentée sur la figure IV.4 échoue clairement à représenter l'exemple avec des composantes monopitch. Elle atteint pourtant une valeur inférieure du coût, ce qui soulève la question du rapport entre la valeur numérique du critère et la pertinence sémantique de la représentation obtenue.

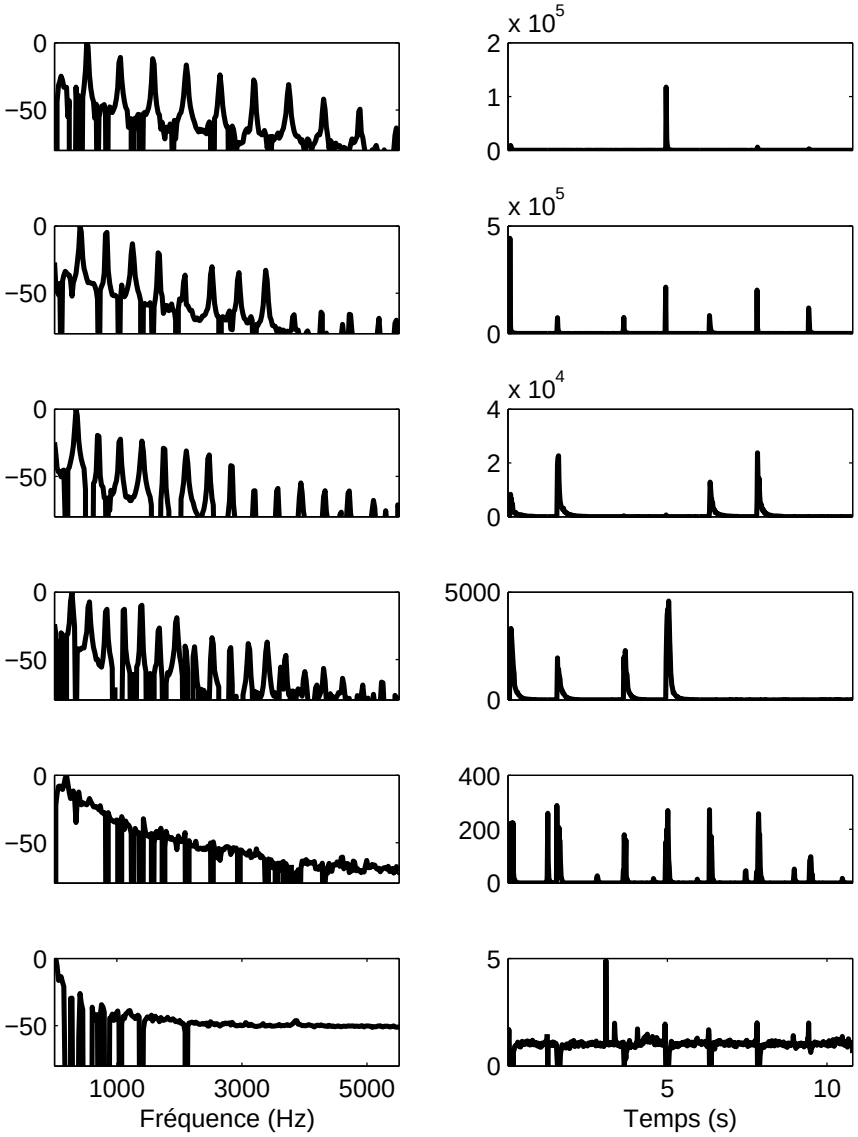


FIGURE IV.3 – Une factorisation réussie, $D_{IS}(\mathbf{V}, \mathbf{WH}) \approx 65000$.

En ce qui concerne les autres coûts, nous observons que les solutions rendues par l'algorithme EUC-NMF sont toutes admissibles, mais que les bases produites sont très bruitées, ce qui laisse craindre

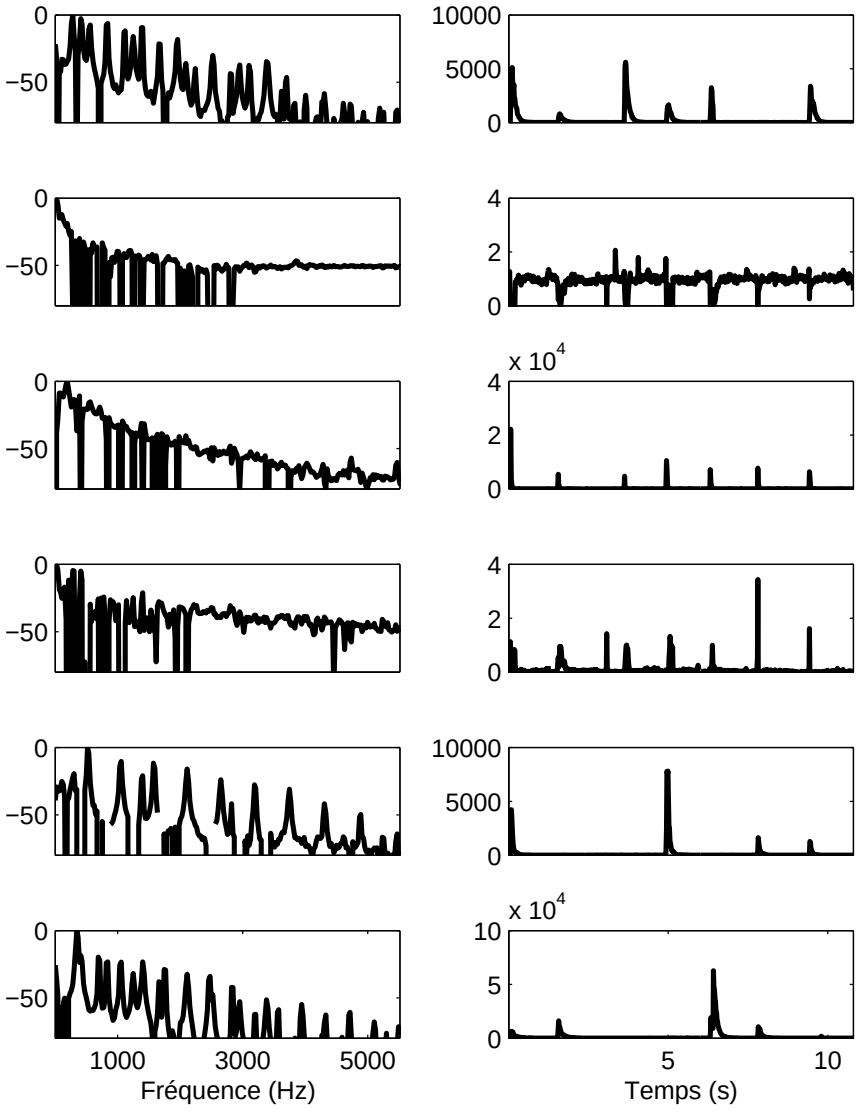


FIGURE IV.4 – Une factorisation ratée, $D_{IS}(\mathbf{V}, \mathbf{WH}) \approx 58000$

une mauvaise séparation et une difficulté à l'estimation du pitch dans des cas de dimensions plus importantes. KL-NMF produit à la fois des solutions admissibles (mais bruitées) et non-admissibles, mais qui sont cette fois corrélées avec la valeur du coût, sans explication évidente.

IV.2.2 Sur données réelles

IV.2.2.1 Expérience

Dans cette section, nous considérons la tâche complète de « WAV vers MIDI » telle que définie à la section I.2.3 (page 22). Davantage de détails sur le système complet de transcription et les métriques d'évaluation seront données dans la section IX.3 (page 138).

Nous calculons les transcriptions MIDI des 30 premières secondes d'un morceau de musique enregistré sur DisKlavier et fourni avec une annotation MIDI de référence [Bello *et al.*, 2006] [2], en considérant les 5 méthodes d'initialisation précédentes. L'expérience est reproduite 10 fois.

IV.2.2.2 Résultats

Nous évaluons à la fois la performance de transcription, son rapport avec la valeur finale de la fonction de coût, et la vitesse de convergence des algorithmes.

La figure IV.5 montre, par exemple, l'évolution de la valeur de la distance euclidienne au cours de l'algorithme en fonction de la méthode d'initialisation employée. Nous pouvons voir que les quatre initialisations conduisent à peu près à la même valeur finale du coût en un nombre comparable d'itérations ; si les initialisations structurées semblent accélérer l'algorithme pendant les premières itérations, la tendance s'inverse rapidement et, dans le cas de l'initialisation au hasard, l'évolution est plus régulière. La forme des courbes suggère que la NMF initialisée par RAND+SKM approche un minimum local vers $\ell = 200$, dont elle parvient à s'échapper.

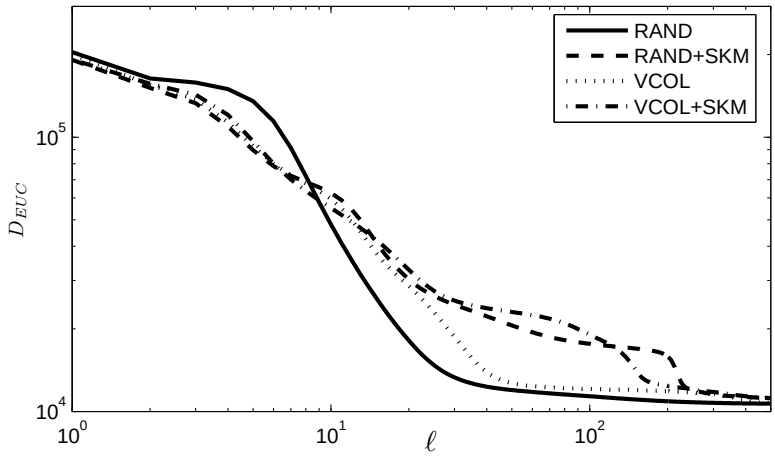


FIGURE IV.5 – Évolution de la distance EUC au cours des 500 premières itérations de NMF, pour différents algorithmes d'initialisation.

2. *Bagatelle n°1*, opus 33, Ludwig van Beethoven. Utilisée avec l'aimable autorisation de Juan Bello *et al.*

Nous examinons ensuite la performance de transcription. Afin d'évaluer le potentiel de chacune des méthodes, nous sélectionnons d'abord la meilleure transcription parmi les 10 expériences, en réglant manuellement le seuil de détection. Ces résultats sont résumés dans la table IV.1. Une fois de plus, il est difficile de dégager une tendance positive de l'initialisation structurée par rapport à l'initialisation au hasard. En revanche, nous remarquons que la NMF par minimisation de la divergence IS est la plus performante (sauf pour VCOL+SKM, mais son score reste comparable à celui obtenu par D_{EUC} sur ce test).

	D_{EUC}	D_{KL}	D_{IS}
RAND	58%	52%	**70**%
RAND+SKM	62%	60%	**82**%
VCOL	64%	58%	**67**%
VCOL+SKM	**63**%	55%	60%
CRO	45%	40%	**74**%

TABLE IV.1 – Performance maximale de transcription pour différents coûts et initialisations.

Enfin, nous nous interrogeons sur une éventuelle corrélation entre la valeur finale du coût et la performance de transcription. Pour ce faire, nous affichons sur la figure IV.6 les nuages de points correspondant à chaque couple coût/initialisation dans le plan coût/F-mesure, pour les 10 expériences et en choisissant un seuil de détection commun maximisant la F-mesure moyenne. Sur cette figure, on constate à nouveau l'absence de corrélation nette entre la valeur finale du coût et la performance de transcription. On remarque aussi qu'à seuil fixé pour tous les morceaux, la divergence IS semble perdre son intérêt sur les autres distances : les performances dans ces conditions sont très variables, et pas clairement meilleures que celles obtenues par minimisation de D_{EUC}.

IV.2.3 Conclusion

On observe que les trois fonctions de coût possèdent des minima locaux, même la distance euclidienne (supposée possiblement exempte de ces minima). L'utilisation d'une initialisation structurée (SKM et CRO) ne garantit pas la convergence vers un éventuel minimum global. Il faut cependant remarquer que les composantes extraites peuvent être sémantiquement pertinentes, et la performance de transcription correcte, même s'il s'agit d'un minimum local : la minimisation optimale du critère n'est donc pas nécessairement un objectif à poursuivre, ce qui motivera au chapitre V (page 79) le recours à des approches contraintes.

IV.3 Algorithme tempéré

Devant l'échec de l'approche par initialisation structurée, nous proposons une autre approche d'évitement des minima locaux, utilisant les distances généralisées précédemment décrites (section III.3.3). Cette section décrit l'algorithme et les résultats expérimentaux auxquels il conduit. Ce travail a donné lieu à publication [Bertin *et al.*, 2009b].

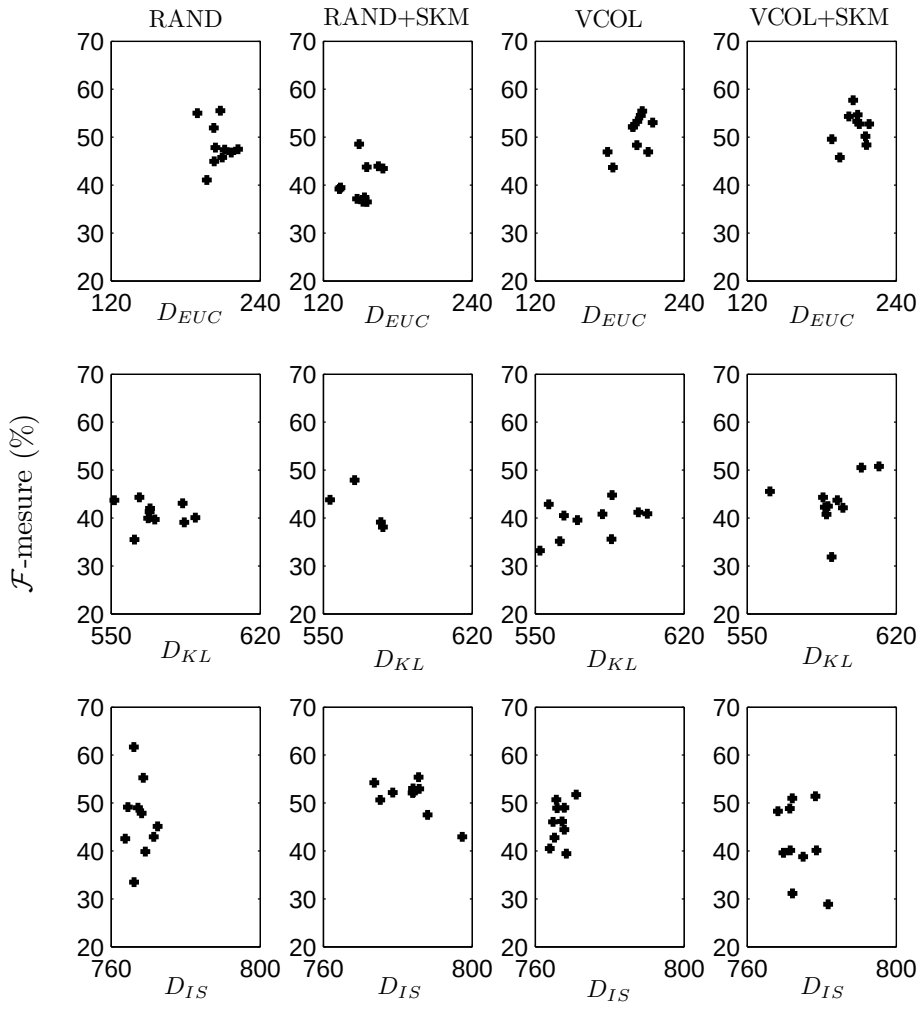


FIGURE IV.6 – F-mesure en fonction de la valeur finale du coût, pour différents coûts et algorithmes d'initialisation.

IV.3.1 Algorithme

Compte-tenu de leur pouvoir de généralisation et d'interpolation entre les trois distances EUC, KL et IS, nous choisissons comme point de départ les β-divergences décrites au paragraphe III.3.3.2 (page 56). En calculant les dérivées partielles $\nabla_{\mathbf{H}} D_\beta(\mathbf{V}|\mathbf{WH})$ (resp. $\nabla_{\mathbf{W}} D_\beta(\mathbf{V}|\mathbf{WH})$) à partir de l'équation (III.19) (page 56), puis en appliquant la règle (III.25) (page 59), on obtient l'algorithme multiplicatif alterné [Cichocki *et al.*, 2006] :

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V} \otimes (\mathbf{WH})^{\cdot[\beta-2]})}{\mathbf{W}^T((\mathbf{WH})^{\cdot[\beta-1]})} \tag{IV.1}$$

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V} \otimes (\mathbf{WH})^{\cdot[\beta-2]})\mathbf{H}}{((\mathbf{WH})^{\cdot[\beta-1]})\mathbf{H}^T} \tag{IV.2}$$

L'application de ces règles laisse cependant craindre des minima locaux, observés notamment section IV.2 dans des cas particuliers. Cependant, comme on a observé que la distance euclidienne ($\beta = 2$) en souffrait moins que le coût IS ($\beta = 0$), on peut imaginer exploiter cette propriété et l'expression générique disponible pour améliorer la minimisation de ce dernier, tout simplement en utilisant le paramètre β comme un paramètre de « température » à faire décroître. L'algorithme dit de « NMF tempérée » que nous proposons s'écrit donc :

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V} \otimes (\mathbf{WH})^{\cdot[\beta(\ell)-2]})}{\mathbf{W}^T((\mathbf{WH})^{\cdot[\beta(\ell)-1]})} \tag{IV.3}$$

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V} \otimes (\mathbf{WH})^{\cdot[\beta(\ell)-2]})\mathbf{H}}{((\mathbf{WH})^{\cdot[\beta(\ell)-1]})\mathbf{H}^T} \tag{IV.4}$$

où ℓ désigne le numéro de l'itération en cours. Il s'agit donc simplement de faire dépendre β de ℓ dans les règles de mise à jour (IV.1) et (IV.2). L'idée consiste à faire débuter l'algorithme à une température favorable (zone où la β-divergence est convexe et/ou moins sensible aux minima locaux) et à faire décroître progressivement cette température jusqu'à atteindre le coût cible, c'est-à-dire la divergence IS dans notre cas. Plus précisément, nous utilisons le schéma de température décrit dans la figure IV.7 ; β prend la valeur β_i pendant ℓ_i iterations, puis décroît suivant un arc de cosinus pendant les ℓ_d itérations suivantes, avant de se maintenir à la valeur finale β_e pendant les ℓ_e dernières itérations. Dans la suite, le préfixe et l'indice ($\beta_e \rightarrow \beta_i$) désigneront une instanciation particulière de ce schéma, décrite par les valeurs initiale et finale de β. Les paramètres ℓ_i, ℓ_d et ℓ_e sont fixés.

IV.3.2 Résultats expérimentaux : données de synthèse

Dans cette partie, nous évaluons la convergence de l'algorithme tempéré et sa capacité à atteindre de meilleurs minima de D_{IS}, en comparaison des algorithmes multiplicatifs à β fixe.

IV.3.2.1 Cadre expérimental

Dans un premier temps, nous utilisons ici des données synthétiques, afin de connaître une certaine forme de vérité-terrain sur les données expérimentales. En effet, nous verrons ultérieurement (chapitre VI) qu'il existe un modèle génératif de données pour lequel la minimisation de la distance IS est pertinente dans un sens statistique. Nous utilisons donc ce modèle (voir équation (VI.15)) pour générer un jeu de données de synthèse. Les matrices $\mathbf{W}_0$ et $\mathbf{H}_0$ (vérité-terrain) sont tirées au hasard et

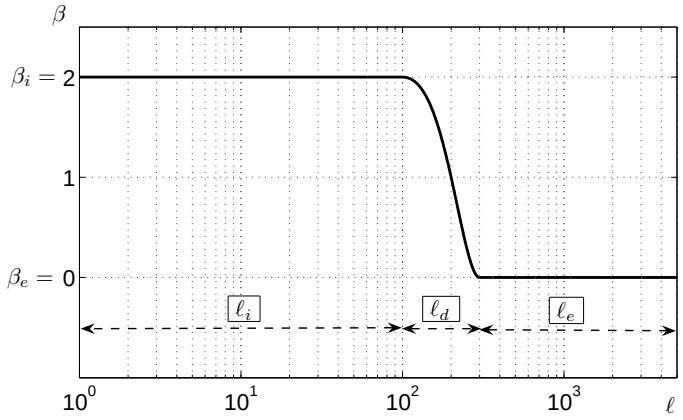


FIGURE IV.7 – Évolution de β en fonction du nombre d'itérations ℓ.

Paramètre	F	K	N	ℓ_i	ℓ_d	ℓ_e
Valeur	50	5	500	100	200	4700

TABLE IV.2 – Paramètres utilisés pour les simulations de la section IV.3.2

fixées. Le bruit multiplicatif $\mathbf{B}$ est généré comme réalisation d'un processus aléatoire de distribution Gamma, de paramètres de forme et d'échelle égaux à 1. La matrice $\mathbf{V}_0$ à factoriser est ensuite formée suivant la relation $\mathbf{V}_0 = (\mathbf{W}_0\mathbf{H}_0) \otimes \mathbf{B}$. On tire ensuite une initialisation (W_i, H_i) au hasard, à partir de laquelle on calcule les factorisations $(W^{\beta_i \to \beta_e}, H^{\beta_i \to \beta_e})$ pour différentes valeurs de β_i et β_e. Pour chaque réalisation V_0, on répète cette expérience 100 fois, pour 100 tirages du couple initial ($\mathbf{W}_i$,$\mathbf{H}_i$). Enfin, ce jeu d'expériences est répété pour 10 réalisations $\mathbf{V}_0$ différentes.

Les dimensions des matrices sont choisies inférieures à l'application aux données musicales pour des raisons de temps de calcul, mais leurs proportions sont compatibles avec le cas réel. Le paramètre de forme du bruit Gamma est choisi suffisamment grand pour que la distribution soit assez piquée autour de 1 (voir figure VII.1, page 108).

Pour chaque réalisation et chaque initialisation, les paires (β_i, β_e) testées sont : (10,0), (2,0), (1,0), (0,0). La table IV.2 résume les valeurs des paramètres utilisés pour la simulation.

IV.3.2.2 Résultats

On considère une factorisation produite par un algorithme et une initialisation donnée comme un « succès » si l'inégalité $D_{IS}^{\beta_e \to \beta_i}(V_0|W_0H_0) \leq D_{IS}^{0 \to 0}(V_0|W_0H_0)$ est vérifiée, puisqu'elle signifie que l'algorithme a atteint une valeur du coût comparable à la distance entre les observations bruitées $\mathbf{V}_0$ et la vérité-terrain $\mathbf{W}_0\mathbf{H}_0$. La table IV.3 présente les résultats des algorithmes testés.

$\beta_i \to \beta_e$	10→0	2→0	1→0
Taux de succès	18	100	98

TABLE IV.3 – Taux de succès (%).

Dans un but illustratif, nous présentons également sur la figure IV.8 l'évolution de la valeur de

$D_{IS}(\mathbf{V}_0|\mathbf{WH})$, en fonction du nombre d'itération, pour des réalisations significatives sélectionnées à la main, avec $\beta_e = 0$ et $\beta_i = 2, 0$.

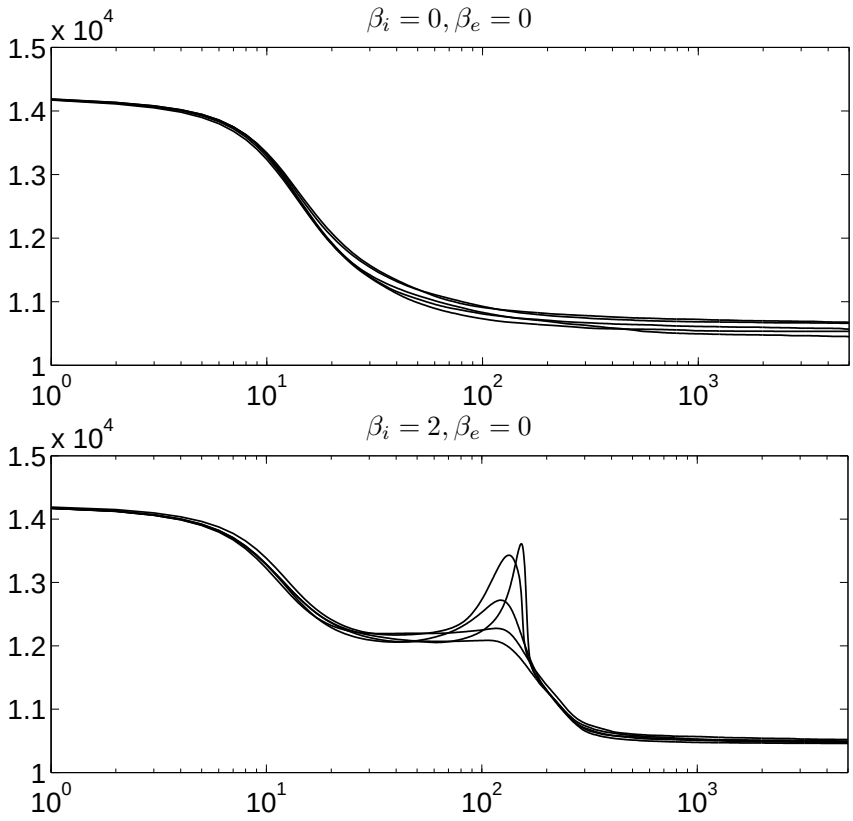


FIGURE IV.8 – Divergence IS *vs.* nombre d'itérations.

IV.3.2.3 Discussion

Comme on peut le voir dans la table IV.3, l'approche tempérée avec $\beta_e = 2$ est particulièrement efficace pour atteindre une valeur finale du coût inférieure à celle obtenue par IS-NMF classique, lorsque les données suivent le modèle statistique sous-jacent. L'algorithme tempéré avec $\beta_e = 1$ semble aussi performant dans cette table, mais il faut cependant souligner que $D_{IS}^{2\to 0}(V_0|WH) \leq D_{IS}^{1\to 0}(V_0|WH)$ dans la plupart des cas. Cette inégalité, ajoutée à la performance médiocre de l'algorithme tempéré $(10 \to 0)$, suggère l'importance de la « zone de convexité » $(1 \leq \beta \leq 2)$ et de son exploitation lors des premières itérations de l'algorithme.

Bien qu'aucun minimum local n'apparaisse clairement dans ce jeu de données, les valeurs finales atteintes par $(0 \to 0)$-NMF sont plus dispersées que celles obtenues par approche tempérée $(2 \to 0)$. Ceci pourrait corroborer l'existence de minima locaux observée dans la section III.2.

$(10 \to 0)$-NMF montre des performances relativement mauvaises. Plusieurs explications peuvent être envisagées. D'une part, la non-convexité du coût d_β pour $\beta > 2$ peut laisser présumer l'effet préjudiciable d'un tel choix. Les données synthétiques utilisées ici possèdent une dynamique faible (comparée à des signaux audionumériques réels), ce qui émousse l'intérêt de la propriété $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$, principalement utile dans des contextes à forte dynamique. De plus, comme on peut le voir sur la figure III.3, d_{10} diverge très rapidement vers l'infini lorsque $|y - x|$ croît, ce qui fera sans doute

diverger l'algorithme si l'initialisation contient des valeurs trop élevées. Enfin, puisqu'aucun minimum local net n'est réellement observé et que nous considérons des dimensions F, K, N relativement faibles, nous pouvons avancer la conclusion que l'approche tempérée à trop haute température pourrait être non seulement inutile mais éventuellement délétère.

En dépit du manque de preuve formelle, nous vérifions ici que D_{IS} est bien décroissante au décours de l'algorithme $(0 \to 0)$-NMF (IS-NMF). Au contraire, on observe des phases de croissance de D_{IS} lors de l'exécution de $(2 \to 0)$-NMF, ce qui est le comportement attendu d'un algorithme tempéré. Nous remarquons que plus le « maximum local » atteint au cours de cette phase (autour de $\ell = 100$) est élevé, et plus le coût finalement atteint est faible. Ceci valide le recours à l'approche tempérée : en permettant à D_{IS} de croître, on atteint des solutions approchées du problème de NMF inaccessibles par les approches usuelles.

IV.3.3 Résultats expérimentaux : données réelles

Dans cette section, nous considérons la tâche complète de « WAV vers MIDI » telle que définie à la section I.2.3 (page 22). Plus de détails sur le système complet de transcription et les métriques d'évaluation seront donnés dans la section IX.3 (page 138).

Nous calculons les transcriptions MIDI de six pièces enregistrées sur DisKlavier, fournies avec leur annotation MIDI de référence, et d'une durée de 30 secondes [Bello *et al.*, 2006] [3]. Ces pièces appartiennent au répertoire classique du piano, mais sont toutefois relativement simples sur le plan de la virtuosité (tempo et polyphonie relativement faible). Chaque pièce est factorisée à partir de 10 initialisations au hasard différentes, et pour différents couples $(\beta_i \to \beta_e)$. Les fichiers MIDI produits et de référence sont alors comparés pour produire des scores moyens de transcription (*cf.* section IX.3.3), page 139. La table IV.4 reporte les résultats obtenus pour six algorithmes avec β variable ou fixe.

$\beta_i \to \beta_e$	10→0	2→0	1→0	0→0	2→2	1→1
Précision	**83.4**	73.6	69.7	77.2	67.8	70.5
Rappel	**79.2**	79.2	73.6	70.6	73.6	65.5
F-measure	**81.3**	76.3	71.6	73.7	70.6	67.9

TABLE IV.4 – Performance moyenne de transcription. (%).

Nous pouvons faire les observations suivantes. Premièrement, tous les algorithmes mettant en jeu la divergence IS, qu'ils soient tempérés ou non, atteignent de meilleures performances de transcription que KL-NMF et EUC-NMF, ce qui confirme l'intérêt de cette divergence pour la représentation de signaux audio. L'approche tempérée avec $\beta_i = 10$ conduit à la meilleure performance de transcription, en dépit de la non-convexité de la β-divergence pour $\beta > 2$. Une explication possible à ce fait étonnant est la propriété $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$; ainsi, les premières itérations de $(10 \to 0)$-NMF mettent l'accent sur la bonne représentation des composantes de forte énergie. On peut alors voir $(10 \to 0)$-NMF comme un algorithme représentant en premier lieu les événements « les plus importants », puis raffinant la qualité de la description à mesure que β décroît. Les données analysées possèdent la forte dynamique typique des signaux audio, ce qui expliquerait la différence de résultats avec les données synthétiques de la question précédente ; ainsi, la représentation prioritaire des composantes de forte énergie pourrait être décisive dans le cas de signaux réels, à forte dynamique. D'autre part, les dimensions du problème

3. Utilisées avec l'aimable autorisation des auteurs.

le rendent plus susceptible de posséder de nombreux minima locaux (comparé au cas synthétique), ce qui motive le choix d'une température initiale élevée.

L'approche tempérée avec $\beta_i = 2$ se classe deuxième dans notre test. Cependant, $(0 \to 0)$-NMF atteint une valeur finale du coût D_{IS} inférieure dans tous les cas. Comme dans la section précédente, aucune corrélation claire n'est observée entre le coût atteint et la performance de transcription.

Un autre résultat remarquable est l'existence d'échecs sévères de certaines réalisations de $(0 \to 0)$-NMF, qui n'apparaissent pas dans la table IV.4 qui présente des performances moyennées, et pour lesquelles la F-mesure est inférieure à 10%. À l'inverse, nous n'avons pas observé de tels échecs pour EUC-NMF, KL-NMF et les approches tempérées. Cependant, si nous excluons ces cas pathologiques du calcul de la performance moyenne, $(0 \to 0)$-NMF est aussi performante que les approches tempérées.

IV.4 Conclusion

Cette étude confirme la pertinence de la divergence IS pour des tâches de traitement du signal audio, et l'amélioration apportée par l'approche tempérée. Cependant, l'absence de corrélation nette entre la valeur du critère et la pertinence musicale de la factorisation obtenue, et la difficulté d'éviter les minima locaux avec certitude suggère l'insuffisance de la seule contrainte de non-négativité, malgré un choix adapté de la fonction de coût. Ces constatations nous amènent à examiner, dans le chapitre suivant, l'ajout de contraintes supplémentaires à la NMF, ce qui a été également récemment suggéré dans [Klingenberg *et al.*, 2009].

Chapitre V

Variantes contraintes de la NMF

Résumé

Où, afin d'améliorer la forme des solutions, l'on s'intéresse à des variantes de la NMF intégrant d'autres contraintes que la non-négativité, en particulier à la contrainte d'harmonicité qui sied à l'analyse de sons musicaux ; ce qui nous amène à proposer un modèle d'harmonicité pour la NMF et un algorithme pour résoudre ce nouveau problème.

V.1 Introduction

Dans le problème de NMF standard, la seule contrainte explicite est la non-négativité de tous les coefficients des matrices en présence. Toute autre propriété de la décomposition, aussi satisfaisante soit-elle, est un effet secondaire incontrôlé. En un certain sens, la relative sémantique qui se dégage de la décomposition, l'efficacité de la séparation de composantes pertinentes, et l'interprétabilité de ces composantes sont simplement de « bonnes nouvelles ». L'idée d'améliorer ce potentiel, en ajoutant des contraintes explicites au problème de factorisation pour renforcer et contrôler ces propriétés, paraît donc naturelle.

Ainsi, plusieurs contraintes ont été introduites dans la littérature, afin que les solutions de la NMF remplissent au mieux certaines attentes. Parmi les contraintes proposées antérieurement, citons la parcimonie [Hoyer, 2004], la localisation spatiale [Li *et al.*, 2001], la décorrélation maximale entre les sources [Zhang et Fang, 2007] ou encore la continuité temporelle [Virtanen, 2007, Chen *et al.*, 2006].

Quelle que soit la contrainte considérée, ces algorithmes partagent une approche commune : l'ajout d'un terme de pénalité. Plutôt que la minimisation exclusive d'un terme D_r quantifiant l'erreur de reconstruction (EUC ou KL le plus souvent), la fonction de coût globale minimisée en pratique inclut un terme D_c quantifiant la propriété que l'on souhaite imposer. Le problème de NMF contraint s'exprime alors comme :

$$\min_{\mathbf{W},\mathbf{H}} D_r(\mathbf{V}|\mathbf{WH}) + \lambda D_c(\mathbf{W},\mathbf{H})$$

où λ est un paramètre de pondération, dont le choix est important et peut s'avérer délicat. En effet notre objectif principal reste d'obtenir une factorisation approchée de $\mathbf{V}$, or à mesure que $\mathbf{WH}$ s'en rapproche, les pénalisations ajoutées prennent relativement de plus en plus de poids, à moins que les termes qu'elles multiplient ne se rapprochent eux aussi suffisamment de zéro. Cela conduit certains auteurs à introduire une pondération variable, dépendant elle-même de $\mathbf{W}$ et/ou de $\mathbf{H}$ (*e.g.* [Virtanen, 2007]).

Dans ce chapitre, nous présentons un panorama des diverses contraintes de la littérature (section V.2), dont nous discutons l'intérêt dans le cas des signaux audio. Nous présentons ensuite dans la section V.3 un modèle et un algorithme permettant de résoudre un problème de NMF où les bases $\mathbf{W}$ sont contraintes à posséder une structure harmonique similaire à celle des spectres de notes de musique, et correspondant donc à la décomposition attendue.

V.2 Panorama des variantes contraintes

V.2.1 Parcimonie

La parcimonie (en anglais *sparsity*) est d'abord définie assez vaguement comme la propriété d'être « souvent nul ». Cette propriété a été largement étudiée et utilisée. Dans le cas le plus général, le problème se pose de la manière suivante :

$$\min D(\mathbf{x}|\mathbf{Ua}) + \|\mathbf{a}\|_0 \tag{V.1}$$

où $\mathbf{x}$ est un vecteur de données, $\mathbf{U}$ un dictionnaire et $\mathbf{a}$ la décomposition de $\mathbf{x}$ sur ce dictionnaire. La norme $\|.\|_0$ consiste simplement à compter le nombre de coefficients non nuls de la décomposition $\mathbf{a}$.

Le plus souvent, le dictionnaire **U** est fixé arbitrairement (ondelettes, transformée en cosinus discrets modifiée) ou appris sur une base de données extérieures. Sa détermination constitue un problème en soi. Il est généralement redondant (*overcomplete*) c'est-à-dire que $K >> \max(F, N)$, d'où l'enjeu de la décomposition parcimonieuse, qui permet de sélectionner seulement quelques vecteurs dans le dictionnaire parmi un très grand nombre — et qui, compte tenu des dimensions du problème, peut rapidement atteindre une importante complexité. Les algorithmes permettant de déterminer une décomposition **a** sont dits « algorithmes de poursuite ». Le plus connu d'entre eux est certainement l'algorithme de *Matching Pursuit* [Mallat et Zhang, 1993], mais de très nombreuses variantes existent, et en particulier des variantes imposant la non-négativité de la décomposition, tels que que la variante non-négative de *Basis Pursuit* proposée dans [Hoyer, 2002] ou FOCUSS+ [Murray et Kreutz-Delgado, 2004]. Enfin, nous pouvons citer [Aharon *et al.*, 2005] qui propose une adaptation non-négative de l'algorithme des K-SVD [Aharon *et al.*, 2006], qui permet à la fois d'apprendre le dictionnaire et de déterminer la décomposition.

Le codage parcimonieux, qui consiste à résoudre le problème (V.1) ou une forme approchée de celui-ci, a été largement traité et utilisé pour la séparation de sources audio, la représentation de signaux musicaux et leur transcription [Benaroya *et al.*, 2003, Abdallah et Plumbley, 2004, Blumensath et Davies, 2006, Daudet et Torrésani, 2006, Leveau, 2007].

Dans sa formulation, ce problème est analogue à la NMF, à ceci près que la nature des observations n'est pas nécessairement la même (dictionnaire de formes d'ondes, transformées en ondelettes ou en cosinus discrets modifiée), et que d'autre part nous considérons un problème de réduction de rang, c'est-à-dire le contraire d'un dictionnaire redondant. La question de la pertinence de cette contrainte dans le cas de la NMF se pose donc. En effet, on peut considérer qu'il s'agit d'une contrainte utile : intuitivement par exemple dans le cas du piano, le clavier comporte 88 touches et les mains du pianiste 10 doigts, le nombre de notes simultanément actives est donc relativement limité par rapport au nombre d'atomes dans la base. Cependant, c'est une contrainte à double tranchant car la représentation la plus parcimonieuse d'une suite d'accords est celle qui consiste à représenter chaque accord comme un atome élémentaire (si l'ordre K le permet), la décomposition à chaque trame ne comportant alors qu'un seul coefficient non nul. D'autre part, l'usage possible de la pédale *forte*, ou d'effets comme la réverbération sur des claviers numériques, augmente bien au-delà de 10 le nombre de notes pouvant être simultanément entendues.

En ce qui concerne la contrainte de parcimonie dans le cadre de la NMF, [Hoyer, 2004] définit une mesure S de la parcimonie d'une ligne h_k comme :

$$S(h_k) = \frac{\sqrt{N} - \left(\sum_{n=1}^{N} |h_{kn}| / \sqrt{\sum_{n=1}^{N} h_{kn}^2} \right)}{\sqrt{N} - 1} \tag{V.2}$$

c'est-à-dire, le rapport entre les normes L^1 et L^2 du vecteur, normalisé entre 0 et 1. Il propose ensuite un algorithme de descente de gradient résolvant le problème (III.1) sous les contraintes d'égalités $\forall k \quad S(h_k) = S_h$, la quantité S_h étant fixée et mesurant donc le degré souhaité de parcimonie. Suivant une approche sensiblement différente, [Eggert et Körner, 2004] propose l'usage d'un terme de pénalité renforçant la parcimonie, qu'il choisit simplement comme la norme L^1 des vecteurs h_k (ce qui est une approche classique du problème de codage parcimonieux (V.1)). [Virtanen, 2007] propose également l'usage d'un critère de parcimonie dans la NMF (utilisée dans un contexte de séparation de sources monocapteur), exprimé sous forme d'un terme de pénalité, qui est une fonction des lignes h_k normalisées par leur variance. Son étude conclut à l'inefficacité de cette contrainte dans le cas de la séparation de sources monocapteur.

V.2.2 Régularité

Dans le problème de NMF standard et dans la plupart de ses variantes, les trames temporelles successives sont considérées comme des observations indépendantes et sans relation, ce qui n'est évidemment pas vrai pour les sons naturels, et en particulier les sons musicaux. Dans le cas d'une note de musique, la plus grande partie de la note (les phases de décroissance, entretien et extinction dans le schéma d'enveloppe dit « ADSR », *cf.* section I.2.2.3 page 21) possède un spectre variant lentement dans le temps. Lorsqu'on exprime ce spectre sous la forme d'un produit entre un spectre $\mathbf{w}_k$ et un gain variable dans le temps h_k, comme cela est fait dans le modèle NMF, cela revient à dire que la ligne h_k est régulière, ou, en d'autres termes, que les coefficients h_{kn} et $h_{k(n-1)}$ ne diffèrent pas « trop » l'un de l'autre. Partant de cette considération, [Virtanen, 2007] et [Chen *et al.*, 2006] introduisent dans la fonction de coût un terme de pénalité prenant en compte cette continuité temporelle. Dans [Virtanen, 2007], ce terme est directement exprimé en fonction des différences $h_{kn} - h_{k(n-1)}$, tandis que [Chen *et al.*, 2006] utilise le quotient entre les variances à court et long terme de h_k. Ces deux termes s'avèrent favorables à la régularité des lignes de $\mathbf{H}$. Une autre approche possible est l'approche statistique issue de [Févotte *et al.*, 2009], qui consiste à poser une distribution a priori sur $\mathbf{H}$ pour obtenir la continuité temporelle, et que nous exposerons en détail au chapitre VII (page 105).

Il est intéressant de remarquer que la non-régularité peut également être un objectif, voir par exemple [Pascual-Montano *et al.*, 2006], suivant les données et l'application. [Pascual-Montano *et al.*, 2006] souligne que la régularité d'un des facteurs (*i.e.* $\mathbf{W}$ ou $\mathbf{H}$) peut renforcer la parcimonie de l'autre facteur, ce qui établit un lien entre ces deux contraintes populaires. À l'inverse, [Virtanen, 2007] combine à la fois les contraintes de parcimonie et de régularité temporelle, mais conclut à la non efficacité de la parcimonie dans le cas particulier traité.

V.2.3 Décorrélation

Pour opérer une réduction de rang pertinente, et éviter de séparer en plusieurs composantes une unité faisant sens, il peut sembler intéressant de chercher à produire des lignes de $\mathbf{H}$ les plus décorrélées possibles. [Li *et al.*, 2001] propose de réaliser une forme affaiblie de cette contrainte, sous la forme d'un terme de pénalité dépendant du produit $\mathbf{H}\mathbf{H}^T$. En effet, renforcer la diagonalité de cette matrice revient intuitivement à faire diminuer la corrélation entre les lignes de $\mathbf{H}$. [Li *et al.*, 2001] et [Raczyński *et al.*, 2007] proposent de considérer les deux quantités :

$$D_{c1}(h_k) = \sum_{j \neq k} \left[\mathbf{H}\mathbf{H}^T\right]_{kj}, \tag{V.3}$$

dépendant des termes non-diagonaux de la matrice de covariance $\mathbf{H}\mathbf{H}^T$, à minimiser, et :

$$D_{c2}(h_k) = \sum_{k} \left[\mathbf{H}\mathbf{H}^T\right]_{kk}, \tag{V.4}$$

termes diagonaux de la même matrice, à maximiser.

[Zhang et Fang, 2007] propose un terme similaire, mais normalisé et faisant intervenir le logarithme des coefficients de $\mathbf{H}\mathbf{H}^T$ et non les coefficients eux-mêmes. La méthode est appliquée à un problème de séparation de sources multi-capteur surdéterminé ($K > F$) et s'avère efficace même lorsque les sources ne sont pas statistiquement indépendantes. Ce travail est à rapprocher de [Zhang et Fang, 2007], qui met au point un algorithme (non multiplicatif) nommé *non-negative least correlated component*

analysis (nLCA) dans un cas identique et en s'appuyant sur une quantité très similaire, mais avec une méthodologie sensiblement différente, inspirée de la géométrie.

Par la contrainte de décorrélation, on espère éviter que plusieurs composantes aient exactement la même enveloppe temporelle, ce qui laisserait penser qu'elle décrivent un même objet sonore qui gagnerait à être représenté par un seul atome (typiquement, séparation des harmoniques d'une même note). Toutefois, les règles de composition de l'harmonie classique peuvent également laisser craindre une certaine corrélation entre les occurrences de notes consonantes ; par exemple dans la partie grave de la tessiture au piano, où il est fréquent de retrouver les mêmes accords plaqués régulièrement. Dans ce cas, la contrainte de décorrélation risque d'amener à regrouper au sein d'une même composante des notes souvent jouées simultanément.

V.2.4 Harmonicité

Comme on a pu l'observer dans la section I.2.2 (page 13), les spectres des parties quasi-stationnaires des notes de musiques sont harmoniques ou quasi-harmoniques, ce qui fonde la notion même de fréquence fondamentale, et, naturellement, de nombreuses méthodes d'estimation de celle-ci. Il paraît donc logique de chercher à représenter le signal sur un dictionnaire dont les éléments présentent cette même propriété. Si, dans le domaine plus général de design de dictionnaire, et en particulier dans le domaine des représentations parcimonieuses (*cf.* section V.2.1), cette propriété est largement recherchée (voir par exemple l'extension des atomes de Gabor proposées dans [Gribonval et Bacry, 2003, Leveau, 2007]), en revanche peu de travaux consacrés à la NMF cherchent à intégrer cette propriété à la base $\mathbf{W}$.

À notre connaissance, [Raczyński *et al.*, 2007] est le seul travail antérieur au nôtre proposant une variante harmonique de la NMF. Dans cette approche, la base $\mathbf{W}$ est directement initialisée comme un dictionnaire de spectres harmoniques (sur l'échelle tempérée), les coefficients correspondant à des fréquences hors du peigne théorique étant mis à zéro. Ceci interdit une forte évolution du dictionnaire, qui n'a que peu de chances de s'adapter réellement aux données, en particulier si l'instrument est accordé différemment des fréquences théoriques. De plus, il existe un risque de dégénérescence de certaines composantes vers des notes en rapport harmonique (typiquement, les octaves supérieures). [Niedermayer, 2008] propose également de forcer l'harmonicité du dictionnaire en l'apprenant sur une base de notes isolées et en interdisant sa modification, ce qui en fait une méthode de « semi-NMF », non adaptative.

V.2.5 Autres

D'autres contraintes sont proposées dans la littérature. Dans le domaine de l'image, des termes de pénalité sont proposés pour produire des bases « localisées », c'est-à-dire dont les coefficients non-nuls sont regroupés sur l'image : [Li *et al.*, 2001] introduit trois termes de pénalité que l'on peut relier aux contraintes de parcimonie et de décorrélation, tandis que [Wang *et al.*, 2004] s'appuie sur la notion d'analyse discriminante de Fisher, qui vise à maximiser la distance inter-classe et minimiser la distance intra-classe.

Dans l'idée de se rapprocher d'autres méthodes de décompositions des signaux et notamment la PCA, [Choi, 2008] propose une variante orthogonale de la NMF, c'est-à-dire une NMF sous la contrainte $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ ou, au choix, $\mathbf{H}\mathbf{H}^T = \mathbf{I}$ suivant les besoins de l'application. Le premier cas n'est pas pertinent en musique, en raison des recouvrements temps-fréquence des notes ; le second est à rapprocher de la contrainte de décorrélation (section V.2.3), sous une forme normalisée. L'algorithme

qu'il propose est directement dérivé des conditions KKT, ce qui peut s'interpréter comme l'ajout d'un terme de pénalité sous la forme d'une fonction des multiplicateurs de Lagrange.

Enfin, signalons une dernière contrainte, l'invariance des bases par translation (*shift-invariance*), utilisée par exemple dans [Kim et Choi, 2006]. Elle consiste à considérer que les spectres des notes d'un même instrument de musique sont identiques à une translation près dans un domaine fréquentiel bien choisi (typiquement, la transformée à Q constant, *cf.* section IX.2.1 page 128), de manière à réduire le nombre de dimensions du problème. Cette contrainte est principalement répandue dans les articles cherchant à résoudre des problèmes de NMF convolutive [Blumensath et Davies, 2006, Smaragdis, 2004, Shashanka *et al.*, 2008a], qui permettent de mieux modéliser l'évolution du spectre d'une note au cours de son déroulement (par exemple au cours d'un *vibrato*) en échange d'une moindre souplesse sur la modélisation des enveloppes spectrales, et que nous ne traiterons pas ici.

V.3 NMF harmonique

Parmi les contraintes évoquées à la section précédente, l'harmonicité paraît être à la fois une contrainte particulièrement pertinente pour l'analyse de signaux musicaux, et peu exploitée dans le cadre de la NMF, certainement car elle n'a pas d'équivalent dans le domaine du traitement de l'image, d'où sont issues nombre des variantes que nous avons présentées. Nous nous y intéressons plus particulièrement ici. Ce travail en collaboration a donné lieu à publication [Vincent *et al.*, 2008] et a été évalué avec un bon classement dans la compétition internationale MIREX [1] [Vincent *et al.*, 2007].

V.3.1 Motivation

Nous avons supposé depuis le début, après observation sur des exemples simples, la capacité de la NMF à extraire des composantes présentant une structure harmonique. Cependant, rien dans le calcul ne l'impose, et on peut s'attendre dans des cas plus complexes (sons réels, polyphonie plus élevée) à ce que certains des spectres de la base ne possèdent pas clairement de hauteur.

Un moyen de forcer l'harmonicité de la base serait de forcer chacune de ses composantes à s'exprimer comme une combinaison linéaire de spectres $P(f)$ ne possédant qu'un pic fréquentiel, les pics constituant la composante totale étant tous situés à des fréquences multiples d'une même fondamentale :

$$\mathbf{w}_k = \sum_{m=1}^{M} e_{mk} P(m f_0(k)) \tag{V.5}$$

Il ne resterait plus qu'à incorporer ce modèle sous forme d'un problème de NMF dans lequel les coefficients à estimer seraient les coefficients e_{mk} plutôt que les bases w_{fk}. Cette méthode a cependant un défaut : si par exemple tous les coefficients e_{mk} pour m impair tendent vers 0, la composante peut dégénérer en une composante à l'octave (fréquence fondamentale double). À l'inverse, on peut fixer d'emblée $\mathbf{W}$ comme un dictionnaire de spectres harmoniques, mais cela limiterait considérablement les capacités d'adaptation aux données (seuls les coefficients $\mathbf{H}$ seraient appris, ce qui conduit à une « semi-NMF »). Un autre solution pourrait être de simplement initialiser $\mathbf{W}$ sous cette forme, puis de laisser

1. L'ensemble des résultats de l'évaluation est disponible à l'adresse :
`http://www.music-ir.org/mirex/2007/index.php/Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results`

l'algorithme NMF se dérouler sans contrainte, ce qui ne garantirait cependant pas l'harmonicité du dictionnaire à l'issue de la factorisation. Pour éviter ces écueils, nous proposons un modèle intermédiaire qui assure l'harmonicité tout en préservant une certaine capacité d'adaptation aux données.

V.3.2 Modèle génératif

Nous imposons aux composantes de la base de s'exprimer comme une combinaison linéaire de spectres harmoniques à bande étroite (*patterns*) qui sont fixés arbitrairement :

$$w_{fk} = \sum_{m=1}^{M} e_{mk} P_{km}(f) \tag{V.6}$$

Les spectres P_{km} sont constitués de quelques partiels adjacents à des fréquences multiples d'une même fondamentale f_0 associée à l'indice k. Ces spectres sont modulés par l'enveloppe spectrale de la sous-bande k. Cette enveloppe spectrale est choisie en suivant une modélisation perceptuelle [Vincent *et al.*, 2008]. La figure V.1 illustre ces patterns pour une note donnée, ainsi que la composante résultante $\mathbf{w}_k$. Ainsi exprimées, les colonnes de $\mathbf{W}$ n'ont plus la possibilité de dégéner vers une octave ou un autre multiple de la fréquence fondamentale de départ ; les coefficients de pondération entre les différents patrons assurent l'adaptabilité du dictionnaire à l'enveloppe spectrale des sons analysés.

V.3.3 Patterns

Dans la suite, nous supposons que les bandes f sont espacées linéairement sur l'échelle ERB (voir section IX.2.1 page 128), définie par :

$$f_{\mathrm{ERB}} = 9.26 \log(0.00437 f_{\mathrm{Hz}} + 1) \tag{V.7}$$

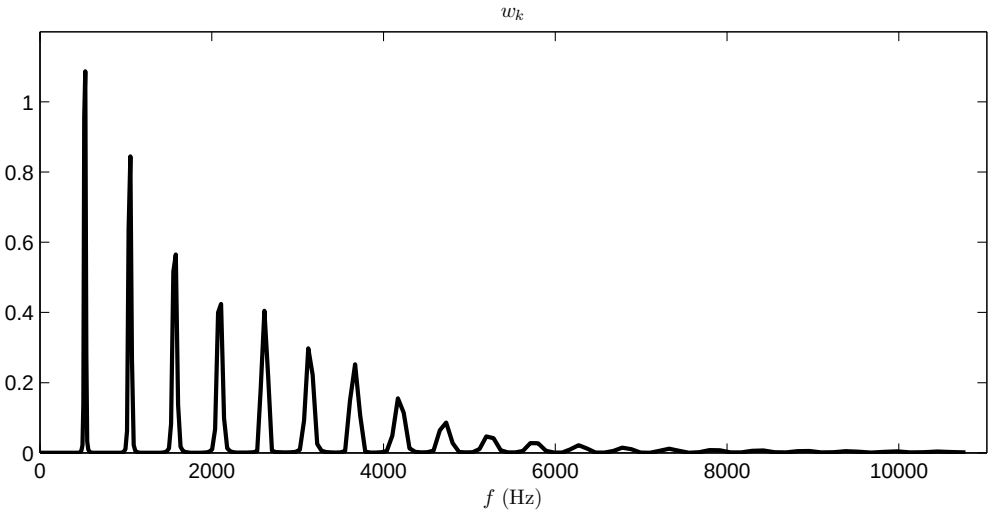
Les fréquences centrales des bandes sont séparées par un pas de Δf ERB. La première bande est centrée en f_0^i et le nombre F de bandes est calculé de manière que la fréquence centrale de la bande la plus haute soit en dessous de la fréquence de Nyquist, avec un nombre maximal de M_{max} sous-bandes pour décrire la composante $\mathbf{w}_k$. Nous définissons un patron P_{km} comme le produit d'un spectre harmonique dont tous les partiels ont une amplitude unitaire, et de la réponse fréquentielle du filtre gammatone [Patterson *et al.*, 1991] de largeur de bande Δf qui modélise la sous-bande f. La réponse fréquentielle d'un filtre gammatone s'écrit :

$$\Gamma(f) = \frac{1}{\left(1 + \frac{(f-f_c)^2}{\gamma^2}\right)^{n/2}} \tag{V.8}$$

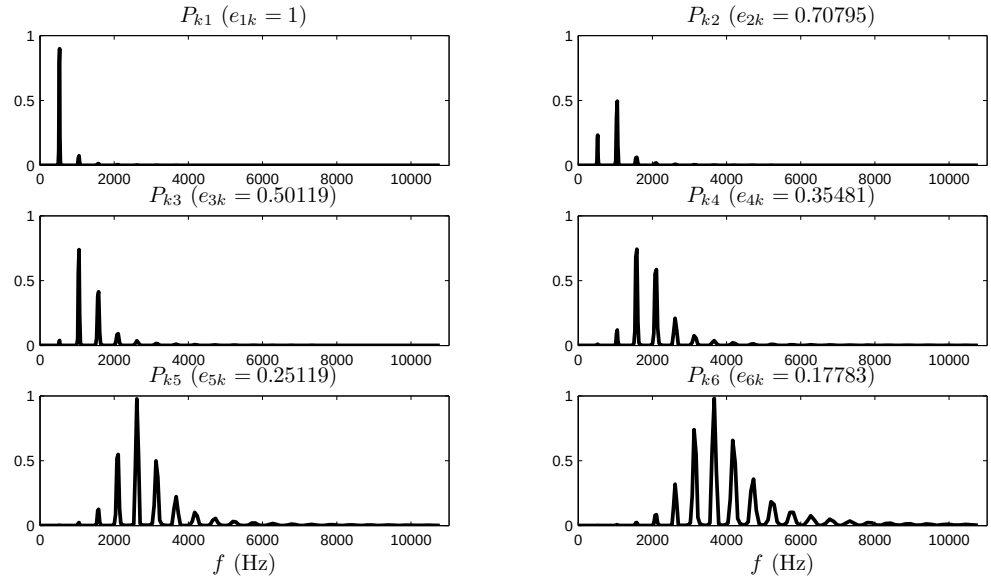
où le coefficient γ est choisi de manière à obtenir une largeur de bande d'un ERB, f_c est la fréquence centrale du filtre et n son ordre (souvent choisi égal à 4). Ce filtre a une réponse impulsionnelle finie, ce qui permet une implantation simple.

Un modèle similaire a été utilisé dans [Virtanen et Klapuri, 2002] dans un contexte de séparation de sources, dans lequel les fréquences fondamentales des notes sont connues mais au cours duquel les enveloppes spectrales sont adaptées séparément à chaque trame temporelle.

La figure V.1 illustre un exemple de patrons pour la note do_4 du piano, avec $M_{max} = 6$ et $\Delta f = 1.75$.



(a) Composante $\mathbf{w}_k$



(b) Patterns correspondants P_{km}

FIGURE V.1 – Exemple de spectre harmonique de base $\mathbf{w}_k$ correspondant à la note do_4 ($C4$, MIDI 72), avec les spectres à bande étroite P_{km} le composant et les coefficients d'enveloppe spectrale e_{mk} (avec $M_{max} = 6$).

La largeur des sous-bandes peut affecter les performances de transcription. Lorsque le spectre à bande étroite $P_{km}(f)$ contient un seul partiel, la base résultante $\mathbf{w}_k$ peut représenter des multiples de la fréquence fondamentale attendue, ce qui déclenche des erreurs d'octave. À l'inverse, si le patron contient trop de partiels, les spectres de bases risquent de ne pas pouvoir s'adapter à l'enveloppe spectrale des instruments transcrits, ce qui peut conduire à la détection de notes erronnées ou à l'oubli de certaines notes.

Ce modèle est pertinent d'un point de vue cognitif : en effet, la perception de hauteur est basée en partie sur la détection de périodicités dans chaque bande auditive [Klapuri et Davy, 2006].

V.3.4 Paramètres supplémentaires

L'ajout de ce modèle nous amène à nous poser des questions nouvelles par rapport à la NMF sans contrainte sur le dictionnaire. En effet, les fréquences des partiels varient suivant le signal analysé, d'une part en raison du *diapason du morceau, d'autre part en raison d'une possible inharmonicité des sons, par exemple s'il s'agit de piano. Dans [Vincent *et al.*, 2007, Vincent *et al.*, 2008], une prise en compte de ces deux phénomènes est proposée. On considère trois modèles possibles pour la fréquence fondamentale des notes. Le premier considère un accord au diapason et suivant exactement la gamme tempérée :

$$f_0 = 440 \times 2^{\frac{p-69}{12}}. \tag{V.9}$$

On considère ensuite un possible désaccord du diapason d'un facteur $q \in [-\frac{1}{2}, \frac{1}{2})$ (plus ou moins un demi-ton), mais en conservant les rapports de fréquences de la gamme tempérée :

$$f_0 = 440 \times 2^{\frac{p+q-69}{12}}. \tag{V.10}$$

Enfin, on envisage un possible désaccord $q_p \in [-\frac{1}{2}, \frac{1}{2})$ dépendant du pitch [Ortiz-Berenguer *et al.*, 2004]

$$f_0 = 440 \times 2^{\frac{p+q_p-69}{12}}. \tag{V.11}$$

Par ailleurs, les fréquences des partiels sont supposés soit parfaitement harmoniques :

$$f_l = (l+1) f_0 \tag{V.12}$$

soit inharmoniques, avec un coefficient d'inharmonicité $b_p \geq 0$ [Fletcher et Rossing, 1998] :

$$f_l = l f_0 \sqrt{\frac{1 + b_p l^2}{1 + b_p}}. \tag{V.13}$$

L'optimisation de ces paramètres et l'étude de l'impact des différents modèles sur la performance de transcription ont été réalisées par Emmanuel Vincent et publiées dans [Vincent *et al.*, 2008]. Nous ne développerons pas ce point ici.

V.3.5 Algorithme

Les coefficients e_{mk} sont appris par NMF tout comme la décomposition $\mathbf{H}$. Les règles de mise à jour sont obtenues en minimisant la même fonction de coût qu'en NMF standard, à ceci près qu'elles sont minimisées par rapport à $\mathbf{E}$ et $\mathbf{H}$ plutôt que $\mathbf{W}$ et $\mathbf{H}$.

V.3.5.1 Distance euclidienne pondérée

Dans cette variante, les paramètres du modèle sont appris en minimisant la sonie du résiduel, mesurée comme une norme euclidienne pondérée [Vincent et Plumbley, 2007] qui s'écrit :

$$D_{WEUC} = \sum_{f,n} g_{fn} \left[\mathbf{V} - \mathbf{WH}\right]_{fn}^2, \tag{V.14}$$

où les poids perceptifs g_{fn} dépendent de l'énergie du coefficient v_{fn}. Ce critère améliore la transcription des notes peu énergétiques, comparée à la norme euclidienne habituelle [Vincent *et al.*, 2007].

Les amplitudes h_{kn} et les coefficients d'enveloppe e_{mk} sont initialisés au hasard et mis à jour alternativement jusqu'à ce qu'un minimum local de D_{WEUC} soit atteint. Puisque le modèle demeure linéaire par rapport à ces coefficients, les mises à jour peuvent être obtenues sous forme multiplicative en utilisant l'équation (III.25) (page 59) :

$$h_{kn} \leftarrow h_{kn} \frac{\sum_f w_{fk} g_{fn} v_{fn}}{\sum_f w_{fk} g_{fn} \sum_j w_{fj} h_{jn}} \tag{V.15}$$

$$e_{mk} \leftarrow e_{mk} \frac{\sum_{f,n} h_{kn} P_{km}(f) g_{fn} v_{fn}}{\sum_{f,n} h_{kn} P_{km}(f) g_{fn} \sum_j w_{fj} h_{jn}}. \tag{V.16}$$

Nous dénotons cet algorithme par l'acronyme « **HEUC-NMF/MU** ».

V.3.5.2 Divergence d'Itakura-Saito

Le modèle peut s'appliquer de la même manière à la minimisation de la divergence IS. Le critère à minimiser s'écrit :

$$D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{IS}(v_{fn}|\sum_{k=1}^{K} w_{fk}h_{kn}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{IS}(v_{fn}|\sum_{k=1}^{K}\sum_{m=1}^{M} h_{kn}e_{mk}P_{km}(f)) \tag{V.17}$$

Nous calculons sa dérivée partielle par rapport au coefficient h_{kn}. Elle s'écrit comme la différence de deux termes positifs :

$$\nabla_{h_{kn}} D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fn}} - \sum_{f=1}^{F} \frac{v_{fn} w_{fk}}{\hat{v}_{fn}^2} \tag{V.18}$$

où $\hat{v}_{fn} = \sum_{k'=1}^{K} w_{fk'} h_{k'n} = \sum_{k'=1}^{K}\sum_{m'=1}^{M} e_{m'k'} P_{k'm'}(f) h_{k'n}$. La dérivée partielle par rapport à e_{mk} s'exprime sous la même forme :

$$\nabla_{e_{mk}} D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} \frac{h_{kn}P_{km}(f)}{\hat{v}_{fn}} - \sum_{f=1}^{F}\sum_{n=1}^{N} \frac{v_{fn}h_{kn}P_{km}(f)}{\hat{v}_{fn}^2} \tag{V.19}$$

Nous pouvons en déduire des règles de mises à jour obtenues par la règle (III.25) (page 59) :

$$h_{kn} \leftarrow h_{kn} \times \frac{\sum_f v_{fn} w_{fk} / \hat{v}_{fn}^2}{\sum_f w_{fk} / \hat{v}_{fn}} \tag{V.20}$$

$$e_{mk} \leftarrow e_{mk} \times \frac{\sum_{f,n} v_{fn} h_{kn} P_{km}(f) / \hat{v}_{fn}^2}{\sum_{f,n} h_{kn} P_{km}(f) / \hat{v}_{fn}} \tag{V.21}$$

Dans la suite, cet algorithme sera désigné par l'acronyme « **H-NMF/MU** » et testé au chapitre X.

Des travaux prolongeant ce modèle, et s'intéressant en particulier à la possibilité d'apprendre les enveloppes des spectres de sous-bandes, ont été menés par Emmanuel Vincent. Notre contribution à ce travail étant mineure, nous n'en reproduisons pas ici les résultats.

Troisième partie

Approche probabiliste

Chapitre VI

De la NMF à l'approche probabiliste

Résumé

Où l'on examine les liens entre NMF et modélisation probabiliste des signaux audio, propose un modèle et démontre son équivalence avec l'IS-NMF, et développe un algorithme d'estimation du maximum de vraisemblance résolvant le problème de NMF dans ce modèle.

VI.1 Introduction

L'usage d'approches statistiques pour modéliser le signal audio et pour s'atteler à la tâche de transcription musicale n'est pas nouveau, comme on l'a vu dans le chapitre II. Le rapprochement entre le cadre statistique et la contrainte de non-négativité est cependant beaucoup plus récent. Il commence par l'importation de la contrainte de non-négativité dans des cadres statistiques plus anciens, pour se poursuivre par des interprétations statistiques de la NMF, jusqu'à ouvrir de nouvelles perspectives algorithmiques pour la résolution du problème de NMF. Un renversement conceptuel apparaît ainsi dans la littérature : alors que les premiers travaux considèrent l'idée d'utiliser la nature non-négative des données pour aider à leur traitement via des méthodes statistiques, les travaux ultérieurs établissent des interprétations, puis des équivalences formelles entre des problèmes d'estimation et de NMF, pour finir par proposer d'utiliser la modélisation et les outils statistiques dans le but de résoudre le problème de NMF. Il s'opère alors un glissement d'une problématique de représentation des données (NMF déterministe) à une problématique davantage orientée « séparation de sources » (NMF probabiliste).

La méthodologie d'analyse du signal audio par modélisation statistique spécifiquement considérée dans cette partie s'appuie sur divers principes généraux :

- Le signal observé $\mathbf{X}$ est considéré comme la réalisation d'une variable aléatoire $\mathbf{x}$;
- Il est modélisé de manière plus ou moins fine, soit directement via un modèle (loi de probabilité), soit à travers d'autres variables aléatoires, dites *latentes* dont la distribution est modélisée ;
- Les développements visent à obtenir une méthode d'estimation des paramètres $\boldsymbol{\theta}$ du modèle ;
- Cette estimation peut être obtenue par la méthode du *maximum de vraisemblance* (MV), qui consiste à maximiser, par rapport à $\boldsymbol{\theta}$, la vraisemblance $p(\mathbf{X}|\boldsymbol{\theta})$ (probabibilité des observations sachant le modèle), ou, le plus souvent, son logarithme, pour des raisons de facilité calculatoire ;
- Elle peut également être obtenue par la méthode du *maximum a posteriori* (MAP), reliée à la précédente en vertu du théorème de Bayes. Elle consiste à maximiser la distribution a posteriori $p(\boldsymbol{\theta}|\mathbf{X})$ du modèle sachant les observations. Cette inversion, dite « bayésienne », permet en outre d'inclure dans la modélisation des connaissances a priori sur les paramètres à estimer, via la distribution a priori $p(\boldsymbol{\theta})$.

Le chapitre est organisé comme suit : la section VI.2 présente un bref état de l'art des modèles statistiques prenant en compte la non-négativité des données, d'une part par l'usage de la contrainte de non-négativité (section VI.2.1), d'autre part en établissant un lien plus ou moins explicite avec le problème de NMF (section VI.2.2). Dans la section VI.3, nous présenterons le modèle que nous avons retenu et établirons son équivalence avec le problème de NMF par minimisation de la distance d'Itakura-Saito. On en déduit dans la section VI.4 un algorithme de type Espérance-Maximisation (EM) qui résout le problème standard de NMF par une estimation du maximum de vraisemblance.

VI.2 État de l'art

VI.2.1 Contrainte de non-négativité dans les approches statistiques

Le rapprochement entre les travaux sur la NMF et l'approche statistique commence par l'« importation » de la contrainte de non-négativité dans des méthodes usuelles de décomposition des signaux dans un formalisme statistique.

À notre connaissance, [Plumbley, 2002] est le premier à s'intéresser à l'analyse en composantes

indépendantes (ICA) de données non-négatives. L'ICA est une technique de réduction de dimensionnalité particulièrement populaire dans le domaine de la séparation aveugle de sources (BSS) ; on en trouvera une présentation très complète par exemple dans [Cardoso, 1998]. Dans le paradigme de la BSS, on observe à chaque trame temporelle F mélanges $[z_1(n) \dots z_F(n)]$ de K variables aléatoires appelées *sources*, $[s_1(n) \dots s_K(n)]$. Les mélanges et les sources sont liés par une matrice de mélange $\mathbf{A}$ invariante et supposée de rang plein, telle que :

$$\mathbf{z} = \mathbf{A}\mathbf{s} \tag{VI.1}$$

où $\mathbf{z} = [z_1 \dots z_F]^T \in \mathbb{R}^F$, $\mathbf{s} = [s_1 \dots s_K]^T \in \mathbb{R}^K$ et $\mathbf{A} \in \mathbb{R}^{F \times K}$. La tâche est dite « sous-déterminée » lorsque $K < F$ (on observe moins de mélanges que de sources).

Cette notation établit clairement un parallèle avec la NMF, mais il faut cependant remarquer que dans le paradigme original, les sources et les mélanges observés évoluent dans le domaine temporel, et la séparation exploite une diversité spatiale (séparation multi-capteurs). Pour mener l'analogie, on pourrait considérer que le domaine transformé dans lequel évolue $\mathbf{V}$ introduit une diversité fréquentielle, les points fréquentiels étant considéré comme des « capteurs ».

[Plumbley, 2002] étudie l'opportunité d'une contrainte de non-négativité dans l'ICA sur le plan théorique uniquement, et suggère la possibilité de développer des algorithmes, ce qui est fait dans [Plumbley, 2003, Oja et Plumbley, 2003, Plumbley, 2004]. Dans ces travaux, une équivalence est établie entre l'indépendance statistique des sources (argument usuel pour résoudre le problème standard d'ICA), et l'affaiblissement de cette hypothèse en hypothèse de décorrélation des sources dans le cas où celles-ci sont non-négatives, ce qui rapproche l'ICA et la PCA (analyse en composantes principales, où les composantes extraites sont décorrélées en terme de variance, mais non indépendantes statistiquement). Cet affaiblissement de l'hypothèse est à rapprocher des variantes de la NMF sous contrainte de décorrélation dans le cadre déterministe (*cf.* section V.2.3, page 82). Enfin, il est à noter que l'hypothèse de non-négativité ne concerne que les sources, et non la matrice de mélange. En ce sens, l'ICA non-négative peut être considérée comme une « semi-NMF » sous contrainte de décorrélation.

Par ailleurs, dans la même mouvance, ces idées sont utilisées dans un cadre connexe, celui du codage parcimonieux [Abdallah et Plumbley, 2004, Abdallah et Plumbley, 2006], la parcimonie (*cf.* section V.2.1, page 80) étant imposée par un a priori sur la distribution des sources (supposée laplacienne). C'est aussi dans le cadre de ce travail que la non-négativité associée à un cadre probabiliste est appliquée à des spectres de puissance de signaux audio dans une tâche d'analyse de musique polyphonique.

La contrainte de non-négativité imposée aux tâches de codage parcimonieux est également largement appliquée dans le domaine de l'image (voir par exemple [Shang, 2008]).

VI.2.2 Interprétations probabilistes de la NMF

Le pont entre NMF et cadre statistique s'établit véritablement dans la littérature lorsqu'un modèle (probabiliste) de l'observation (à valeurs non-négatives) est posé, et qu'il est établi une équivalence formelle entre la minimisation de la fonction de coût dans le problème de NMF et l'estimation des paramètres du modèle dans le problème statistique. Ces équivalences sont recensées de manière synthétique dans [Févotte et Cemgil, 2009].

Par exemple, [Virtanen *et al.*, 2008] pose le modèle suivant sur le spectrogramme d'amplitude $|\mathbf{X}|$:

$$|\mathbf{x}_n| = \sum_{k=1}^{K} |\mathbf{c}_{kn}| \tag{VI.2}$$

sous l'hypothèse que chaque variable latente $\mathbf{c}_{kn}(f)$ est distribuée suivant une loi de Poisson généralisée (notée $\mathcal{P}$) :

$$|\mathbf{c}_{kn}(f)| \sim \mathcal{P}(\mathbf{c}_{kn}(f)|w_{fk}h_{kn}) \tag{VI.3}$$

$$\mathcal{P}(u|\lambda) = \exp(-\lambda)\,\frac{\lambda^u}{u!} \tag{VI.4}$$

La somme de variables aléatoires poissonniennes étant elle-même distribuée suivant une loi de Poisson, il en découle que $|x_{fn}| \sim \mathcal{P}(\sum_{k=1}^{K} w_{fk}h_{kn})$. La log-vraisemblance $-\log p(\mathbf{X}|\mathbf{W},\mathbf{H})$ s'écrit donc simplement grâce à l'expression de la loi de Poisson (VI.4), et on vérifie qu'elle est égale, à une constante près, à la distance de Kullback-Leibler $D_{KL}(\,|\mathbf{X}|\,|\,\mathbf{W}\,\mathbf{H})$ (III.5). Ceci établit l'équivalence entre estimation du MV dans le modèle (VI.2) et résolution de la KL-NMF. Après factorisation, les composantes estimées sont formées en utilisant la phase des observations [Virtanen, 2007], de manière que

$$\hat{c}_{kfn} = w_{fk}\,h_{kn}\,\arg(x_{fn}), \tag{VI.5}$$

où $\arg(x)$ désigne la phase du scalaire complexe x.

Cette approche mérite quelques commentaires. D'une part, la distribution de Poisson est originellement définie seulement sur des entiers, ce qui altère la possibilité d'interpréter statistiquement la KL-NMF de données non dénombrables telles que les spectres audio (on pourrait cependant envisager une mise à l'échelle appropriée et une quantification très fine pour réduire ce problème). D'autre part, cette approche contraint la non-négativité d'une manière relativement arbitraire, en prenant la valeur absolue de $\mathbf{X}$. La méthode de reconstruction force les composantes à posséder la même phase que les observations et la reconstruction des composantes n'est ni fondée statistiquement, ni conservative, *i.e.* $\mathbf{x}_n \approx \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}$. Notons au passage que la recontruction de Wiener est utilisée dans le problème de KL-NMF du spectrogramme d'amplitude $|\mathbf{X}|$ par [Smaragdis, 2007], qui le présente comme un filtrage dans le domaine spectral et qui met en exergue le caractère conservatif de l'approche.

Un autre exemple d'approche statistique de la résolution du problème de NMF et de ses approches contraintes peut être trouvé dans [Schmidt et Laurberg, 2008]. Dans le modèle proposé, les facteurs $\mathbf{W}$ et $\mathbf{H}$ sont exprimés comme deux fonctions f_h et f_w (dites « fonctions liantes ») de variables latentes gaussiennes, avec des hypothèses relativement faibles sur ces deux fonctions. Cette approche peut être vue comme une généralisation de celle que nous développerons dans ce chapitre, pour des choix appropriés de f_h et f_w.

Dans le contexte bien différent de l'analyse de données textuelles, [Ding *et al.*, 2006,Ding *et al.*, 2008] démontre une équivalence entre la NMF de la matrice de co-occurrence et son « analyse sémantique latente probabiliste » (PLSA, également nommée PLSI pour *Probabilistic Latent Semantic Indexing*, suivant le contexte), une technique statistique d'apprentissage non supervisé introduite dans [Hofmann, 1999]. Une discussion détaillée de l'équivalence entre PLSI et NMF pourra être consultée dans [Rigouste, 2006, Gaussier et Goutte, 2005].

VI.3 Modèle et équivalence

VI.3.1 Somme de Gaussiennes

Dans le reste de ce chapitre, on considère plus particulièrement le modèle génératif suivant : $\forall n = 1, \ldots, N$

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{kn} \in \mathbb{C}^{F \times 1} \tag{VI.6}$$

où les variables latentes $\mathbf{c}_{kn}$ sont distribuées suivant une loi gaussienne :

$$\mathbf{c}_{kn} \sim \mathcal{N}_c(0, h_{kn}\ \mathrm{diag}\,\mathbf{w}_k), \tag{VI.7}$$

$\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est la distribution gaussienne complexe multivariée de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$:

$$\mathcal{N}_c\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \det(\pi\, \boldsymbol{\Sigma}^{-1})\, \exp -(\mathbf{x} - \boldsymbol{\mu})^H\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{x} - \boldsymbol{\mu}) \tag{VI.8}$$

Les composantes $\mathbf{c}_{1n}, \ldots, \mathbf{c}_{Kn}$ sont supposées mutuellement indépendantes, et à k fixé les $\mathbf{c}_{kn}$ sont indépendantes et identiquement distribuées (iid). Nous noterons désormais $\mathbf{V}$ la matrice de coefficients : $v_{fn} = |x_{fn}|^2$.

Le modèle génératif (VI.6) a été introduit par [Benaroya *et al.*, 2003, Benaroya *et al.*, 2006] dans un cadre de séparation de sources audio mono-capteur. Dans ce contexte, l'observation $\mathbf{x}_n = [x_{1n}, \ldots, x_{Fn}]^T$ est la transformée de Fourier à court-terme (TFCT) d'un signal audionumérique x, $n = 1, \ldots, N$ étant l'indice de trame temporelle et $f = 1, \ldots, F$ l'indice de point fréquentiel. Les auteurs supposent que le signal $\mathbf{x}$ résulte de la somme de deux sources, $\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2$ et les TFCT de ces sources sont modélisées par :

$$\mathbf{s}_{1n} = \sum_{k=1}^{K_1} \mathbf{c}_{kn} \quad \texttt{et} \quad \mathbf{s}_{2n} = \sum_{k=K_1+1}^{K_1+K_2} \mathbf{c}_{kn}, \tag{VI.9}$$

avec $K_1 + K_2 = K$. Ainsi, la TFCT de chaque source est modélisée comme la somme de composantes élémentaires caractérisées par leur densité spectrale de puissance (DSP) $\mathbf{w}_k$, modulée dans le temps par des coefficients d'activation dépendants du temps (enveloppes temporelles) h_{kn}. Les DSPs caractérisant chaque source sont apprises sur des données d'entraînement, puis le spectrogramme résultant du mélange, $|\mathbf{X}|^2$, est décomposé sur le dictionnaire (désormais connu) $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{K_1}, \mathbf{w}_{K_1+1}, \ldots, \mathbf{w}_{K_1+K_2}]$. Cependant, dans ces articles, les DSPs et les enveloppes temporelles sont estimées séparément, suivant des approches ad-hoc (les DSPs sont apprises par quantification vectorielle) et l'équivalence entre l'estimation du MV et l'IS-NMF n'est pas exploitée.

La modélisation des TFCT de signaux audio à l'aide de gaussiennes complexes a été très largement utilisée en traitement du signal, et a prouvé son efficacité dans de nombreuses applications, en particulier dans le domaine du débruitage (voir par exemple [Cohen et Gannot, 2007]). Cependant, tandis que les travaux de débruitage posent le plus souvent une trame d'observation $\mathbf{x}_n$ comme la somme des contributions d'une seule source et d'un bruit additif présents à la même trame, le paradigme considéré ici étend profondément cette modélisation, en permettant que l'observation soit la somme de *plusieurs* composantes gaussiennes, possédant chacune leur covariance.

Avec cette modélisation gaussienne, la non-négativité apparaît naturellement à travers l'équivalence du problème en terme d'identification des variances. L'égalité des phases entre les composantes

et les observations est une conséquence du filtrage de Wiener dans le cadre de la modélisation gaussienne. La reconstruction des composantes est à la fois fondée statistiquement et conservative (voir équation (VI.13) et suivante).

VI.3.1.1 Équivalence NMF/MV

Nous posons et démontrons le théorème suivant [Févotte *et al.*, 2009] :

Théorème 1 (IS-NMF comme estimateur du MV dans un modèle de somme de gaussiennes). L'estimation au sens du maximum de vraisemblance des matrices $\mathbf{W}$ et $\mathbf{H}$ à partir de l'observation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ est équivalente à la résolution du problème de NMF : $\mathbf{V} \approx \mathbf{WH}$, par minimisation de la divergence d'Itakura-Saito (III.6).

Démonstration. Sous les hypothèses précédentes, l'opposé de la log-vraisemblance $C_{MV}(\mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{X}|\mathbf{W}, \mathbf{H})$ s'écrit :

$$C_{MV}(\mathbf{W}, \mathbf{H}) = -\sum_{n=1}^{N} \sum_{f=1}^{F} \log \mathcal{N}_c \left(x_{fn} | 0, \sum_k w_{fk}\, h_{kn} \right) \tag{VI.10}$$

$$= NF \log \pi + \sum_{n=1}^{N} \sum_{f=1}^{F} \log \left(\sum_k w_{fk}\, h_{kn} \right) + \frac{|x_{fn}|^2}{(\sum_k w_{fk}\, h_{kn})} \tag{VI.11}$$

$$\stackrel{\text{c}}{=} \sum_{n=1}^{N} \sum_{f=1}^{F} d_{IS} \left(|x_{fn}|^2 \mid \sum_k w_{fk}\, h_{kn} \right) \tag{VI.12}$$

Ainsi, la minimisation de $C_{MV}(\mathbf{W}, \mathbf{H})$ suivant les variables $\mathbf{W}$ et $\mathbf{H}$ (estimateur du maximum de vraisemblance) se réduit au problème de NMF $\mathbf{V} \approx \mathbf{WH}$ par minimisation de la divergence d'Itakura-Saito. Remarquons que le théorème 1 reste vrai dans le cas de gaussiennes à valeurs réelles ; dans ce cas, $C_{MV}(\mathbf{W}, \mathbf{H})$ et $D_{IS}(\mathbf{V}|\mathbf{WH})$ sont égales à une constante additive et un facteur multiplicatif 1/2 près. □

Le modèle génératif (VI.6) peut également être considéré comme une généralisation de modèles bien connus de signaux composites. Par exemple, on trouvera dans [Feder et Weinstein, 1988] des travaux d'inférence bayésienne de composantes superposées possédant une structure gaussienne. Dans ce travail, l'auteur suppose que les composantes sont stationnaires, et simplement modélisées par leur DSP $\mathbf{w}_k$, elle-même représentée *via* un ensemble de paramètres d'intérêt $\boldsymbol{\theta}_k$ à estimer. Le modèle (VI.6) étend ce modèle en ajoutant les paramètres d'amplitude $\mathbf{H}$. Cette extension a un inconvénient de taille : en faisant dépendre du nombre total de trames, N, le nombre de paramètres à estimer $(F\,K + K\,N)$, le modèle ne permet pas de conserver les propriétés d'optimalité asymptotique de l'estimation du MV. Cependant, il faut noter que c'est précisément l'introduction de ces paramètres d'amplitude qui permet de considérer $\mathbf{W}$ comme un ensemble de paramètres éventuellement identifiables. En effet, si les h_{kn} sont fixés à 1, la variance de $\mathbf{x}_n$ se réduit à $\sum_k \mathbf{w}_k$ (pour tout n), c'est-à-dire la somme des paramètres restant à estimer, ce qui rendrait bien entendu les DSP $\mathbf{w}_k$ non identifiables. [1]

Un autre point intéressant de l'équivalence IS-NMF / estimation du MV est la possibilité de *reconstruire* les composantes $\mathbf{c}_{kn}$, dans un sens d'optimalité statistique, ce qui n'est pas forcément le cas

1. Les paramètres du modèle $\boldsymbol{\theta}$ sont dits identifiables ssi, à des permutations d'indice près, il existe un unique $\boldsymbol{\theta}$ tel que x suit le modèle VI.6.

lorsque la NMF est réalisée à partir d'autres fonctions de coût. En effet, si $\mathbf{W}$ et $\mathbf{H}$ sont connus, un estimateur des moindres carrés (MMSE) peut-être obtenu par filtrage de Wiener, de manière que :

$$\hat{c}_{kfn} = \frac{w_{fk}\, h_{kn}}{\sum_{l=1}^{K} w_{fl}\, h_{ln}} x_{fn} \tag{VI.13}$$

Les gains de Wiener somment à 1 pour une entrée donnée d'indice (f, n) ; la décomposition est donc conservative :

$$\mathbf{x}_n = \sum_{k=1}^{K} \hat{\mathbf{c}}_{kn}. \tag{VI.14}$$

Une des conséquences de la reconstruction par filtrage de Wiener est l'égalité des phases entre toutes les composantes $\hat{c}_{kfn}$ et le mélange x_{fn}.

VI.3.2 Bruit multiplicatif

Théorème 2 (IS-NMF comme estimateur du MV en présence de bruit multiplicatif). Considérons temporairement un autre modèle génératif :

$$\mathbf{V} = (\mathbf{WH}) \otimes \mathbf{B} \tag{VI.15}$$

où $\mathbf{B}$ est un bruit multiplicatif indépendant et identiquement distribué suivant une loi Gamma de moyenne 1. Alors, l'estimation par maximum de vraisemblance de $\mathbf{W}$ et $\mathbf{H}$ est équivalente à la résolution du problème de NMF : $\mathbf{V} \approx \mathbf{WH}$, par minimisation de la divergence d'Itakura-Saito III.6.

Démonstration. Soient $\{b_{fn}\}$ les coefficients de la matrice $\mathbf{B}$. On rappelle la notation $\hat{\mathbf{V}} = \mathbf{WH}$. Nous avons $v_{fn} = \hat{v}_{fn}\, b_{fn}$, avec $p(b_{fn}) = \mathcal{G}(b_{fn}|\alpha, \beta)$, $\mathcal{G}(u|\alpha, \beta)$ étant la fonction densité de probabilité (fdp) :

$$\mathcal{G}(u|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, u^{\alpha-1}\, \exp(-\beta\, u),\ u \geq 0 \tag{VI.16}$$

Sous l'hypothèse de bruit iid, l'opposé de la log-vraisemblance $C_{MV}(\mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{V}|\mathbf{W}, \mathbf{H})$ s'écrit :

$$\begin{aligned} C_{MV}(\mathbf{W}, \mathbf{H}) &= -\sum_{f,n} \log p(v_{fn}|\hat{v}_{fn}) && \text{(VI.17)} \\ &= -\sum_{f,n} \log \mathcal{G}\left(v_{fn}/\hat{v}_{fn}|\alpha, \beta\right)/\hat{v}_{fn} && \text{(VI.18)} \\ &\stackrel{\text{c}}{=} \beta \sum_{f,n} \frac{v_{fn}}{\hat{v}_{fn}} - \frac{\alpha}{\beta} \log \frac{v_{fn}}{\hat{v}_{fn}} && \text{(VI.19)} \end{aligned}$$

Le quotient α/β n'est autre que la moyenne de la distribution Gamma. Lorsqu'il vaut 1, on obtient l'égalité de $C_{MV}(\boldsymbol{\theta})$ et de $D_{IS}(\mathbf{V}|\hat{\mathbf{V}}) = D_{IS}(\mathbf{V}|\mathbf{WH})$ à une constante et un facteur multiplicatif près. □

Cette nouvelle équivalence offre en particulier une nouvelle explication à l'invariance par homothétie de la divergence d'Itakura-Saito : le bruit joue comme un facteur d'échelle sur $\hat{v}_{fn}$. Parallèlement, EUC-NMF est équivalente à l'estimation par MV en présence de bruit gaussien iid additif. L'influence des

bruits additifs est plus importante sur les coefficients de faible amplitude de $\hat{\mathbf{V}}$ (c'est-à-dire, dans les cas de faible RSB), par rapport aux coefficients de plus forte amplitude. Dans le cas de KL-NMF, elle ne correspond ni à un bruit additif, ni à un bruit multiplicatif, mais peut s'exprimer comme une estimation du MV dans du bruit poissonnien. En résumé, nous avons :

$$\text{EUC-NMF} : \quad p(v_{fn}|\hat{v}_{fn}) = \quad \mathcal{N}(v_{fn}|\hat{v}_{fn}, \sigma^2), \tag{VI.20}$$

$$\text{KL-NMF} : \quad p(v_{fn}|\hat{v}_{fn}) = \quad \mathcal{P}(v_{fn}|\hat{v}_{fn}), \tag{VI.21}$$

$$\text{IS-NMF} : \quad p(v_{fn}|\hat{v}_{fn}) = \quad \frac{1}{\hat{v}_{fn}}\,\mathcal{G}\left(\frac{v_{fn}}{\hat{v}_{fn}}|\alpha, \frac{1}{\alpha}\right), \tag{VI.22}$$

et dans tous les cas, $\mathbb{E}\left[v_{fn}|\hat{v}_{fn}\right] = \hat{v}_{fn}$.

Le théorème 2 peut être rapproché de la « mesure d'erreur statistiquement motivée » proposée dans [Abdallah et Plumbley, 2004], et qui s'avère être la divergence d'Itakura-Saito, dans un contexte très similaire au nôtre (le codage parcimonieux non-négatif, voir par exemple [Plumbley *et al.*, 2006]). En soulignant l'invariance de cette métrique par homothétie, ce travail amène [Virtanen, 2007] à considérer, sans l'identifier comme telle, la divergence d'Itakura-Saito comme fonction de coût de la NMF, dans un contexte de séparation de source audio mono-capteur. Cependant, l'algorithme est appliqué sur le spectrogramme d'amplitude, et non de puissance, perdant ainsi l'interprétation statistique. Les sources sont reconstruites via l'équation (VI.5) plutôt que par filtrage de Wiener, avec une perte d'optimalité.

VI.4 Algorithmes EM et SAGE pour la NMF non contrainte

Dans cette section, nous décrivons un algorithme de type EM [Dempster *et al.*, 1977] permettant d'estimer les paramètres $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, à partir du modèle et du formalisme introduits dans le théorème 1.

VI.4.1 Cadre de l'algorithme SAGE

En particulier, la structure additive du modèle génératif (VI.6) permet la mise à jour *séparée* des paramètres décrivant chaque composante $\mathbf{C}_k \stackrel{\text{def}}{=} [\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}]$ en utilisant une variante d'EM nommée SAGE (*Space Alternating Generalized EM*) [Fessler et Hero, 1994]. SAGE est une extension des algorithmes EM, conçue pour des modèles de données présentant des structures remarquables, parmi lesquels les données générées par superposition de composantes. SAGE est connu pour converger plus rapidement (en nombre d'itérations) que les algorithmes EM usuels. Cependant, une itération de SAGE est en général plus coûteuse en temps de calcul qu'une itération EM, étant donné qu'il requiert de mettre à jour les statistiques exhaustives « plus souvent » *(cf. infra)*.

Soit une partition de l'espace des paramètres, $\boldsymbol{\theta} = \bigcup_{k=1}^{K} \boldsymbol{\theta}_k$, telle que :

$$\boldsymbol{\theta}_k = \{\mathbf{w}_k, h_k\}, \tag{VI.23}$$

Rappelons que $\mathbf{w}_k$ est la k^e colonne de $\mathbf{W}$ et h_k la k^e ligne de $\mathbf{H}$. Afin d'utiliser l'algorithme SAGE, il faut choisir pour chaque sous-ensemble de paramètres $\boldsymbol{\theta}_k$ un espace de données latentes (*hidden data space*) complet pour le sous-ensemble considéré. Ici, on choisit simplement comme espace latent les variables $\mathbf{C}_k \stackrel{\text{def}}{=} [\mathbf{c}_{k1}, \ldots, \mathbf{c}_{kN}]$.

On construit ensuite une fonctionnelle $Q_k^{MV}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')$, de la même manière que dans les algorithmes EM mais ne dépendant que du sous-ensemble $\boldsymbol{\theta}_k$. Cette fonctionnelle s'écrit comme l'espérance conditionnelle de l'opposé de la log-vraisemblance des variables latentes $\mathbf{C}_k$:

$$Q_k^{MV}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\mathbf{C}_k|\boldsymbol{\theta}_k)\, p(\mathbf{C}_k|\mathbf{X},\boldsymbol{\theta}')\, d\mathbf{C}_k. \tag{VI.24}$$

Une itération ℓ de l'algorithme SAGE consiste à calculer (étape « E ») et à minimiser (étape « M ») la fonctionnelle Q_k^{MV} pour chaque valeur de k. Remarquons que contrairement à un algorithme EM « classique », au cours duquel les paramètres ne sont mis à jour qu'une fois par itération et ne dépendent que des valeurs à l'itération précédente, dans le cas de SAGE $\boldsymbol{\theta}'$ contient les valeurs les plus à jour des paramètres (mises à jour pendant l'itération en cours, nommément les $\boldsymbol{\theta}_j$ pour $j < k$). Cela amène un léger surcroît de coût de calcul (*cf. supra*).

VI.4.2 Convergence

[Fessler et Hero, 1994] établit des propriétés de convergence sous certaines conditions. Il convient de s'assurer que ces hypothèses sont vérifiées, et de préciser quelles propriétés de convergence sont établies. Pour ce faire, l'espace de données latentes doit vérifier la condition (il est dit complet, ou admissible) :

$$p(\mathbf{X},\mathbf{D}_k;\theta) = p(\mathbf{X}|\mathbf{D}_{-k};\theta_k)p(\mathbf{X};\theta), \tag{VI.25}$$

où $\mathbf{D}_{-k} = \{\mathbf{D}_j\}_{j\neq k}$. Cela revient à dire que la distribution conditionnelle $p(\mathbf{X}|\mathbf{D}_{-k})$ est indépendante de θ_k. Ceci est vrai sous l'hypothèse d'indépendance mutuelle des lignes h_k et des colonnes $\mathbf{w}_k$. Le réalisme de cette hypothèse d'indépendance pourrait cependant être discutée, compte-tenu de la nature des données.

Dans ces conditions, le cadre offert par l'usage de SAGE garantit la croissance monotone du critère $C^{MV}(\boldsymbol{\theta})$. De plus, [Fessler et Hero, 1994] prouve l'existence d'une région de convergence monotone en norme, c'est-à-dire que suivant ces mises à jour, $\boldsymbol{\theta}$ converge en norme vers un minimum local, pourvu que l'algorithme ait été initialisé dans un voisinage approprié de ce minimum. Cependant, cette preuve est restreinte à l'intérieur du domaine de définition, n'autorisant donc pas de coefficients nuls, et ne résout pas le problème de l'initialisation.

VI.4.3 Étape E

Le calcul de la fonctionnelle $Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')$ implique la détermination de deux quantités : la log-vraisemblance des données latentes $\log p(\mathbf{C}_k|\boldsymbol{\theta}_k)$, et la distribution a posteriori des données latentes, $p(\mathbf{C}_k|\mathbf{X},\boldsymbol{\theta}')$.

Le modèle (VI.7) nous permet d'écrire la vraisemblance des données latentes :

$$-\log p(\mathbf{C}_k|\boldsymbol{\theta}_k) = -\sum_{n=1}^{N}\sum_{f=1}^{F} \log \mathcal{N}_c(c_{kfn}|0, h_{kn}\, w_{fk}) \tag{VI.26}$$

$$\stackrel{\text{c}}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\, h_{kn}\right) + \frac{|c_{kfn}|^2}{w_{fk}\, h_{kn}}. \tag{VI.27}$$

La distribution a posteriori $p(\mathbf{C}_k|\mathbf{X},\boldsymbol{\theta}')$ s'obtient en utilisant la méthodologie proposée dans [Benaroya *et al.*, 2003] et basée sur un filtrage de Wiener. En effet, à k fixé, en écrivant $\mathbf{x}_n = \mathbf{c}_{kn} + \sum_{j\neq k} \mathbf{c}_{jn}$ et sous l'hypothèse d'indépendance des $\mathbf{c}_{jn}$ et de leur distribution gaussienne iid, nous pouvons voir l'estimation de la distribution de $\mathbf{c}_{kn}$ comme un problème de séparation de deux sources gaussiennes, dont l'estimateur bayésien optimal est donné par le filtrage de Wiener [Wiener, 1949] :

$$\mathbb{E}\left[\mathbf{c}_{kn}(f)|\mathbf{x}_n\right] = \left(\frac{var(\mathbf{c}_{kn}(f))}{var(\mathbf{c}_{kn}(f) + \sum_{j\neq k}\mathbf{c}_{jn}(f))}\right) x_n(f) \tag{VI.28}$$

$$var(\mathbf{c}_{kn}(f)|\mathbf{x}_n(f)) = \left(\frac{var(\mathbf{c}_{kn}(f))}{var(\mathbf{c}_{kn}(f) + \sum_{j\neq k}\mathbf{c}_{jn}(f))}\right) \sum_{j\neq k} var(\mathbf{c}_{jn}(f)) \tag{VI.29}$$

Ainsi, en utilisant (VI.7), on obtient la distribution a posteriori des données latentes :

$$p(\mathbf{C}_k|\mathbf{X},\boldsymbol{\theta}') = \prod_{n=1}^{N}\prod_{f=1}^{F} \mathcal{N}_c(c_{k,fn}|\mu_{k,fn}^{post'}, \lambda_{k,fn}^{post'}), \tag{VI.30}$$

où $\mu_{k,fn}^{post'}$ et $\lambda_{k,fn}^{post'}$ sont donnés par :

$$\mu_{k,fn}^{post'} = \frac{w'_{fk}\,h'_{kn}}{\sum_l w'_{fl}\,h'_{ln}}\, x_{fn}, \tag{VI.31}$$

$$\lambda_{k,fn}^{post'} = \frac{w'_{fk}\,h'_{kn}}{\sum_l w'_{fl}\,h'_{ln}} \sum_{l\neq k} w'_{fl}\,h'_{ln}. \tag{VI.32}$$

Les variables primées désignent les valeurs les plus à jour des paramètres.

Enfin, l'étape E en elle-même est réalisée en prenant l'espérance de (VI.27) conditionnellement à la distribution a posteriori (*cf.* définition (VI.24)), ce qui nous conduit à :

$$Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\,h_{kn}\right) + \frac{|\mu_{k,fn}^{post\,\prime}|^2 + \lambda_{k,fn}^{post\,\prime}}{w_{fk}\,h_{kn}} \tag{VI.33}$$

$$\stackrel{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}(|\mu_{k,fn}^{post\,\prime}|^2 + \lambda_{k,fn}^{post\,\prime} \mid w_{fk}\,h_{kn}). \tag{VI.34}$$

VI.4.4 Étape M

L'étape de maximisation composante par composante revient à un problème de NMF d'ordre 1 :

$$\min_{\mathbf{w}_k,\, h_k \geq \mathbf{0}} D_{IS}(\hat{\mathbf{V}}'_k \,|\, \mathbf{w}_k\, h_k) \tag{VI.35}$$

Entrée : Matrice à coefficients positifs ou nuls $\mathbf{V}$
Sortie : Matrices à coefficients positifs ou nuls $\mathbf{W}$ et $\mathbf{H}$ telles que $\mathbf{V} \approx \mathbf{WH}$
Initialiser $\mathbf{W}$ et $\mathbf{H}$ avec des valeurs positives ou nulles
for $i = 1 : n_{iter}$ **do**
  for $k = 1 : K$ **do**
    Calculer $\mathbf{G}_k = \frac{\mathbf{w}_k h_k}{\mathbf{WH}}$ `% gain de Wiener`
    Calculer $\mathbf{V}_k = \mathbf{G}_k^{\cdot[2]}.\mathbf{V} + (1 - \mathbf{G}_k).(\mathbf{w}_k h_k)$ `% Puissance a posteriori de` $\mathbf{C}_k$
    $h_k \leftarrow \frac{1}{F}(\mathbf{w}_k^{\cdot[-1]})^T \mathbf{V}_k$ `% Mise à jour de la` k^e `ligne de` $\mathbf{H}$
    $\mathbf{w}_k \leftarrow \frac{1}{N}\mathbf{V}_k\,(h_k^{\cdot[-1]})^T$ `% Mise à jour de la` k^e `colonne de` $\mathbf{W}$
    Normaliser $\mathbf{w}_k$ et h_k
  end for
end for

TABLE VI.1 – IS-NMF/EM.

où $\hat{\mathbf{V}}'_k$ désigne la reconstruction $\hat{\mathbf{V}}_k$ calculée à partir des paramètres à jour $\boldsymbol{\theta}'$. Dans ce cadre à une composante, les dérivées partielles de la fonctionnelle s'expriment simplement comme :

$$\nabla_{h_{kn}} Q_k^{MV}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \frac{F}{h_{kn}} - \frac{1}{h_{kn}^2}\sum_{f=1}^{F}\frac{\hat{v}'_{k,fn}}{w_{fk}}, \tag{VI.36}$$

$$\nabla_{w_{fk}} Q_k^{MV}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \frac{N}{w_{fk}} - \frac{1}{w_{fk}^2}\sum_{n=1}^{N}\frac{\hat{v}'_{k,fn}}{h_{kn}}. \tag{VI.37}$$

Leur annulation est alors immédiate, et conduit aux règles de mise à jour :

$$h_{kn}^{(\ell+1)} = \frac{1}{F}\sum_{f}\frac{\hat{v}'_{kfn}}{w_{fk}^{(\ell)}}, \tag{VI.38}$$

$$w_{fk}^{(\ell+1)} = \frac{1}{N}\sum_{n}\frac{\hat{v}'_{kfn}}{h_{kn}^{(\ell+1)}}, \tag{VI.39}$$

Ces règles garantissent $Q_k^{MV}(\mathbf{w}_k^{(i+1)}, h_k^{(i+1)}|\boldsymbol{\theta}') \leq Q_k^{MV}(\mathbf{w}_k^{(i)}, h_k^{(i)}|\boldsymbol{\theta}')$, donc la décroissance de la fonctionnelle à chaque itération. Elles peuvent être écrites sous forme matricielle. L'algorithme résultant est présenté dans la table VI.1. Dans la suite, on le désignera de manière abrégée par « IS-NMF/EM ».

VI.4.5 Commentaires

IS-NMF/EM et IS-NMF/MU (voir section III.4) ont la même complexité algorithmique par itération : $\mathcal{O}(12\,FKN)$. Cependant, ils peuvent conduire à des temps de calcul différents, comme on le verra chapitre VIII, pour deux raisons : une vitesse de convergence (en nombre d'itérations) différente, et les possibilités variables d'astuces d'implémentation Matlab (boucles vs. produits matriciels ou de Hadamard).

Les propriétés de l'algorithme SAGE [Fessler et Hero, 1994] garantissent la convergence de l'algorithme IS-NMF/EM vers un point stationnaire de la fonction de coût $D_{IS}(\mathbf{V}|\mathbf{WH})$. Cependant, il ne

peut converger que vers un point de l'intérieur du domaine de définition des paramètres, autrement dit, $\mathbf{W}$ et $\mathbf{H}$ ne peuvent contenir de coefficients strictement nuls. En effet, dans l'équation (VI.35), si l'une des variables w_{fk} ou h_{kn} tend vers 0, le coût $d_{IS}(v_{fn}|w_{fk}h_{kn})$ tend vers l'infini.

L'algorithme SAGE a été utilisé dans le contexte de la séparation de source monocapteur par [Ozerov *et al.*, 2007], pour l'estimation de paramètres d'un modèle plus ou moins relié au modèle (VI.6) utilisé ici. En effet, ces auteurs s'intéressent au problème de séparation voix chantée/accompagnement musical en recourant à un modèle génératif de la forme $\mathbf{x}_n = \mathbf{c}_{V,n} + \mathbf{c}_{M,n}$, où la première composante modélise la voix tandis que la seconde représente la musique. Chaque composante se voit ensuite attribuée un modèle de mélange gaussien (GMM). Les paramètres du GMM modélisant la voix sont appris sur des données d'entraînement, tandis que les paramètres liés à la musique sont adaptés depuis les données observées. Bien que reliés, les modèles GMM et NMF demeurent assez fondamentalement différents. Le premier exprime le signal comme la somme de deux composantes pouvant chacune prendre différents états, tandis que le second l'exprime comme somme de K composantes, dont chacune représente un seul objet. Aucun des deux modèles n'est *a priori* meilleur que l'autre : ils s'intéressent à des caractéristiques différentes du signal. On peut imaginer que les deux modèles puissent être utilisés simultanément dans une approche mixte dans le cadre global de SAGE, par exemple en modélisant la voix par un GMM (*i.e.*, une composante spécifique, prenant différents états) et la musique par un modèle NMF (*i.e.*, un signal composite, comportant plusieurs objets sonores), une idée par exemple exploitée dans [Durrieu *et al.*, 2008].

Chapitre VII

Contraintes de régularité temporelle et d'harmonicité

Résumé

Où l'on étend et adapte le modèle probabiliste précédent pour résoudre un problème de NMF avec contraintes d'harmonicité et de régularité, grâce un algorithme original de type Espérance-Maximisation.

VII.1 Introduction

Le passage au cadre statistique et l'estimation du MV pour résoudre le problème de NMF offre des propriétés formelles intéressantes, mais ne présente pas particulièrement d'intérêt informatique pratique, l'algorithme multiplicatif restant en général plus efficace et moins coûteux en temps de calcul. Cependant, un des intérêts majeurs de l'approche statistique est la possibilité de basculer de l'estimation du MV à l'estimation du *maximum a posteriori* (MAP), grâce à la règle de Bayes qui s'exprimera dans notre cas sous la forme :

$$p(\mathbf{W},\mathbf{H}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{W},\mathbf{H})p(\mathbf{W})p(\mathbf{H})}{p(\mathbf{X})} \tag{VII.1}$$

Cette formulation offre la possibilité de choisir des distributions a priori $p(\mathbf{W})$ et $p(\mathbf{H})$, et ainsi d'induire certaines propriétés que nous attendons de la solution ; autrement dit, d'imposer d'une nouvelle manière des contraintes à la NMF.

Dans ce chapitre, nous proposons d'intégrer à la modélisation statistique le modèle harmonique (V.6) proposé précédemment dans un cadre déterministe (section V.3, page 84), et d'utiliser le choix de l'a priori $p(\mathbf{H})$ pour imposer la régularité des enveloppes temporelles. L'intérêt particulier de ces deux contraintes a été exposé dans la section V.2 (page 80). La section VII.2 présente le modèle choisi pour imposer cette régularité temporelle, ainsi que l'estimateur MAP résultant et l'algorithme qui le détermine. On intègre ensuite au modèle la contrainte d'harmonicité, dans la section VII.3.

VII.2 Régularité temporelle

L'intérêt de la contrainte de régularité et son implantation dans le cadre déterministe ont été évoqués section V.2.2. En ce qui concerne les approches probabilistes, [Shashanka *et al.*, 2008b] propose une version régularisée de KL-NMF dans un cadre bayésien, et qui inclut également des contraintes de parcimonie. Nous présentons ici notre approche de la régularité temporelle, publiée dans [Févotte *et al.*, 2009].

VII.2.1 Modèles

On se place dans un cadre bayésien, où $\mathbf{W}$ et $\mathbf{H}$ ont des distributions a priori $p(\mathbf{W})$ et $p(\mathbf{H})$ indépendantes et connues. Nous cherchons un estimateur joint de $\mathbf{W}$ et $\mathbf{H}$ maximisant la probabilité a posteriori. Cela se traduit par un critère à minimiser que nous écrivons :

$$C_{MAP}(\mathbf{W},\mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{W},\mathbf{H}|\mathbf{X}) \tag{VII.2}$$

$$\stackrel{\text{c}}{=} D_{IS}(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) \tag{VII.3}$$

Si l'on utilise des a priori indépendants de la forme $p(\mathbf{W}) = \prod_k p(\mathbf{w}_k)$ et $p(\mathbf{H}) = \prod_k p(h_k)$, l'algorithme SAGE présenté dans le chapitre VI précédent peut encore être utilisé pour l'estimation du MAP. Les fonctionnelles en jeu vont alors s'écrire :

Loi	Densité de probabilité	Mode	Moyenne	Variance
Gamma	$\mathcal{G}(u\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, u^{\alpha-1}\, \exp(-\beta\, u)$	$(\alpha-1)/\beta$	α/β	α/β^2
Inverse-Gamma	$\mathcal{IG}(u\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, u^{-(\alpha+1)}\, \exp(-\frac{\beta}{u})$	$\beta/(\alpha+1)$	$\beta/(\alpha-1)$	$\beta^2/(\alpha-1)^2(\alpha-2)^2$

TABLE VII.1 – Lois Gamma et inverse-Gamma.

$$Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\boldsymbol{\theta}_k|\mathbf{C}_k)\, p(\mathbf{C}_k|\mathbf{X},\boldsymbol{\theta}')\, d\mathbf{C}_k \tag{VII.4}$$

$$\stackrel{\text{c}}{=} Q_k^{MV}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') - \log p(\mathbf{w}_k) - \log p(h_k) \tag{VII.5}$$

L'étape E est inchangée, puisqu'elle consiste toujours à calculer $Q_k^{MV}(\mathbf{w}_k, h_k|\boldsymbol{\theta}')$, de la même manière que dans la section VI.4. Seule l'étape de maximisation va être modifiée par l'introduction de contraintes via les termes $-\log p(\mathbf{w}_k)$ et/ou $-\log p(h_k)$.

Plus précisement, dans la suite, nous considérons des a priori sous forme de chaîne de Markov, qui favoriseront la régularité des lignes de $\mathbf{H}$. Nous ne poserons pas d'a priori sur $\mathbf{W}$ (qui sera donc estimé par MV comme précédemment). Cependant, il faut noter que la méthodologie présentée ici pourrait parfaitement s'appliquer à $\mathbf{W}$, que la structure d'a priori sur $\mathbf{W}$ ou $\mathbf{H}$ soit la même ou totalement différente. Nous pouvons également remarquer que, puisque les composantes sont toutes traitées séparément, on pourrait de la même manière poser des modèles différents pour chacune d'elles (en utilisant par exemple des GMM pour certaines d'entre elles, comme évoqué à la fin du chapitre VI).

Nous supposons que la distribution des h_k possède une structure de chaîne de Markov :

$$p(h_k) = p(h_{k1}) \prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)}), \tag{VII.6}$$

où $p(h_{kn}|h_{k(n-1)})$ est une fonction densité de probabilité (fdp) dont le mode est atteint en $h_{k(n-1)}$. La motivation de ce choix est de contraindre chaque coefficient h_{kn} à ne pas différer significativement de sa valeur à la trame précédente $n-1$, ce qui devrait favoriser la régularité de la ligne h_k de proche en proche. On propose de considérer deux choix possibles de fdp, pour $n = 2, \ldots, N$,

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha, (\alpha+1)\, h_{k(n-1)}) \tag{VII.7}$$

et

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{G}(h_{kn}|\alpha, (\alpha-1)/h_{k(n-1)}) \tag{VII.8}$$

où $\mathcal{G}(x|\alpha,\beta)$ est la distribution Gamma et $\mathcal{IG}(x|\alpha,\beta)$ la distribution inverse-Gamma. La valeur des densités de probabilité associées ainsi que les valeurs du mode et de la variance correspondantes sont rappelées dans la table VII.1 ; les figures VII.1 et VII.2 illustrent ces fdp pour quelques valeurs de α.

On peut vérifier dans la table VII.1 que les distributions proposées dans les équations (VII.7) et (VII.8) atteignent effectivement leur mode en $h_{kn} = h_{k(n-1)}$. α est un paramètre dit « de forme », qui contrôle la dispersion de la densité autour de son mode. Une grande valeur de α réduit la dispersion et donc contraint davantage h_k à être régulier, tandis qu'une faible valeur de α rend l'a priori plus

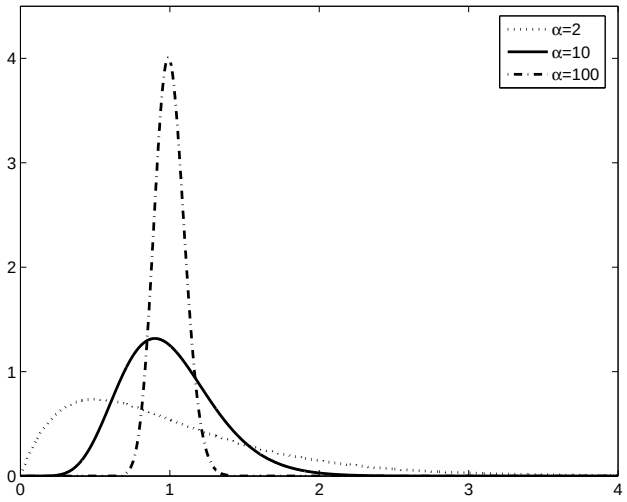


FIGURE VII.1 – Densité de probabilité de lois Gamma $\mathcal{G}(u|\alpha,\beta)$ de moyenne 1, α variable.

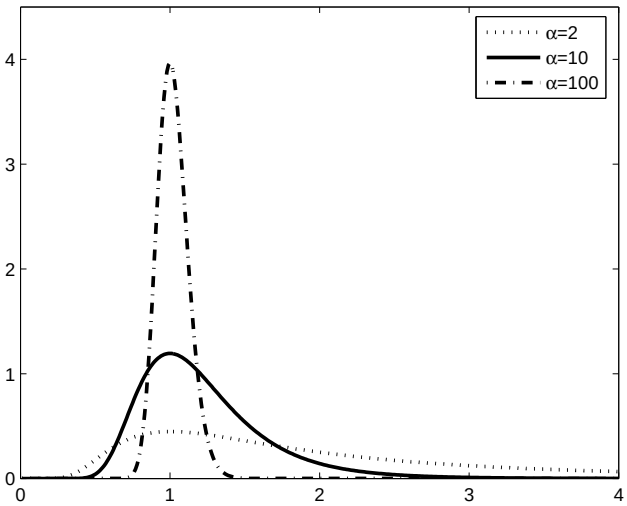


FIGURE VII.2 – Densité de probabilité de lois inverse-Gamma $\mathcal{IG}(u|\alpha,\beta)$ de moyenne 1, α variable.

dispersé et donc moins contraignant. Les deux distributions, Gamma et inverse-Gamma deviennent très proches pour de grandes valeurs de α, comme on peut le voir sur les figures VII.1 et VII.2.

Le choix de cet a priori est motivé par plusieurs raisons. D'une part, la non-négativité y est naturellement assurée. D'autre part, l'a priori est conjugué pour le modèle d'observation gaussien, ce qui amène une simplicité calculatoire et la possibilité d'une résolution analytique. Enfin, il paraît adapté pour la modélisation des enveloppes temporelles du signal musical : il favorise la régularité grâce au choix approprié du mode, et l'asymétrie de la distribution autour de son mode contraint davantage de régularité sur les parties décroissantes ($h_{kn} \leq h_{k(n-1)}$) que sur les parties croissantes ($h_{kn} \geq h_{k(n-1)}$) de l'enveloppe. Ce dernier point est particulièrement intéressant, car nous voulons favoriser la régularité des moments de silence et des parties stationnaires des notes, sans pour autant trop défavoriser les attaques.

Les a priori posés sur $\mathbf{H}$ ne sont complètement définis qu'après avoir fixé une distribution pour le début de la chaîne de Markov. On choisit ici l'a priori « non-informatif [1] » de Jeffreys $p(h_{k1}) \propto 1/h_{k1}$.

VII.2.2 Algorithmes

En prenant désormais en compte la structure (VII.6), la dérivée de $Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}')$ par rapport à h_{kn} devient, $\forall n = 2, \ldots, N-1$:

$$\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \nabla_{h_{kn}} Q_k^{ML}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') - \nabla_{h_{kn}} \log p(h_{k(n+1)}|h_{kn}) - \nabla_{h_{kn}} \log p(h_{kn}|h_{k(n-1)}) \quad \text{(VII.9)}$$

En remplaçant chaque terme par sa valeur, on montre que cette dérivée peut s'écrire :

$$\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \frac{-F}{h_{kn}^2}(p_2\, h_{kn}^2 + p_1\, h_{kn} - p_0) \quad \text{(VII.10)}$$

où p_0, p_1 et p_2 sont des coefficients scalaires dépendant de la loi utilisée (Gamma ou inverse-Gamma), et dont la valeur est donnée dans la table VII.2 [2]. Ainsi, la mise à jour des h_{kn} se ramène à la résolution d'une équation polynômiale du second degré. Ce polynôme possède une unique racine positive, que nous écrivons sous une forme évitant une possible division par zéro :

$$h_{kn}^{(l+1)} = \frac{2p_0}{\sqrt{p_1^2 + 4p_2p_0} + p_1} \quad \text{(VII.11)}$$

Les coefficients h_{k1} et h_{kN} des bords de la chaîne de Markov ont des mises à jour spécifiques, qui s'expriment cependant comme la solution de l'annulation de polynômes d'ordre 1 ou 2, dont les coefficients sont également indiqués dans la table VII.2.

Nous pouvons remarquer que la différence des mises à jour entre la loi Gamma et la loi inverse-Gamma se résume principalement à intervertir les positions de $h_{k(n-1)}$ et $h_{k(n+1)}$ dans les expressions de p_0 et p_2. De ce fait, utiliser un a priori Gamma sur une chaîne anticausale (« partant de la fin ») $p(h_k) = \prod_{n=1}^{N-1} p(h_{kn}|h_{k(n+1)})\, p(h_{kN})$ et de paramètre de forme α est exactement équivalent (en termes de règles de mises à jour par maximum a posteriori) à l'utilisation d'une chaîne inverse-Gamma causale (« partant

1. Une loi a priori est dite non informative lorsqu'elle est construite uniquement à partir des observations, sans usage d'un paramètre. Le lecteur pourra se référer à [Robert, 2007]

2. Les valeurs de p_0, p_1, p_2 sont communes à chaque $n \in [2 \ldots N-1]$ et diffèrent au bord de la chaîne de Markov ($n = 1$ et $n = N$). Ils dépendent bien sûr de k, n et ℓ ; nous ne reportons pas cette dépendance dans la notation pour des questions de lisibilité.

inverse-Gamma			
	p_2	p_1	p_0
h_{k1}	$(\alpha+1)/h_{k2}$	$F-\alpha+1$	$-F\,\tilde{h}_{k1}^{ML}$
h_{kn}	$(\alpha+1)/h_{k(n+1)}$	$F+1$	$-F\,\tilde{h}_{kn}^{ML}-(\alpha+1)\,h_{k(n-1)}$
h_{kN}	0	$F+\alpha+1$	$-F\,\tilde{h}_{kN}^{ML}-(\alpha+1)\,h_{k(N-1)}$
Gamma			
	p_2	p_1	p_0
h_{k1}	0	$F+\alpha+1$	$-F\,\tilde{h}_{k1}^{ML}-(\alpha-1)\,h_{k2}$
h_{kn}	$(\alpha-1)/h_{k(n-1)}$	$F+1$	$-F\,\tilde{h}_{kn}^{ML}-(\alpha-1)\,h_{k(n+1)}$
h_{kN}	$(\alpha-1)/h_{k(N-1)}$	$F-\alpha+1$	$-F\,\tilde{h}_{kN}^{ML}$

TABLE VII.2 – Coefficients du polynôme d'ordre 2 à annuler pour mettre à jour h_{kn} dans l'IS-NMF bayésienne avec prior en chaîne de Markov. $\tilde{h}_{kn}^{ML}$ désigne la mise-à-jour du MV donnée par l'équation (VI.38).

du début », *cf.* équation(VII.6)) de paramètre de forme $\alpha-2$. Inversement, une chaîne inverse-Gamma anticausale et de paramètre α est équivalente à une chaîne Gamma causale de paramètre $\alpha+2$.

Récemment, les auteurs de [Virtanen *et al.*, 2008] ont considéré l'usage de chaînes Gamma pour régulariser la KL-NMF. La modélisation qu'ils proposent est cependant différente de la nôtre. L'a priori Gamma utilisé est construit sur un mode hiérarchique, c'est-à-dire en introduisant des variables auxiliaires supplémentaires qui assurent que les a priori sont bien conjugués pour le modèle de Poisson des observations. L'estimation des facteurs est ensuite réalisée par l'approche habituelle de descente multiplicative de gradient, et les résultats en séparation de source monocapteur sont présentés à partir de la factorisation du spectrogramme d'amplitude $|\mathbf{X}|$, les composantes étant reconstruites suivant la méthode (VI.5).

Dans la suite, nous désignerons cet algorithme régularisé en temps par l'abréviation **S-NMF/EM**.

VII.3 Ajout de la contrainte harmonique

Le second objectif de notre modélisation est d'imposer par ailleurs une contrainte harmonique sur le dictionnaire $\mathbf{W}$, de la même manière que dans le paradigme déterministe précédemment exposé (voir section V.3, page 84). Dans cette section, nous présentons le modèle retenu (section VII.3.1) et proposons deux algorithmes d'inférence dans ce modèle : un estimateur du MV, ne tenant compte que de la contrainte d'harmonicité (section VII.3.2), puis un estimateur du MAP (section VII.3.3) tenant aussi compte des a priori de régularité temporelle que nous avons présentés dans la section précédente.

VII.3.1 Adaptation du modèle

L'usage direct de la formulation (V.6) (page 85) dans le modèle (VI.6) (page 97) est possible, mais amène des difficultés calculatoires que l'on peut éviter en adaptant le modèle. À cet effet, nous posons le modèle alternatif suivant :

$$\mathbf{x}_n = \sum_{k=1}^{K}\sum_{m=1}^{M} \mathbf{d}_{kmn} \tag{VII.12}$$

avec

$$\begin{aligned} \mathbf{x}_n &\in \mathbb{C}^F \\ \mathbf{d}_{kmn} &\sim \mathcal{N}(0, h_{kn} e_{mk} \mathrm{diag}(\mathbf{P}_{km})) \\ \mathbf{P}_{km} &= [P_{km}(1) ... P_{km}(F)]^T \end{aligned}$$

Les gabarits spectraux $\mathbf{P}_{km}$ sont fixés comme à la section V.3.

En supposant l'indépendance des $\mathbf{d}_{kmn}$, et en posant l'égalité $\mathbf{c}_{kn} = \sum_m \mathbf{d}_{kmn}$, nous pouvons vérifier que :

$$\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn} \sum_m e_{mk} \mathrm{diag}(\mathbf{P}_{km})) \tag{VII.13}$$

ce qui assure bien l'équivalence entre le modèle (VII.12) d'une part, et d'autre part le modèle (VI.6) où l'on poserait directement la loi (VII.13).

VII.3.2 Algorithme sans contrainte temporelle

Comme précédemment, nous pouvons déterminer les paramètres du modèle par un estimateur du maximum de vraisemblance, obtenu en utilisant l'algorithme SAGE. Les paramètres à estimer sont désormais $\boldsymbol{\theta} = \{\mathbf{E}, \mathbf{H}\}$, puisque les gabarits $\mathbf{P}_{km}$ sont fixés. Le critère de MV s'écrit :

$$C_{MV}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log p(\mathbf{X}|\boldsymbol{\theta}) \tag{VII.14}$$

Nous partitionnons de nouveau l'espace des paramètres en sous-ensembles disjoints : $\boldsymbol{\theta}_k = \{\{e_{mk}\}_m, h_k\}$ réalisant $\boldsymbol{\theta} = \bigcup_{k=1}^K \boldsymbol{\theta}_k$. Cette partition et la forme additive du modèle (VII.12), où les variables latentes sont supposées indépendantes, permettent d'utiliser comme précédemment le formalisme SAGE. L'espace de données latentes associé à chaque sous-ensemble de paramètres $\boldsymbol{\theta}_k$ est $\mathbf{D}_k = [\mathbf{D}_{k1} \dots \mathbf{D}_{kN}]$, où $\mathbf{D}_{km} = [\mathbf{d}_{km1} \dots \mathbf{d}_{kmN}] \in \mathbb{C}^{F \times N}$. Les fonctionnelles à maximiser dans le cadre de SAGE sont les espérances conditionnelles des log-vraisemblances des $\mathbf{D}_k$ pour chaque k :

$$Q_k^{MV}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) \stackrel{\text{def}}{=} \int_{\mathbf{D}_k} \log p(\mathbf{D}_k|\boldsymbol{\theta}_k) p(\mathbf{D}_k|\mathbf{V}, \boldsymbol{\theta'}) d\mathbf{D}_k \tag{VII.15}$$

où $\boldsymbol{\theta'}$ contient les valeurs estimées les plus à jour des paramètres.

Grâce au modèle additif (VII.12), nous pouvons remarquer que la fonctionnelle Q_k^{MV} peut s'exprimer comme une somme (sur l'indice m) de fonctionnelles auxiliaires Q_{km}^{MV} :

$$Q_{km}^{MV}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'}) \stackrel{\text{def}}{=} \int_{\mathbf{D}_{km}} \log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km}) p(\mathbf{D}_{km}|\mathbf{V}, \boldsymbol{\theta'}) d\mathbf{D}_{km} \tag{VII.16}$$

où nous définissons de nouveaux sous-ensembles $\boldsymbol{\theta}_{km} = \{e_{mk}, h_k\}$. Remarquons que, bien que les $\{\boldsymbol{\theta}_{km}\}_{km}$ ne constituent pas une partition de $\boldsymbol{\theta}$, l'hypothèse d'indépendance des d_{kmn} nous permet d'écrire que $Q_k^{MV}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = \sum_m Q_{km}^{MV}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'})$, facilitant ainsi les calculs sans perdre le cadre théorique de SAGE. Le problème se réduit ainsi à maximiser chaque $Q_{km}^{MV}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'})$ par rapport à e_{mk}, et leurs sommes $Q_k^{MV}(\boldsymbol{\theta}_k|\boldsymbol{\theta'})$ par rapport à h_{kn}, itérativement. La maximisation de ces fonctionnelles garantit la croissance du critère $C_{MV}(\boldsymbol{\theta})$ [Fessler et Hero, 1994].

À chaque itération et pour chaque k, on calcule les fonctionnelles Q_{km}^{MV}. Comme précédemment, elles sont calculées en deux étapes. La log-vraisemblance des données latentes s'écrit :

$$\log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km}) = \log \prod_{n=1}^{N} \prod_{f=1}^{F} p(d_{kmn}(f)|\boldsymbol{\theta}_{km}) \tag{VII.17}$$

Rappelons la loi (VII.13) : $d_{kmn}(f) \sim \mathcal{N}(0, h_{kn}e_{mk}P_{km}(f))$, donc :

$$\log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km}) \stackrel{c}{=} -\sum_{n=1}^{N}\sum_{f=1}^{F} \log(h_{kn}e_{mk}P_{km}(f)) + \frac{|d_{kmn}(f)|^2}{h_{kn}e_{mk}P_{km}(f)}. \tag{VII.18}$$

Le second terme à calculer est la distribution a posteriori des données latentes $p(\mathbf{D}_{km}|\mathbf{V}, \boldsymbol{\theta'})$. Comme à la section VI.4.3 (page 101), nous utilisons le filtre de Wiener proposé dans [Benaroya *et al.*, 2003] en écrivant $\mathbf{x}_n = \mathbf{d}_{kmn} + \sum\sum_{(k',m')\neq(k,m)} \mathbf{d}_{k'm'n}$. Suivant cette méthode, la moyenne et la variance a posteriori de $d_{kmn}(f)$ s'écrivent :

$$\mu_{kmn}^{post}(f) = \frac{h_{kn}e_{mk}P_{km}(f)}{\hat{v}_{fn}} x_{fn} \tag{VII.19}$$

$$\lambda_{kmn}^{post}(f) = \frac{h_{kn}e_{mk}P_{km}(f)}{\hat{v}_{fn}} \sum_{(k',m')\neq(k,m)} \sum h_{k'n}e_{k'm'}P_{k'm'}(f) \tag{VII.20}$$

Nous prenons ensuite l'espérance de la log-vraisemblance conditionnellement à la distribution a posteriori, ce qui conduit à :

$$Q_{km}^{MV}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'}) = -\sum_{n=1}^{N}\sum_{f=1}^{F} \log(h_{kn}e_{mk}P_{km}(f)) + \frac{\left|\mu_{kmn}^{post}(f)\right|^2 + \lambda_{kmn}^{post}(f)}{h_{kn}e_{mk}P_{km}(f)} \tag{VII.21}$$

On annule les gradients $\nabla_{e_{mk}} Q_{km}^{MV}$ et $\nabla_{h_{kn}} \sum_m Q_{km}^{MV}$, ce qui conduit aux règles de mises à jour :

$$h_{kn}^{(\ell+1)} = \frac{1}{FM} \sum_f \sum_m \frac{\left|\mu_{kmn}^{post'}(f)\right|^2 + \lambda_{kmn}^{post'}(f)}{e_{mk}^{(l)}P_{km}(f)} \tag{VII.22}$$

$$e_{mk}^{(\ell+1)} = \frac{1}{FN} \sum_n \sum_f \frac{\left|\mu_{kmn}^{post'}(f)\right|^2 + \lambda_{kmn}^{post'}(f)}{h_{kn}^{(l+1)}P_{km}(f)} \tag{VII.23}$$

où l'exposant $'$ indique, comme précédemment, que $\lambda_{kmn}^{post'}$ et $\mu_{kmn}^{post'}$ sont calculés avec les valeurs les plus récemment estimées des paramètres $\mathbf{E}$ et $\mathbf{H}$.[3]

Cette forme laisse apparaître de possibles erreurs numériques si $h_{kn} = 0$ ou $e_{mk} = 0$. Ceci peut être évité en remplaçant λ_{kmn}^{post} et μ_{kmn}^{post} par leurs expressions (VII.19) et (VII.20). Nous obtenons les mises à jour implémentées en pratique :

3. La mise à jour de tous les h_{kn} implique a priori une double boucle sur les indices k et n. Dans le formalisme SAGE, nous mettons à jour séparément chaque ligne h_k l'une après l'autre, *mais*, lors de chacune de ces mises à jour, tous les h_{kn} pour n de 1 à N sont mis à jour simultanément. Avec une notation plus lourde mais plus explicite, cela signifie qu'à l'itération $(\ell+1)$, le coefficient $h_{kn}^{(\ell+1)}$ est déterminé à partir des $h_j^{(\ell+1)}$ pour tout $j < k$ et des $h_{kp}^{(\ell)}$ pour tout p. Cette précision est nécessaire pour garantir les propriétés de convergence de SAGE.

$$h_{kn}^{(\ell+1)} = h_{kn}^{(\ell)} \left(1 + \frac{1}{FM} \sum_f \sum_m \frac{h_{kn}^{(\ell)} e_{mk}^{(\ell)} P_{km}(f)}{\hat{v}_{fn}} \left(\frac{v_{fn}}{\hat{v}_{fn}} - 1 \right) \right) \tag{VII.24}$$

$$e_{mk}^{(\ell+1)} = e_{mk}^{(\ell)} \left(1 + \frac{1}{FN} \sum_n \sum_f \frac{h_{kn}^{(\ell+1)} e_{mk}^{(\ell)} P_{km}(f)}{\hat{v}_{fn}} \left(\frac{v_{fn}}{\hat{v}_{fn}} - 1 \right) \right) \tag{VII.25}$$

L'exposant ℓ dénote la valeur à l'itération ℓ et $\hat{v}_{fn}$ est la reconstruction courante de v_{fn}, *i.e.* $\hat{v}_{fn} = \sum_{k=1}^{K} \sum_{m=1}^{M} h_{kn} e_{mk} P_{km}(f) = \sum_{k=1}^{K} w_{fk} h_{kn}$ avec les valeurs les plus à jour des paramètres (soit (ℓ), soit $(\ell+1)$, selon les valeurs disponibles).

Les propriétés de convergence sont les mêmes qu'à la section VI.4.2 (page 101).

VII.3.3 Algorithme avec contrainte temporelle

Comme précédemment, l'estimation du MV de $\mathbf{E}$ et $\mathbf{H}$ présentée ci-dessus n'a pas d'intérêt pratique par rapport à l'algorithme multiplicatif correspondant H-NMF/MU mais offre les mêmes avantages, à savoir le cadre théorique assurant certaines propriétés formelles de convergence, et la possibilité d'inclure des a priori sur les paramètres et de contraindre par ce moyen les solutions de la NMF d'une manière élégante.

Nous nous plaçons toujours dans le modèle harmonique (VII.12), et utilisons le théorème de Bayes pour passer d'une estimation du MV à une estimation du MAP, le critère à minimiser étant désormais :

$$C_{MAP}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log p(\boldsymbol{\theta}|\mathbf{V}) \tag{VII.26}$$

$$\stackrel{\text{c}}{=} C_{MV}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{VII.27}$$

Nous utilisons la structure markovienne (VII.6) et l'a priori inverse-Gamma (VII.7). Les paramètres α_k sont ici fixés arbitrairement suivant le degré de régularité souhaité (plus α_k est grand, plus h_k sera contraint, cf. figure VII.2), mais nous pourrions considérer la possibilité de les apprendre. Notons également que rien n'impose qu'ils aient tous la même valeur. Dans ce travail, nous n'imposons pas d'a priori sur $\mathbf{E}$.

Comme précédemment dans le modèle non harmoniquement contraint, l'a priori respecte la structure $p(\mathbf{H}) = \prod_{k=1}^{K} p(h_k)$, permettant d'utiliser à nouveau le formalisme de SAGE. La fonctionnelle (VII.15) à minimiser s'écrit désormais :

$$Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\text{c}}{=} \sum_{m=1}^{M} Q_{km}^{MV}(e_{mk}, h_k|\boldsymbol{\theta}') + \log p(h_k) \tag{VII.28}$$

Q_{km}^{MV} est inchangée et il suffit d'incorporer les contributions de l'a priori pour obtenir notre nouvel algorithme. Nous écrivons donc $Q_k^{MAP} = \sum_{m=1}^{M} Q_{km}^{MV} + \log p(h_k)$, somme de la fonctionnelle associée au MV et des contributions dues à l'a priori. Pour $n = 2 \dots N-1$:

$$\nabla_{h_{kn}} Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = \nabla_{h_{kn}} \left(\sum_{m=1}^{M} Q_{km}^{MV}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta}') \right) + \nabla_{h_{kn}} \left(\log p(h_{k(n+1)}|h_{kn}) + \log p(h_{kn}|h_{k(n-1)}) \right) \tag{VII.29}$$

Puisque $\log \mathcal{IG}(u|\alpha,\beta) \stackrel{c}{=} \alpha \log \beta - (\alpha+1)\log u - \beta/u$, nous avons :

$$\nabla_{h_{kn}} Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = -\frac{\alpha_k+1}{h_{k(n+1)}} - \frac{FM+1}{h_{kn}} + \frac{1}{h_{kn}^2}\left(\sum_{f=1}^{F}\sum_{m=1}^{M}\frac{\left|\mu_{kmn}^{post}(f)\right|^2 + \lambda_{kmn}^{post}(f)}{e_{mk}P_{km}(f)} + (\alpha_k+1)h_{k(n-1)}\right)$$

Ainsi, ce gradient est proportionnel à un polynôme du second ordre en h_{kn} :

$$\begin{aligned}\nabla_{h_{kn}} Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) &= \frac{-FM}{h_{kn}^2}\left(p_2 h_{kn}^2 + p_1 h_{kn} - p_0\right)\\ \text{avec } p_2 &= \frac{1}{FM}\frac{\alpha_k+1}{h_{k(n+1)}}\\ p_1 &= 1+\frac{1}{FM}\\ p_0 &= \tilde{h}_{kn} + \frac{\alpha_k+1}{FM}h_{k(n-1)}\end{aligned}$$

où $\tilde{h}_{kn}$ est l'estimateur du MV dans le modèle harmonique (voir équation (VII.22)). Pour $n = N$, le terme $p(h_{k(n+1)}|h_{kn})$ est simplement annulé dans l'équation (VII.29). Pour $n = 1$, la structure en chaîne de Markov impose de choisir un a priori de début de chaîne $p(h_{k1})$. Nous choisissons comme précédemment l'a priori non-informatif de Jeffreys : $p(h_{k1}) \propto 1/h_{k1}$. Les gradients correspondants à ces extrémités de la chaîne s'écrivent :

$$\begin{cases}\nabla_{h_{k1}} Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = -\dfrac{FM}{h_{k1}} + \dfrac{1}{h_{k1}^2}\left(\displaystyle\sum_{f=1}^{F}\sum_{m=1}^{M}\frac{\left|\mu_{kmn}^{post}(f)\right|^2 + \lambda_{kmn}^{post}(f)}{e_{mk}P_{km}(f)}\right) - \dfrac{\alpha_k-1}{h_{k1}} - \dfrac{\alpha_k+1}{h_{k2}}\\ \nabla_{h_{kN}} Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = -\dfrac{FM}{h_{kN}} + \dfrac{1}{h_{kN}^2}\left(\displaystyle\sum_{f=1}^{F}\sum_{m=1}^{M}\frac{\left|\mu_{kmn}^{post}(f)\right|^2 + \lambda_{kmn}^{post}(f)}{e_{mk}P_{km}(f)} + (\alpha_k+1)h_{k(N-1)}\right) - \dfrac{\alpha_k+1}{h_{kN}}\end{cases}$$

Ainsi, nous pouvons présenter d'une manière générique l'estimation de **H** comme la résolution d'équations du premier ou second degré :

$$\nabla_{h_{kn}} Q_k^{MAP}(e_{mk}, h_k|\boldsymbol{\theta'}) = \frac{-FM}{h_{kn}^2}(p_2\, h_{kn}^2 + p_1\, h_{kn} - p_0) \tag{VII.30}$$

où les coefficients p_0, p_1 et p_2 prennent les valeurs indiquées dans la table VII.3 [4].

La règle de mise-à-jour résultante est donnée par l'unique racine positive du polynôme, que nous écrivons comme précédemment :

$$h_{kn}^{(l+1)} = \frac{2p_0}{\sqrt{p_1^2 + 4p_2p_0} + p_1} \tag{VII.31}$$

En l'absence d'a priori sur **E**, nous l'estimons par MV suivant la règle (VII.25) qui demeure inchangée. L'algorithme résultant sera désigné dans la suite par l'abréviation **HS-NMF/EM**.

4. Les valeurs de p_0, p_1, p_2 sont communes à chaque $n \in [2 \dots N-1]$ et diffèrent au bord de la chaîne de Markov ($n = 1$ et $n = N$). Ils dépendent bien sûr de k, n et l ; nous ne reportons pas cette dépendance dans la notation pour des questions de lisibilité.

	p_0	p_1	p_2
$n=1$	$\tilde{h}_{k1}$	$1+\frac{1-\alpha_k}{FM}$	$\frac{1}{FM}\frac{\alpha_k+1}{h_{k2}}$
$n=2\dots N-1$	$\tilde{h}_{kn}+\frac{\alpha_k+1}{FM}h_{k(n-1)}$	$1+\frac{1}{FM}$	$\frac{1}{FM}\frac{\alpha_k+1}{h_{k(n+1)}}$
$n=N$	$\tilde{h}_{kN}+\frac{(\alpha_k+1)}{FM}h_{k(N-1)}$	$1+\frac{1+\alpha_k}{FM}$	0

TABLE VII.3 – Coefficients du polynôme d'ordre 2 à annuler pour mettre à jour h_{kn} dans l'IS-NMF bayésienne harmonique avec a priori en chaîne de Markov. $\tilde{h}_{kn}^{ML}$ désigne la mise-à-jour du MV donnée par l'équation (VII.24).

VII.3.4 Variante multiplicative

De même que nous disposons, pour l'IS-NMF non contrainte, d'un algorithme de descente de gradient multiplicatif, IS-NMF/MU, et d'un algorithme fondé sur un modèle statistique, IS-NMF/EM, nous pouvons envisager de proposer un algorithme multiplicatif pour la NMF harmonique et régularisée, qui serait obtenu par l'heuristique (III.25) en dérivant directement le critère $C_{MAP}(\boldsymbol{\theta})$.

Nous utilisons directement les règles génériques (III.25) (page 59) en choisissant comme fonction de coût à minimiser le critère a posteriori :

$$-C^{MAP}(\boldsymbol{\theta}) = D_{IS}(\mathbf{V}|\mathbf{WH}) - \sum_{k=1}^{K}\log(p(h_k)), \tag{VII.32}$$

où la contribution de l'a priori temporel sur $\mathbf{H}$ peut être vu comme un terme de pénalité. Les mises à jour sur $\mathbf{E}$ sont inchangées et nous obtenons de nouvelles règles multiplicatives pour $\mathbf{H}$. Cependant, des expériences de simulation préliminaires ont révélé que sous ce schéma de mise à jour, le critère ne décroît pas de manière monotone [5]. Par conséquent, nous proposons d'élever le quotient dans (III.25) à une certaine puissance $\eta \in]0,1[$, dont le rôle est similaire à celui du pas dans les descentes de gradient. Nous obtenons les règles suivantes : pour $n = 2\dots N-1$,

$$h_{kn} \leftarrow h_{kn} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{fn}w_{fk}}{\hat{v}_{fn}^2} + \frac{(\alpha_k+1)h_{k,n-1}}{h_{kn}^2}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fn}} + \frac{1}{h_{kn}} + \frac{\alpha_k+1}{h_{k,n+1}}}) \right)^{\eta} \tag{VII.33}$$

Aux bords de la chaîne de Markov :

$$h_{k1} \leftarrow h_{k1} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{f1}w_{fk}}{\hat{v}_{f1}^2} + \frac{\alpha_k}{h_{k,1}}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{f1}} + \frac{1}{h_{k1}} + \frac{\alpha_k+1}{h_{k,2}}} \right)^{\eta} \tag{VII.34}$$

$$h_{kN} \leftarrow h_{kN} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{fN}w_{fk}}{\hat{v}_{fN}^2} + \frac{(\alpha_k+1)h_{k,N-1}}{h_{kN}^2}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fN}} + \frac{\alpha_k+1}{h_{k,N}}} \right)^{\eta} \tag{VII.35}$$

5. En réalité, la pénalité prend le pas sur la reconstruction. C'est certainement ce qui conduit [Virtanen, 2007] à pondérer le terme de pénalité par un poids variable en fonction des données.

Cet algorithme, soumis à publication [Bertin *et al.*, 2009b], sera désigné par l'acronyme « HS-NMF/MU ».

VII.4 Ajout de composantes libres

Au cours de ce chapitre, nous avons imposé au dictionnaire $\mathbf{W}$ d'être purement harmonique, interdisant la représentation de toute autre forme de composante, telle du bruit résiduel ou l'attaque des notes. C'est d'autant plus dommage que les propriétés d'invariance par homothétie de la distance IS sont justement très profitables pour la bonne représentation de ces composantes de faible énergie. Comme on l'observera dans la section IX.2.2 (page 130), D_{IS} a justement « besoin » d'un nombre de composantes légèrement supérieur à l'ordre optimal pour D_{EUC} ou D_{KL} car elle représente toujours ces composantes résiduelles et transitoires, et ce même si cela la « prive » de colonnes pour représenter toutes les notes.

Heureusement, le schéma de mise à jour offert par SAGE permet aisément l'ajout de composantes non contraintes au dictionnaire, puisque les composantes sont mises à jour l'une après l'autre indépendamment, k après k. Il suffit donc d'ajouter au dictionnaire harmonique des composantes non contraintes, mises à jour suivant les règles classiques multiplicatives ou suivant les règles obtenues par MV. On peut ainsi espérer rendre plus propres les composantes harmoniques, tout en capturant une information éventuellement utile sur les transitoires et le bruit. Cet algorithme sera désigné par « **PHS-NMF/EM** » (NMF partiellement harmonique et régularisée).

Chapitre VIII

Quelques résultats de l'approche probabiliste

Résumé

Où, par le biais d'expériences de simulations simples sur des données de synthèse, nous évaluons la consistance des modèles et algorithmes proposés, leurs propriétés de convergence et leur robustesse au mauvais choix de certains paramètres.

VIII.1 Introduction

Avant d'évaluer les algorithmes proposés dans un cadre réel de transcription musicale, des expériences de simulation peuvent nous informer sur l'efficacité de ces algorithmes et leurs propriétés. L'utilisation de données de synthèse permet la connaissance précise d'une vérité-terrain, et la possibilité de contrôler indépendamment chaque paramètre, à la fois lors de la synthèse des données et de leur analyse. Dans ce court chapitre, nous proposons d'évaluer à la fois la consistance des algorithmes (leur capacité à recouvrer les paramètres de données générées suivant le modèle qu'ils utilisent), leurs propriétés de convergence (vitesse et existence de minima locaux), mais aussi l'influence d'une mauvaise estimation d'un paramètre, en particulier l'ordre K et le paramètre de forme α_k des distributions inverse-Gamma. Cette évaluation est conduite en générant des données de synthèse suivant les modèles et de dimensions réduites par rapport à l'application réelle, afin de gagner en possibilités de visualisation et en charge de calcul.

Le section VIII.2 décrit précisément la génération des données et le protocole suivi pour la réalisation des expériences. Les résultats et conclusions de ces simulations sont présentées en section VIII.3.

VIII.2 Cadre expérimental

VIII.2.1 Génération des données

Nous utilisons le modèle génératif (VII.12) pour produire un jeu de données de synthèse. Afin de réduire le coût de calcul et de visualiser plus facilement le résultat, nous choisissons des dimensions réduites pour le problème. L'intervalle des patrons harmoniques est limité à une seule octave ($K \stackrel{\text{def}}{=} K_0 = 12$, notes du do_4 au si_4) et le nombre de filtres du banc ERB (*cf.* section IX.2, page 128) est réduit à $F = 65$, ce qui fournit une résolution fréquentielle suffisante pour cette octave. Nous prenons $N = 200$, ce qui conserve les proportions entre les dimensions du problème dans le cas réel. Enfin, nous fixons le paramètre de forme de la loi inverse-Gamma à $\alpha_k = 10$ pour toutes les composantes k.

Les coefficients e_{mk} sont tirés aléatoirement entre 0 et 1 suivant une distribution uniforme. Les coefficients h_{kn} sont générés suivant le prior en chaîne de Markov (VII.6). Pour ce faire, nous utilisons la propriété suivante : une variable aléatoire Y est de loi inverse-Gamma ssi la variable $X = 1/Y$ est de loi Gamma de mêmes paramètres. Ceci permet la génération de réalisations auxiliaires $g_{kn} \sim \mathcal{G}(\alpha_k, (\alpha_k + 1)g_{k,n-1})$ grâce à la fonction `gamrnd` de Matlab. Pour toutes les composantes, nous choisissons $\alpha_k \stackrel{\text{def}}{=} \alpha_0 = 10$.

Une fois les réalisations $\mathbf{E}$ et $\mathbf{H}$ tirées, l'observation $\mathbf{V}_0$ est formée en suivant le modèle (VII.12), c'est-à-dire en tirant les gaussiennes $\mathbf{d}_{kmn}$ connaissant leurs moyennes et variances, puis en formant la matrice $\mathbf{V}_0 = |X|^2$. Dans la suite, nous signalerons les vérités-terrains par l'indice 0 ($\mathbf{W}_0$, $\mathbf{H}_0$, etc.). En particulier, nous définissons $D_0 \stackrel{\text{def}}{=} D_{IS}(\mathbf{V}_0|W_0\mathbf{H}_0)$ et $C_0 \stackrel{\text{def}}{=} C^{MAP}(\mathbf{W}_0, \mathbf{H}_0)$.

VIII.2.2 Expériences réalisées

Nous factorisons l'observation $\mathbf{V}_0$ en faisant varier l'algorithme utilisé et les valeurs de différents paramètres, notamment K et α_k, afin d'évaluer l'impact d'un mauvais choix de ces paramètres lors de l'analyse. Nous nous intéressons à l'évolution des valeurs des critères pendant l'exécution des algorithmes, la variation de leur valeur finale en fonction des paramètres, ainsi qu'à l'aspect des composantes

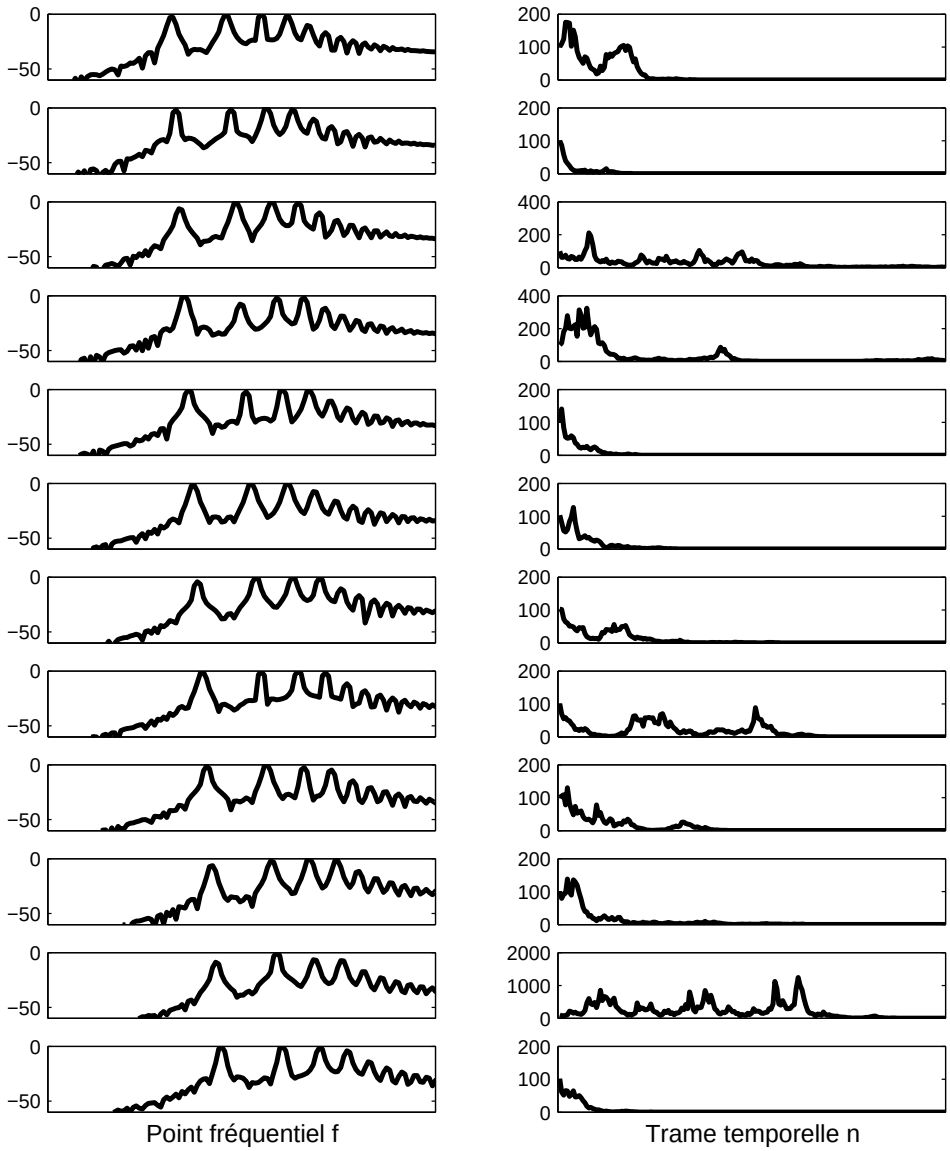


FIGURE VIII.1 – Exemple de données de synthèse générées suivant le modèle (VII.12).

séparées.

VIII.3 Résultats

VIII.3.1 Convergence, vitesse de convergence

Nous suivons et affichons les valeurs des deux critères D_{IS} et C^{MAP} à chaque itération. Nous cherchons en particulier à évaluer l'impact des contraintes sur la vitesse de convergence, l'équilibre entre erreur de reconstruction et critère pénalisé, l'existence éventuelle de minima locaux. On compare également les versions multiplicative et EM de la NMF avec contrainte harmonique et temporelle.

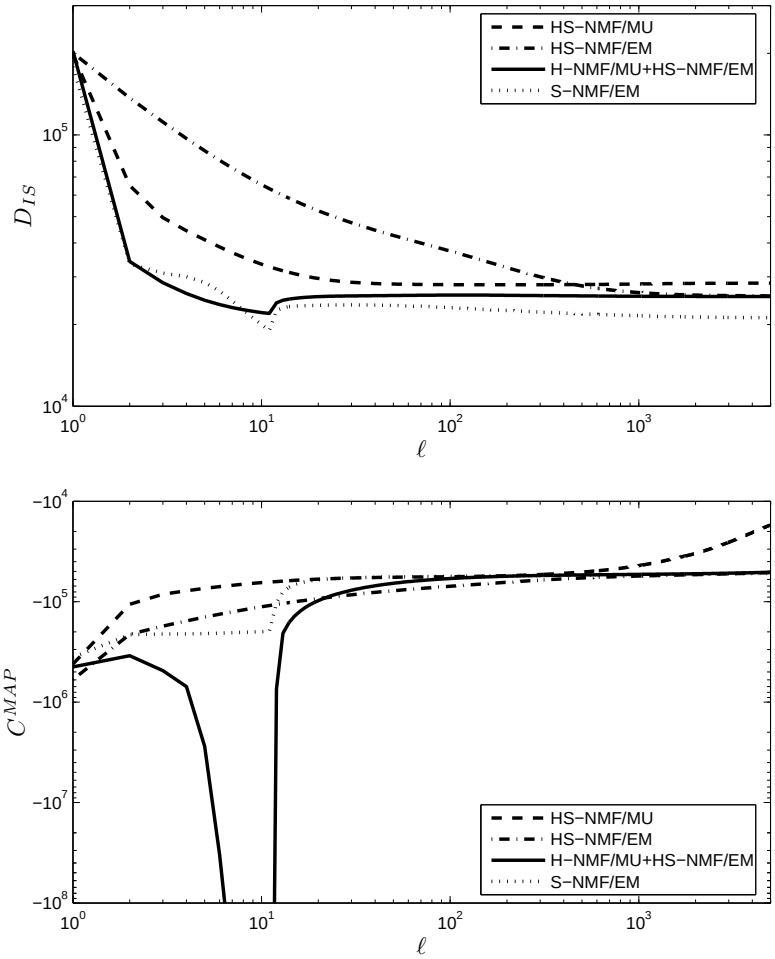


FIGURE VIII.2 – Évolution des critères D_{IS} (en haut) C^{MAP} (en bas) en fonction du nombre d'itérations.

Sur l'exemple présenté à la figure VIII.2, nous voyons que la vitesse initiale de convergence d'une part, et le nombre d'itérations au bout duquel la valeurs des critères se stabilisent d'autre part, sont sensiblement différentes suivant les algorithmes. Les valeurs finales des critères sont comparables, pour des séparations satisfaisantes dans les 4 cas. Comme on pouvait s'y attendre, les algorithmes maximisant le critère C^{MAP} ne font pas décroître D_{IS} de manière monotone.

VIII.3.2 Importance de l'initialisation

L'initialisation de HS-NMF/EM par quelques itérations (ici, 10) de l'algorithme multiplicatif H-NMF/MU accélère considérablement la convergence de l'algorithme, comme on peut le voir sur la figure VIII.2. L'algorithme avec initialisation multiplicative atteint rapidement une valeur faible et stable de D_{IS}, puis emploie les itérations suivantes à affiner la représentation par la contrainte de régularité temporelle, faisant croître C^{MAP}. La chute importante de C^{MAP} au cours de la phase d'initialisation (tronquée sur la figure pour conserver la lisibilité) signale que l'algorithme complet se préoccupe d'abord de minimiser l'erreur de reconstruction, puis de régulariser les enveloppes, ce qui est souhaitable. À l'inverse, le même algorithme initialisé directement au hasard fait décroître lentement l'erreur de reconstruction, laissant à penser que les deux termes (erreur et régularité temporelle) sont en compétition. De plus, même si la valeur finale du critère est la même, il faut noter que les composantes séparées par l'algorithme avec phase d'initialisation multiplicative ont un aspect bien plus proche de la vérité-terrain et moins bruité que celles obtenues sans (ce qui confirme expérimentalement la théorie, qui prévoit que les algorithmes SAGE convergent vers un minimum local, pourvu qu'ils soient initialisés dans un certain voisinage de ce minimum).

VIII.3.3 Robustesse au choix de l'ordre

Le nombre de composantes à la synthèse étant fixé à $K_0 = 12$, on fait varier le nombre de composantes K à l'analyse de $K = 6$ à $K = 24$ par pas de 2, ce qui revient de proche en proche à enlever la note la plus grave et la plus aiguë du dictionnaire (lorsque $K < K_0$) ou au contraire à ajouter une note supplémentaire à chaque extrémité du dictionnaire.

Sur la figure VIII.3, on peut observer la valeur des critères de reconstruction et de régularité obtenus par HS-NMF/EM, en fonction de l'erreur commise sur le choix de l'ordre. Les lignes pointillées repèrent les « valeurs-cibles » des critères (D_0 et C_0) [1]. La divergence d'Itakura-Saito est robuste à une surestimation de l'ordre K ; plus élevée quand $K < K_0$, elle reste quasi constante lorsque $K \geq K_0$. En revanche, le critère régularisé C^{MAP} présente un net maximum lorsque $K = 10$ (sa valeur en $K = K_0 = 12$ étant cependant presque égale). On pourrait voir cette propriété comme une heuristique d'estimation de l'ordre K optimal (à condition de disposer conjointement d'une estimation de α_0).

VIII.3.4 Robustesse au choix du paramètre de forme

Les données étant générées avec $\alpha_0 = 10$, on évalue l'impact d'une mauvaise estimation de ce paramètre lors de la factorisation en faisant varier α dans l'analyse.

On calcule la valeur du coût théorique à minimiser (c'est-à-dire $C^{MAP}(\boldsymbol{\theta})$ avec $\alpha = 10$). Le cas $\alpha_k = 0$ correspond au cas non régularisé (aucune contrainte sur $\mathbf{H}$). Le résultat est présenté sur la figure VIII.4. À partir de $\alpha_k = 5$ et jusqu'à $\alpha_k = 50$, ce coût est quasi stable et supérieur à la valeur-cible D_0 (tracée en pointillés), ce qui suggère une bonne robustesse de l'algorithme au choix de ce paramètre.

Nous affichons également, sur la figure VIII.5, l'activation temporelle d'une composante ($k = 11$) pour différentes valeurs de α_{11}, ainsi que la vérité-terrain.

1. Notons qu'en raison du mode de génération des données, faisant intervenir le tirage de quantités aléatoires, cette valeur n'est pas à proprement parler une vérité-terrain, ni nécessairement un optimum, ce qui explique qu'elle puisse être dépassée par la valeur réelle du critère. Elle constitue seulement un ordre de grandeur indicatif.

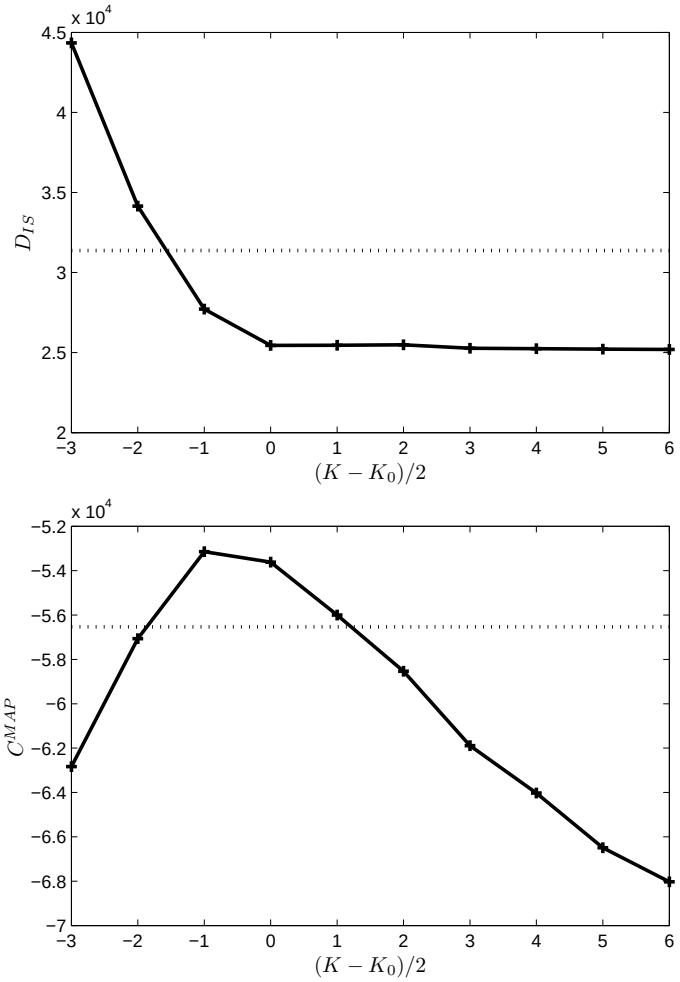


FIGURE VIII.3 – Valeurs finales des critères en fonction de l'erreur d'estimation $K - K_0$.

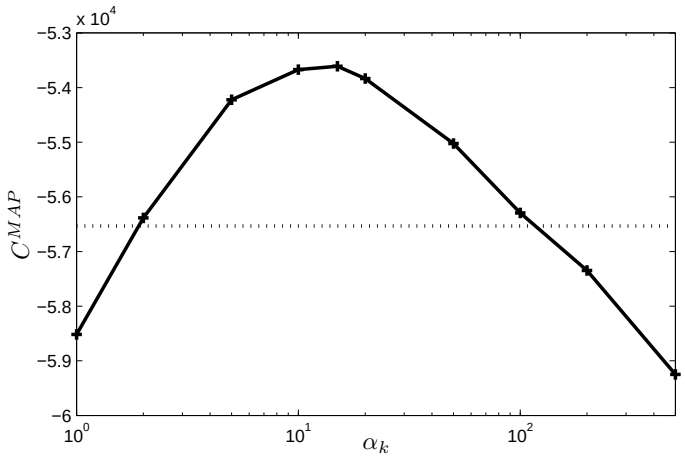


FIGURE VIII.4 – Valeur finale du coût en fonction de α_k.

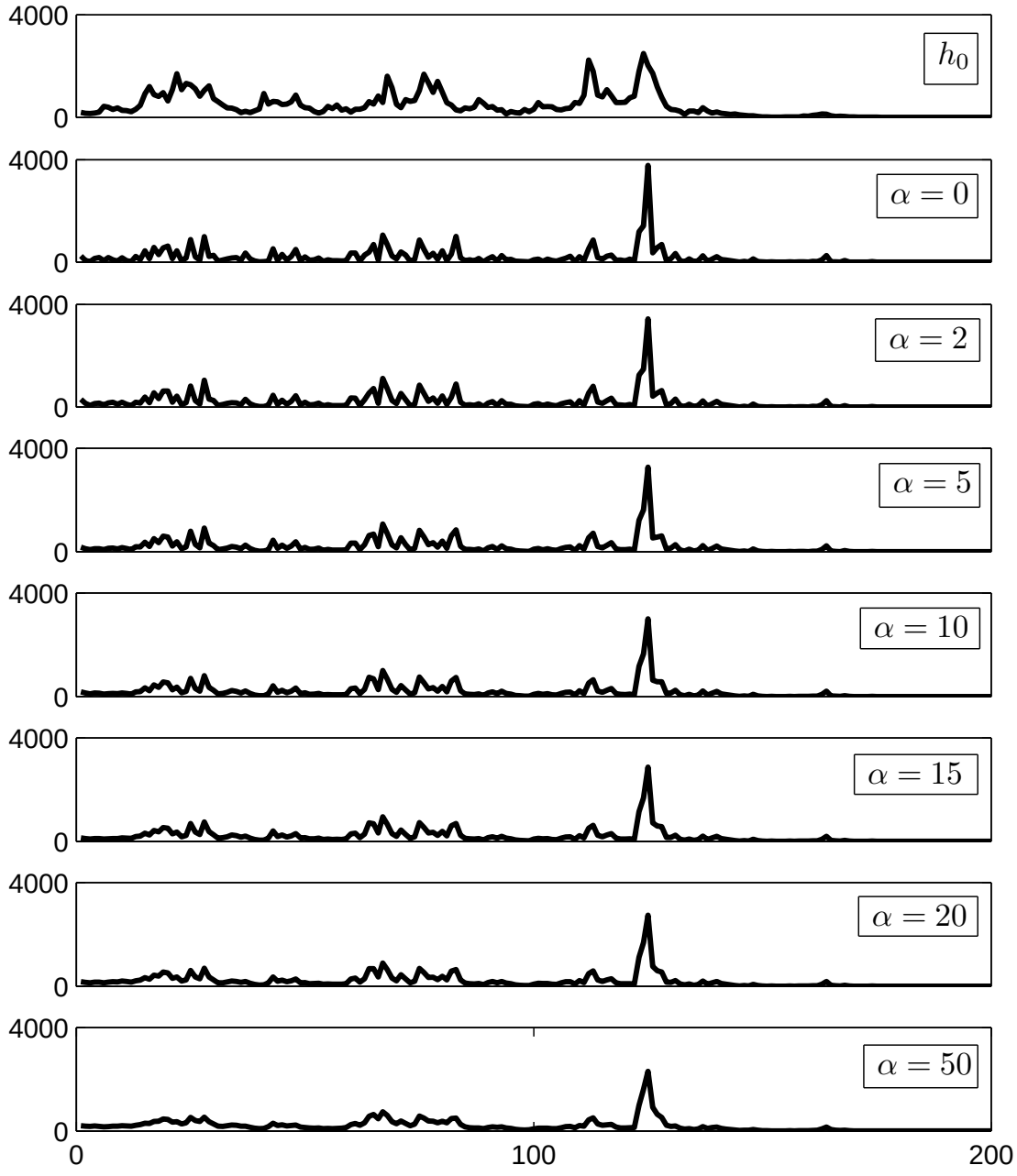


FIGURE VIII.5 – Une composante originale et les composantes recouvrées correspondantes, pour différentes valeurs de α.

Si nous pouvons deviner une régularité grandissante à mesure qu'α_k croît, celle-ci reste discrète, ce qui laisse supposer une bonne marge de manœuvre pour le choix de α_k.

VIII.3.5 Composantes libres

À l'expérimentation, il s'avère que les composantes libres sont mises à jour beaucoup plus rapidement que les composantes harmoniques ; en effet, l'algorithme multiplicatif IS-NMF/MU qui les adapte converge en beaucoup moins d'itérations que HS-NMF/EM qui met à jour les atomes harmoniques. De ce fait, les composantes libres capturent rapidement une grande partie de l'information, qui n'est du coup pas représentée par le reste du dictionnaire. Il en résulte une décomposition très médiocre.

Par conséquent, nous proposons d'utiliser plutôt la version multiplicative HS-NMF/MU, plus rapide, de manière à équilibrer les contributions des parties contrainte et libre du dictionnaire. Cet algorithme sera désigné par l'acronyme « **PHS-NMF/MU** » et évalué directement sur l'application de transcription dans la partie suivante.

Quatrième partie

Application à la transcription

Chapitre IX

Tâche, protocole et évaluation

Résumé

Où l'on présente le protocole expérimental choisi pour évaluer les performances des algorithmes de NMF précédemment mis au point, intégrés au sein d'un système de transcription complet qui est également présenté.

IX.1 Introduction

Les algorithmes de NMF développés au cours des chapitres précédents produisent une représentation $\mathbf{W}, \mathbf{H}$ de mi-niveau intéressante, mais qui ne fournit pas une transcription telle que nous l'avons définie au chapitre I. Dans la section IX.2 de ce chapitre, nous détaillons le système complet de transcription qui utilise ces algorithmes pour convertir la forme d'onde en fichier MIDI. Nous en profitons pour présenter et motiver les choix de différents modules et paramètres de ce système, tels que la représentation temps-fréquence à factoriser ou l'ordre du modèle. Nous présentons ensuite, dans la section IX.3, le protocole que nous adoptons pour évaluer les performances de ces systèmes.

IX.2 Système complet de transcription

À notre connaissance, le principe général de ce système et une partie de ses modules ont été proposés pour la première fois dans nos travaux [Bertin *et al.*, 2007]. Nous décrivons ici ses parties et discutons nos choix.

IX.2.1 Représentation temps-fréquence

Le premier choix à faire est celui de la représentation $\mathbf{V}$. Au cours de ce mémoire, nous avons évoqué et utilisé les deux principaux, la transformée de Fourier à court-terme (TFCT) et la représentation en bandes rectangulaires équivalentes (ERB), mais d'autres choix sont possibles.

IX.2.1.1 Transformée de Fourier à court-terme

La transformée de Fourier à court-terme d'un signal x décrit le contenu fréquentiel de x au cours du temps. Elle est calculée comme la transformée de Fourier discrète de trames successives, qui sont de courts segments de signal multipliés par une fenêtre w de la même longueur. Ainsi, chaque point temps-fréquence de la TFCT est donné par :

$$TFCT[x](f,t) = \sum_{n=0}^{N-1} x(n+t)w(n)e^{-i2\pi\frac{fn}{N}} \tag{IX.1}$$

En prenant le module, ou le module au carré de chacun de ces coefficients, nous obtenons une représentation temps-fréquence $\mathbf{V}$ (spectrogramme d'amplitude ou de puissance) factorisable par NMF.

Le problème bien connu de la TFCT est celui du compromis de résolution temps-fréquence. En effet, la résolution fréquentielle de cette transformée est directement liée à la taille de la fenêtre et à la largeur du lobe principal de son spectre. Ce compromis est problématique pour des signaux de musique. Nous pouvons nous en convaincre en prenant quelques ordres de grandeurs. Dans un morceau comportant des doubles croches (1/4 de temps) à un tempo de 120 à la noire, huit notes sont jouées chaque seconde, ce qui nécessite une résolution temporelle d'au plus 125 ms, soit 5512 échantillons d'un signal à 44100 Hz (fréquence d'échantillonnage du CD). Par ailleurs, la différence de fréquences entre le premier *do* du piano et le *do*♯ qui le suit immédiatement est de 2 Hz (voir table B.1 page 188), ce qui ne peut être résolu qu'avec un nombre de points fréquentiels supérieur à 22050. Ces deux conditions sont incompatibles, ce qui impose de faire un compromis.

Toutefois, les notes les plus graves de la tessiture du piano sont rarement employées, et on peut espérer que la NMF ne souffre pas autant du compromis temps-fréquence que d'autres applications, étant donné les approximations sur lesquelles elle repose.

IX.2.1.2 Transformée en bandes rectangulaires équivalentes

Afin de préserver une résolution minimale d'un demi-ton sur l'ensemble de la tessiture, la TFCT doit être calculée avec une fenêtre relativement longue (nous avons proposé 64 ms dans [Bertin *et al.*, 2007]), ce qui implique à la fois une résolution temporelle médiocre, et un coût de calcul long et superflu étant donné la redondance de l'information en hautes fréquences sur de telles fenêtres. De plus, les fréquences des notes de musique sont réparties sur une échelle non linéaire (en gamme tempérée, on passe d'une note à la suivante en multipliant sa fréquence par $2^{1/12}$). On risque donc d'avoir beaucoup trop de points fréquentiels en haute fréquence par rapport aux notes à discriminer.

Une représentation de dimensions plus réduites, offrant une meilleure résolution temporelle dans le registre hautes fréquences, peut être obtenue en choisissant une échelle non linéaire de fréquences. Nous utilisons la représentation proposée dans [Vincent, 2006] (dans le cadre de modèles spécifiques par instruments) et motivée par des arguments perceptifs. Le signal est filtré par un banc de filtres de 257 fenêtres de Hanning modulées par une sinusoïde, et dont les fréquences centrales sont comprises entre 5 Hz and 10.8 kHz et linéairement espacées sur l'échelle des Bandes Rectangulaires Équivalentes (ERB, [Zwicker et Fastl, 1999, van de Par *et al.*, 2002]). L'équivalence entre fréquences en échelle ERB et exprimées en Hertz est définie par la relation :

$$f_{\mathrm{ERB}} \stackrel{\mathrm{def}}{=} 9.26 \log(0.00437 f_{\mathrm{Hz}} + 1) \tag{IX.2}$$

Le gain de la réponse fréquentielle G_i du i^e filtre peut être calculé analytiquement comme une combinaison de sinus cardinaux. La longueur de chaque filtre est choisie de manière que la bande-passante de son lobe principal soit égale à 4 fois la différence entre sa fréquence centrale et celles des filtres adjacents, ce qui introduit une forme de lissage spectral favorable à la NMF. Après filtrage, chaque sous-bande est segmentée en trames disjointes de 23 ms, et l'énergie dans chaque trame est calculée.

Sur les expériences menées dans [Vincent *et al.*, 2008], nous avons pu constater que ce choix ne modifiait pas significativement la performance de transcription (comparée à la TFCT), mais réduisait de manière importante le temps de calcul.

IX.2.1.3 Autres

Il existe une très grande quantité de représentations temps-fréquence d'un signal. Un inventaire exhaustif de ces représentations est par exemple disponible dans [Demars, 2004]. Dans la littérature consacrée à la NMF appliquée au signal audio, on peut notamment trouver l'utilisation du spectrogramme réalloué et de la transformée à Q constant, que nous décrivons brièvement.

Spectrogramme réalloué Si une fréquence du signal se trouve exactement au milieu de deux fréquences centrales de la TFCT, son énergie va être dissipée sur deux points fréquentiels. L'idée du spectrogramme réalloué [Auger et Flandrin, 1995] est d'ajuster les points temps-fréquences au signal en considérant les points voisins et en déplaçant le point de calcul au « centre de gravité local » de

l'énergie du signal. Son calcul s'appuie sur les dérivées en temps et en fréquence de la phase de la TFCT. Le spectrogramme réalloué a été utilisé comme entrée d'un système de NMF appliqué à la séparation de voix parlée dans [Segbroeck et hamme, 2009].

Transformée à facteur de qualité constant Il est possible de représenter le spectre non plus selon une échelle linéaire en fréquence, mais plutôt selon une échelle logarithmique. C'est le principe des analyses à Q constant, où Q désigne le facteur de qualité d'un filtre, c'est à dire le rapport $F/\Delta F$ entre sa fréquence propre et sa largeur de bande [Brown, 1991]. L'avantage de cette méthode vient du fait qu'un spectre harmonique présente, en échelle logarithmique, une structure toujours identique quelle que soit sa fréquence fondamentale, avec simplement un décalage de l'origine. Ainsi leur structure est indépendante du fondamental dont la valeur se retrouve uniquement dans la position de la première raie (à $\log(f_0)$). Le calcul du spectre à Q constant peut être effectué de différentes manières (FFT, transformée en ondelettes). Cependant, il n'est pas nativement exprimé sous une forme matricielle, car les fenêtres d'analyse des différentes bandes sont de longueur différente. L'utilisation de cette transformée pour la NMF demande donc de la modifier pour la rendre matricielle. De plus, les fenêtres, très longues dans les basses fréquences, introduisent un important pré-écho. Pour ces raisons, nous préférons la transformée ERB précédente qui constitue un meilleur compromis entre la TFCT et le banc à Q constant.

IX.2.2 Choix de l'ordre du modèle

Puisque nous cherchons à obtenir un dictionnaire $\mathbf{W}$ de notes, l'intuition suggère de choisir pour ordre K du modèle une valeur que l'on peut penser proche du nombre de notes contenues dans le morceau, de manière à obtenir une représentation de $|\mathbf{X}|^2$ porteuse de sens, où chaque composante contiendrait exactement une hauteur musicale. On peut vérifier sur un exemple simple que cette intuition est trompeuse.

En effet, si l'on analyse l'exemple désormais canonique (figure II.9 page 41) avec les trois distances usuelles et différentes valeurs de l'ordre K, nous nous apercevons que le choix intuitif $K = 4$ n'est pas le meilleur. Par exemple, sur la figure IX.1 qui représente la factorisation obtenue par IS-NMF, on observe que les composantes extraites n'ont pas de structure clairement monopitch ; la première semble mêler une note avec un phénomène transitoire et est activée à chaque attaque quel que soit le pitch, les deux suivantes sont des mélanges de hauteur, tandis que la dernière a le spectre d'un bruit résiduel (avec une amplitude très inférieure aux trois autres, ce qui est clairement une conséquence de l'invariance par homothétie de la divergence IS).

EUC-NMF et KL-NMF produisent des résultats comparables : deux composantes monopitch, une composante mêlant les deux hauteurs restantes, et une composante transitoire. À l'ordre $K = 5$, KL-NMF parvient à séparer les 4 notes, avec une cinquième composante représentant les attaques, mais dont l'activation est bruitée (oscillations d'amplitude). Étonnamment, EUC-NMF ne sépare pas les quatre notes ; le dictionnaire conserve une composante multipitch, et utilise deux atomes pour représenter plus finement l'une des notes déjà séparées. En ce qui concerne IS-NMF, dont le résultat à $K = 5$ est présenté sur la figure IX.2, nous pouvons voir que deux composantes monopitch apparaissent, mais la séparation n'est toujours pas complète ; il y a désormais deux composantes non harmoniques, l'une représentant le bruit résiduel, l'autre les transitoires d'attaque.

La séparation complète des quatre notes est obtenue à $K = 6$ pour la distance euclidienne et la divergence IS (figure IX.3). En revanche, KL-NMF produit une composante harmonique dont le pitch

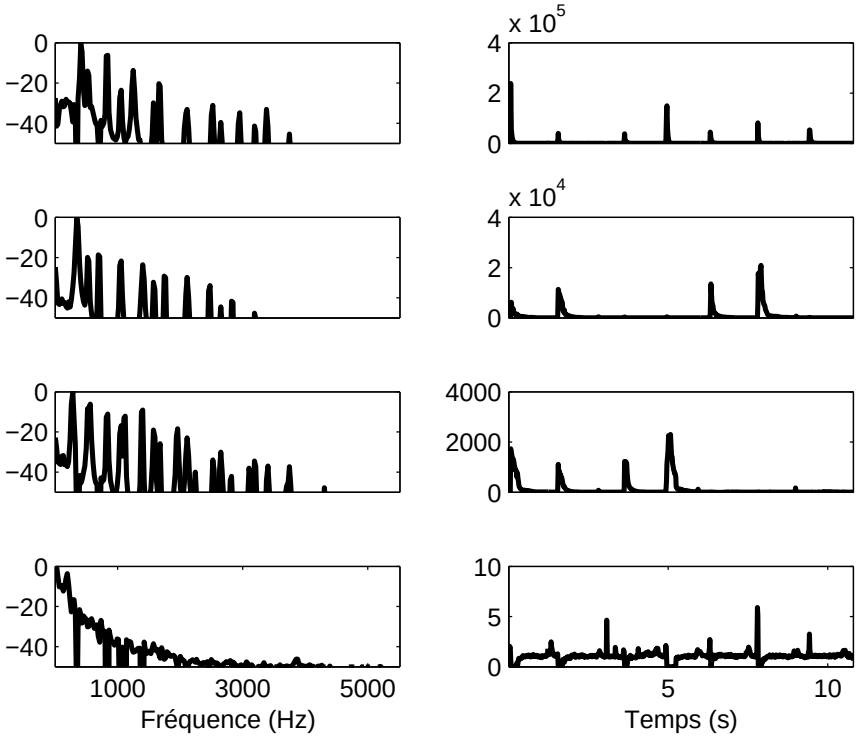


FIGURE IX.1 – IS-NMF/MU avec $K = 4$.

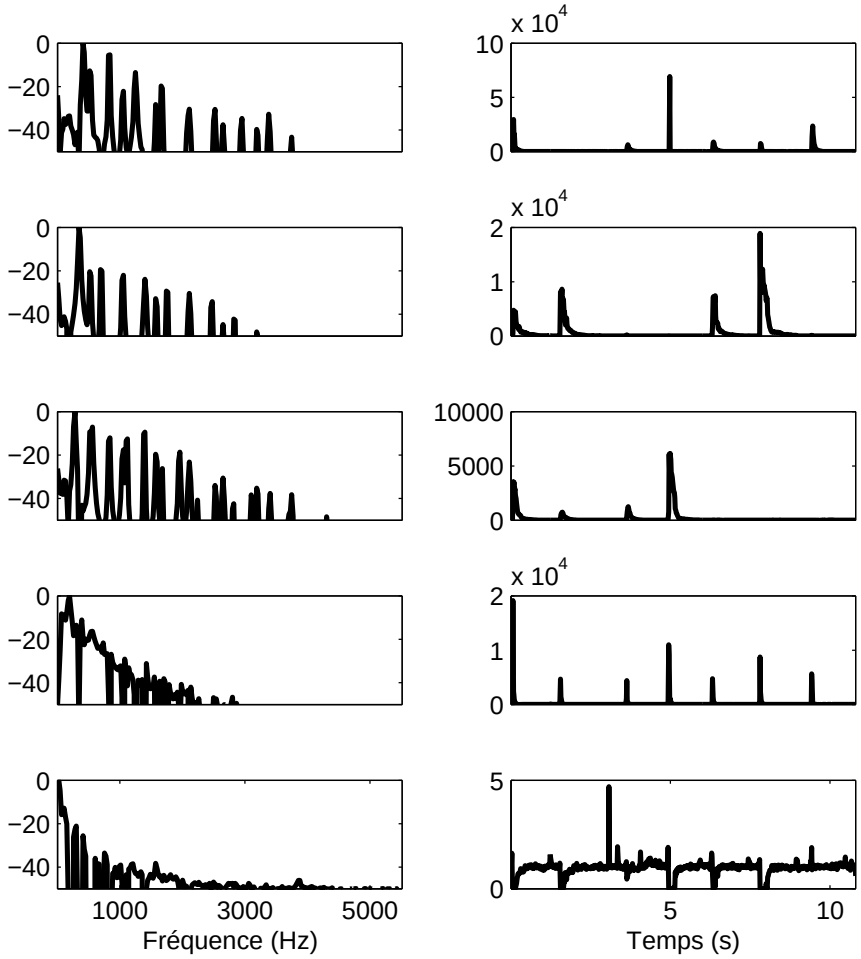


FIGURE IX.2 – IS-NMF/MU avec $K = 5$.

est absent de la vérité-terrain, sans interprétation évidente. IS-NMF produit une factorisation à l'aspect visuel beaucoup plus propre que les deux autres coûts.

Lorsque l'on continue de faire croître K au-delà de 6, EUC-NMF et KL-NMF tendent à séparer une composante en plusieurs (par exemple, une composante représentera la phase de décroissance, une autre la phase d'entretien de la note). Ce n'est pas surprenant, puisque l'on sait que l'enveloppe spectrale de la note évolue au cours du temps, avec en particulier une extinction précoce des partiels d'ordre élevé dans le cas du piano (le modèle NMF est ici mis en défaut). En revanche, lorsque l'on augmente l'ordre de décomposition IS-NMF, les composantes ajoutées sont principalement utilisées pour raffiner la représentation des bruits de faible énergie.

Cette étude tend à favoriser une surestimation de l'ordre du modèle par rapport au nombre de notes attendues, en particulier lorsque la divergence IS est utilisée. Dans le cas non contraint, le choix $K = 88$ (nombre de touches du clavier de piano) paraît raisonnable, puisqu'il est rare que toutes les notes du piano soient jouées dans un même morceau. Dans le cas où l'on utilise la contrainte d'harmonicité, la question est évidemment différente : le choix des patrons P_{km} et l'initialisation imposent un choix plus strict, mais la contrainte prévient le genre de phénomènes observés ci-dessus (séparation d'une note en plusieurs composantes, fusion de deux notes en une seule...). Dans [Vincent *et al.*, 2007], il est observé que l'augmentation de la taille du dictionnaire à deux composantes par hauteur n'influe pas significativement les performances. Nous choisissons simplement $K_h = 88$, soit une composante par touche de piano.

Le comportement de IS-NMF tend cependant à convaincre de l'importance de laisser suffisamment de composantes libres pour représenter les transitoires et le bruit résiduel. Dans la suite, nous avons donc fixé le nombre de composantes libres des dictionnaires harmoniques à $K_f = 12$, ce qui porte le nombre total de composantes à $K = 100$.

IX.2.3 Factorisation

Les différents algorithmes de factorisation mis en œuvre sont résumés dans la table IX.2 (page 141).

IX.2.4 Post-traitement

Une fois les facteurs $\mathbf{W}$ et $\mathbf{H}$ obtenus, il est nécessaire de les convertir sous forme de notes de musiques caractérisées par leur hauteur sur l'échelle MIDI et leurs instants de début et de fin.

Les spectres de la base peuvent posséder ou non une hauteur. Dans le cas des NMF avec contrainte harmonique, seules les éventuelles composantes libres peuvent ne pas être harmoniques. Les autres peuvent être directement étiquetées par la f_0 utlisée pour l'initialisation des patrons P_{km}. Dans le cas des NMF dont le dictionnaire est libre, rien ne peut être dit a priori sur la présence d'une ou plusieurs hauteurs dans la base ; il faut donc les estimer.

Le produit spectral utilisé dans notre premier système [Bertin *et al.*, 2007] présente un défaut : il risque d'échouer si certains partiels ont une amplitude nulle ou très faible. Or, en cas de dictionnaire libre, une même note peut être représenté par plusieurs éléments de la base possédant la même fondamentale mais des partiels d'amplitudes différentes. Pour pallier ce problème, nous proposons d'utiliser une technique basée sur le filtre sinusoïdal en peigne de [Vincent et Plumbley, 2005]. On calcule les produits scalaires des colonnes $\mathbf{w}_k$ avec des filtres en peigne à toutes les f_0 de la gamme chromatique sur le registre du piano, c'est-à-dire pour des fréquences comprises entre 27 et 4200 Hz ($p_{min} = 21$ et $p_{max} = 108$, voir table B.1 page 188). La fréquence f_0 est celle du filtre pour lequel le produit scalaire

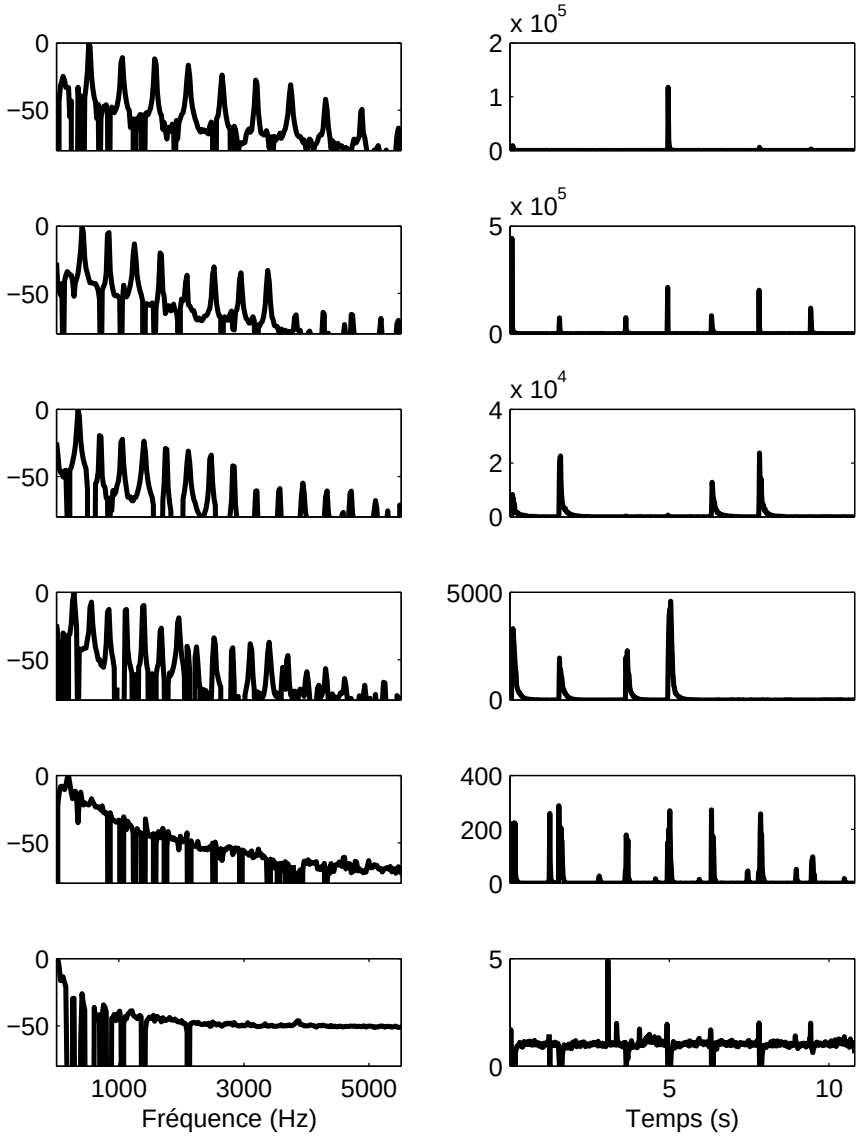


FIGURE IX.3 – IS-NMF/MU avec $K = 6$.

est minimal :

$$f_0^k = \arg\min_{f_0} \sum_{f=1}^{F} w_{fk}^2 \left[1 - \cos(2\pi\nu_f / f_0)\right] \tag{IX.3}$$

où ν_f est la fréquence centrale du f^e filtre ERB. Les fréquences sont ensuite quantifiées en pitch MIDI avec une résolution d'un demi-ton, suivant la formule :

$$f_0 = 2^{\frac{p_{MIDI}-69}{12}} \times 440 \tag{IX.4}$$

Si les produits scalaires d'une composante avec chaque filtre en peigne sont tous égaux, on considère que la composante ne possède pas de hauteur. Rien n'empêchant une composante de posséder plusieurs hauteurs, on pourrait s'interroger sur l'opportunité d'effectuer une estimation multipitch à ce stade. En pratique, cela s'est avéré inutile : pourvu que l'ordre K soit suffisant, la NMF produit peu de composantes multipitch, et, lorsqu'on est dans ce cas, une des hauteurs est nettement prédominante, les autres étant par ailleurs représentées dans d'autres composantes du dictionnaire.

Une fois les hauteurs attribuées aux composantes, une unique enveloppe temporelle est associée à chaque hauteur discrète, en sommant les lignes de $\mathbf{H}$ qui sont associées aux composantes de la base possédant cette hauteur, et en prenant la racine carrée de leur puissance totale dans chaque trame temporelle. Ces enveloppes $\bar{\mathbf{H}}$ sont ensuite traitées pour détecter les attaques des notes : pour ce faire, nous utilisons un simple seuillage ligne par ligne. Une note est détectée dans la k^e enveloppe et à la n^e trame temporelle si son amplitude $\bar{h}_{kn}$ vérifie :

$$\bar{h}_{kn} \geq 10^{-A_{dB}/20} \max_{jm} \bar{h}_{jm} \tag{IX.5}$$

où A_{dB} est un seuil choisi manuellement. Les notes dont la durée est inférieure à 50 ms sont supprimées de la transcription.

IX.2.5 Résumé

Munis de tous ces éléments, il ne nous reste plus qu'à schématiser l'ensemble du système de transcription, dont l'entrée sera le fichier son et la sortie le fichier MIDI. Pour ce faire, nous utilisons dans cette section les outils de la méthode SADT (*Structured Analysis Design Technique*, technique structurée d'analyse et de conception). [1]

Cette technique d'analyse systémique hiérarchique propose de représenter chaque système et ses composantes par une boîte exprimant sa fonction, et quatre types d'entrées-sorties à l'emplacement normalisé autour de la boîte, définies par le code dit « MECS » (Moyens-Entrées-Contrôle-Sorties). La figure IX.4 présente cette forme générique. Le lecteur intéressé pourra se reporter à [Lissandre, 1990].

Nous pouvons représenter les deux premiers niveaux hiérarchiques de notre système de transcription. Au niveau le plus haut (dit « étage A moins zéro »), la fonction dite « globale » du système est de transcrire la musique. Son schéma SADT est donné figure IX.5. En détaillant davantage le système en parties fonctionnelles, nous obtenons le diagramme de la figure IX.6, qui détaille les fonctions principales du système, les entrées-sorties de chaque élément et les moyens et données de contrôle de ces boîtes. Nous obtenons ainsi un schéma global décrivant le fonctionnement du système.

1. ®SADT est une marque déposée de SofTech (USA) et d'IGL Technologie (France), développée aux USA par Doug Ross en 1977 et introduite en Europe à partir de 1982 par Michel Galiner.

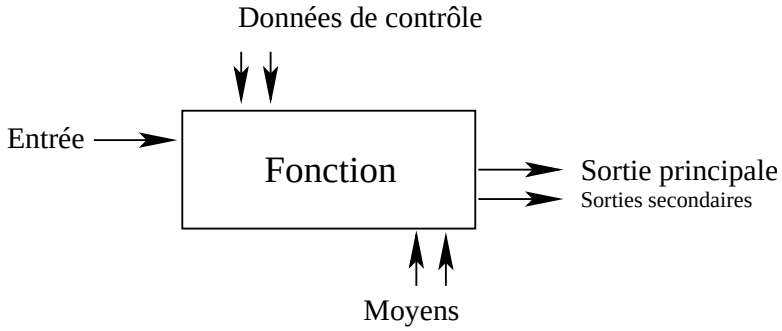


FIGURE IX.4 – Forme générique d'une boîte élémentaire SADT.

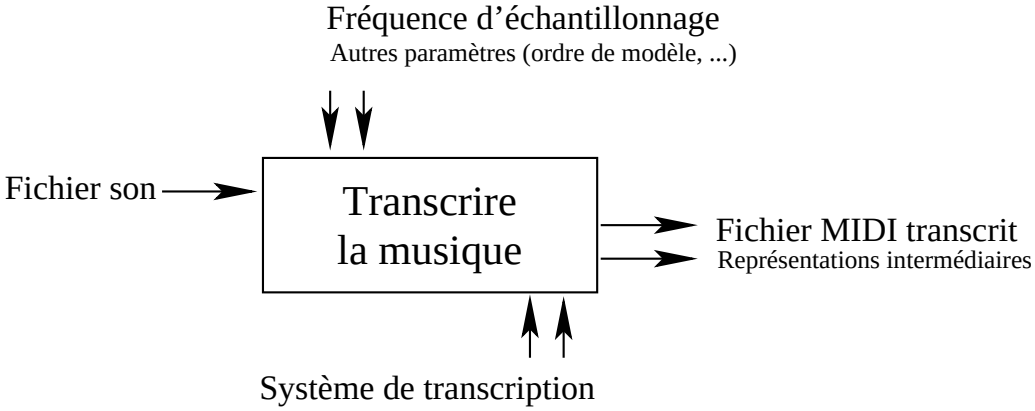


FIGURE IX.5 – Étage SADT A⟍0 du système de transcription.

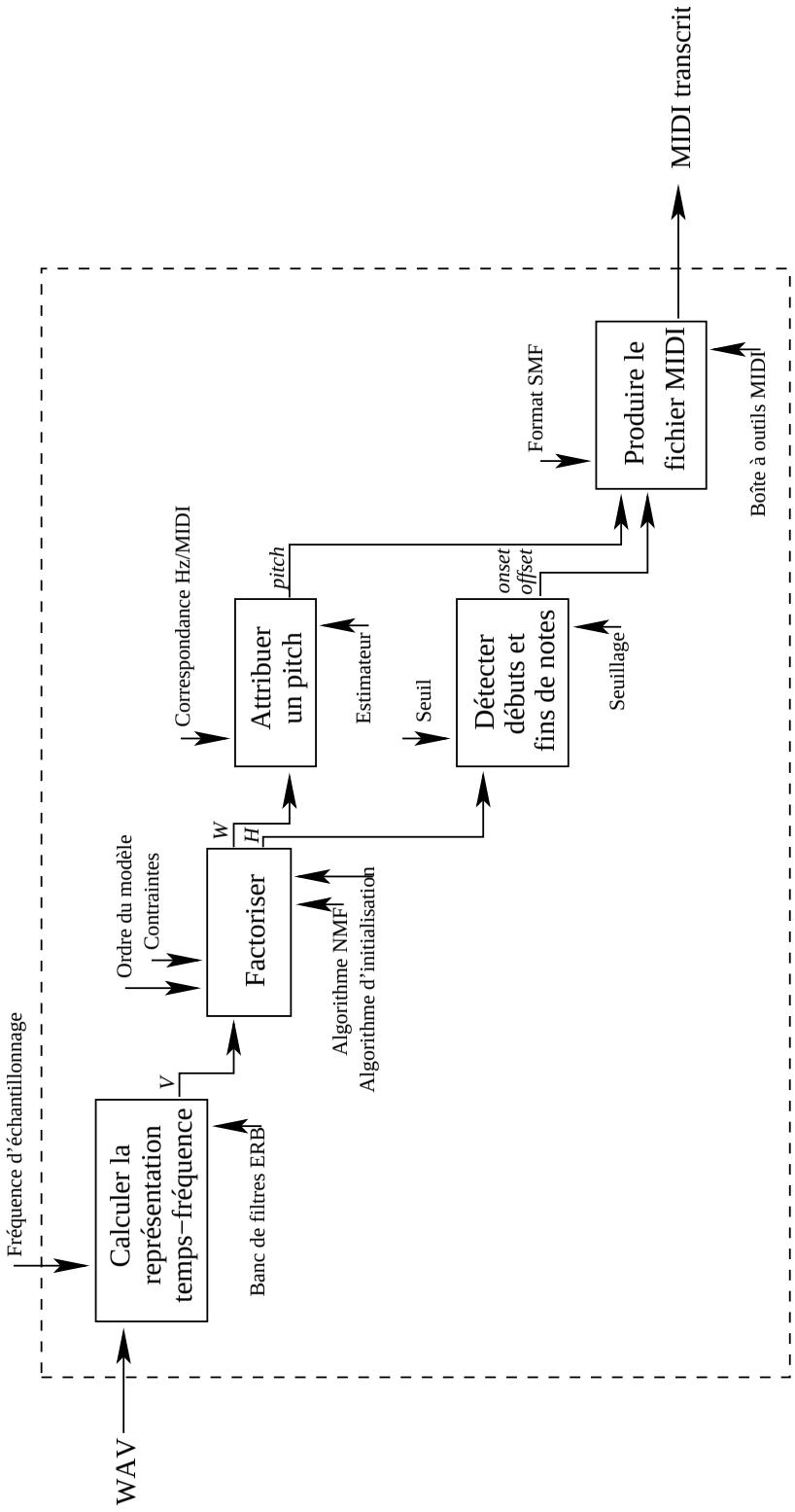


FIGURE IX.6 – Étage SADT A0 d'un système de transcription NMF.

IX.3 Description du cadre expérimental

IX.3.1 Tâche

Nous nous donnons pour objectif la tâche restreinte présentée à la section I.2.3 (page 22), c'est-à-dire la production d'une liste de notes de musique caractérisées par leur hauteur musicale (sur l'échelle chromatique, donc avec une précision d'un demi-ton) et leurs instants de début et de fin en secondes.

IX.3.2 Bases de données

Afin d'évaluer et de quantifier les performances de transcription, nous avons besoin d'un ensemble de pièces musicales accompagnées d'une annotation MIDI précise de leur contenu. Les moyens les plus simples d'acquérir de telles données sont de deux ordres : l'enregistrement d'un instrument MIDI (l'acquisition de l'audio et du MIDI étant simultanés), ou bien la synthèse de son à partir de fichiers MIDI préalablement disponibles. Pour des raisons de réalisme du timbre et de facilité d'acquisition, le piano est un instrument de choix : de nombreux synthétiseurs logiciels de très haute qualité sont disponibles dans le commerce, d'une part, et d'autre part un piano acoustique peut être équipé de systèmes mécaniques et électroniques permettant de déclencher les touches et d'enregistrer une sortie MIDI, tout en conservant le timbre d'un vrai instrument acoustique.

C'est le procédé adopté dans [Emiya, 2008, Emiya *et al.*, 2009] pour constituer la base de données MAPS (*MIDI-Aligned Piano Sounds*). Cette base, librement distribuée, inclut des enregistrements de notes isolées, d'accords tonals et aléatoires et de pièces complètes issues du répertoire classique du piano. Les sons sont soit produits par des logiciels de synthèse de très bonne qualité, soit enregistrés sur un Yamaha DisKlavier (piano droit équipé). De cette base très complète, nous avons extrait deux sous-ensembles de données pour nos évaluations : un sous-ensemble synthétique (logiciel Akoustik Piano de Native Instruments, pré-réglage « Bechstein Bach », dont les échantillons sont issus d'un piano Bechstein 280, désignée par l'abréviation **AkPnBcht**), et un sous-ensemble issus des enregistrements de DisKlavier réalisés à Télécom ParisTech (**ENSTDkAm**). Chaque sous-ensemble est composé de 30 pièces de 30 secondes chacune (les morceaux originaux de MAPS ont été tronqués).

IX.3.3 Évaluation des performances

La seule observation des factorisations extraites et de leur allure sur des exemples simples, telle que nous l'avons faite jusqu'ici, ne peut suffire à établir clairement l'intérêt d'une méthode de transcription en conditions réelles. Si l'illustration des performances sur quelques exemples, comme dans [Moorer, 1975, Walmsley *et al.*, 1999, Smaragdis et Brown, 2003, Davy *et al.*, 2006], permet de mettre en lumière certaines erreurs typiques et de se faire une idée du fonctionnement du système, l'évaluation quantitative demeure un critère de poids dans l'évaluation des performances. Nous nous limiterons ici aux évaluations basées sur un dénombrement des notes détectées ou manquantes, qui sont les plus répandues. Signalons toutefois que des méthodes plus perfectionnées, et perceptivement motivées sont discutées dans [Emiya, 2008, Daniel *et al.*, 2008].

Pour définir des critères quantitatifs d'évaluation, on détermine d'abord l'ensemble TP des notes correctement estimées (*true positive*), l'ensemble FP des notes ajoutées (*false positive*, ou fausses alarmes), et l'ensemble FN des notes oubliées (*false negative*). Une note est considérée comme correcte si son pitch (en numérotation MIDI) est identique à celui d'une note de la vérité-terrain, et si son onset a lieu dans un certain intervalle de temps autour de cette note (50 ms pour [Bello *et al.*, 2006, Vincent

et al., 2008], 70 ms pour [Dixon, 2000],Dixon [2000], 128 ms pour [Bertin *et al.*, 2007], 150 ms pour [Ryynänen et Klapuri, 2005]). Une fausse alarme est une note transcrite alors qu'elle est absente de la référence, tandis qu'une note est oubliée si elle est présente dans la vérité-terrain mais pas transcrite. En fonction des cardinaux de ces ensembles, on définit alors deux critères complémentaires, le rappel (*recall*) $\mathcal{R}$ et la précision (*precision*) $\mathcal{P}$ [van Rijsbergen, 1979] :

$$\mathcal{R} \stackrel{\text{def}}{=} \frac{\#TP}{\#TP + \#FN} \tag{IX.6}$$

$$\mathcal{P} \stackrel{\text{def}}{=} \frac{\#TP}{\#TP + \#FP} \tag{IX.7}$$

Le rappel reflète la proportion de notes correctes parmi les notes originales, alors que la précision reflète la proportion de notes correctes parmi les notes transcrites. Les deux critères peuvent être synthétisés en un seul pour obtenir une note globale, par exemple via la F-mesure $\mathcal{F}$ [van Rijsbergen, 1979] définie par :

$$\mathcal{F} \stackrel{\text{def}}{=} \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R} + \mathcal{P}} \tag{IX.8}$$

De manière relativement équivalente, on peut également définir une note globale $\mathcal{A}$, appelée score [Dixon, 2000] ou *accuracy* [Poliner et Ellis, 2007] par :

$$\mathcal{A} \stackrel{\text{def}}{=} \frac{\#TP}{\#TP + \#FN + \#FP} \tag{IX.9}$$

Remarquons que ces scores, utilisés notamment par [Bello *et al.*, 2006, Ryynänen et Klapuri, 2005, Bertin *et al.*, 2007, Vincent *et al.*, 2008] se placent au niveau de la note. Alternativement, on peut définir des mesures similaires à l'échelle de la trame, en considérant l'absence ou la présence de fréquences fondamentales à chaque trame. Ces métriques sont adoptées par exemple dans [Plumbley *et al.*, 2006, Poliner et Ellis, 2007]. Ce cas semble davantage conçu pour l'évaluation des algorithmes d'estimation de fréquences fondamentales que pour les algorithmes de transcription, mais peut apporter des informations complémentaires (et un score plus flatteur).

[Raphael, 2002, Poliner et Ellis, 2007, Kameoka, 2007] propose des métriques similaires mais un peu plus détaillées, en faisant la distinction entre les notes manquantes, les notes substituées (notes de début et de fin similaire mais présentant une erreur de hauteur), et les fausses alarmes, définies comme les notes incorrectes ne pouvant pas être considérées comme des notes substituées. Ces métriques sont à rapprocher des distances d'édition [Mongeau et Sankoff, 1990]. Dans ce dernier cas, la définition d'une note correctement estimée dépend d'un seuil de tolérance sur la fréquence fondamentale (le demi-ton en général, *cf.* correspondance entre fréquence fondamentale et notes dans l'annexe B page 187), sur l'instant d'attaque et éventuellement sur l'instant d'extinction de la note.

[Ryynänen et Klapuri, 2005] propose également un critère d'évaluation de la durée transcrite. Pour chaque note correctement transcrite, on définit le taux de recouvrement (*overlap ratio*) o_{note} entre la note originale et la note transcrite comme étant le rapport entre la longueur de l'intersection des supports temporels des deux notes et celle de leur union :

$$o_{note} = \frac{\min(t_{off}) - \max(t_{on})}{\max(t_{off}) - \min(t_{on})} \tag{IX.10}$$

où t_{on} et t_{off} sont les couples de temps d'attaque (resp. d'extinction) de la note originale et de la note transcrite correspondante. Le taux de recouvrement moyen (**Mean Overlap Ratio** ou $\mathcal{MOR}$) est la moyenne des taux de recouvrement de toutes les notes correctement transcrites. L'annotation et l'estimation de t_{off} sont délicates, en particulier en cas d'utilisation de la pédale *forte* ou d'existence d'une forte réverbération. Cela impose de prendre la vérité-terrain et les scores résultants avec précaution, et explique que dans la littérature et les compétitions internationales, le $\mathcal{MOR}$ est pris en compte comme un score complémentaire aux scores principaux.

IX.4 Algorithmes testés

Nous résumons ici l'ensemble des algorithmes utilisés pour la réalisation des expériences de ce chapitre.

IX.4.1 Algorithmes de référence

Nous choisissons trois algorithmes de l'état de l'art à confronter à nos approches. Ils résumés dans la table IX.1, Marolt'04 est un système fondé sur un ensemble de réseaux de neurones, mis au point avec une base d'apprentissage et spécifiquement conçu pour le piano. C'est le plus informé de nos systèmes de référence. Virtanen'07 est un système basé sur la NMF du spectrogramme d'amplitude, par minimisation de la divergence KL sous une contrainte de régularité temporelle imposée par un terme de pénalité. Il s'agit donc d'une approche aveugle du problème. Enfin, Emiya'08 est un système incluant une pré-segmentation, l'estimation de fréquences fondamentales multiples trame par trame, puis la fusion des informations de hauteurs grâce à un modèle HMM dont les transitions sont apprises sur des fichiers MIDI.

Désignation	Description	Référence	Informations
Marolt'04	Réseaux de neurones	[Marolt, 2004]	Apprentissage des réseaux sur une base de données de notes isolées
Virtanen'07	NMF multiplicative avec contrainte de régularité temporelle imposée *via* un terme de pénalité	[Virtanen, 2007]	Aucune
Emiya'08	Pré-segmentation, estimation jointe des fréquences fondamentales multiples et post-traitement par HMM	[Emiya, 2008]	Apprentissage des transitions du HMM sur une base de données de fichiers MIDI

TABLE IX.1 – Algorithmes de référence.

IX.4.2 Algorithmes originaux

La table IX.2 résume l'ensemble des algorithmes basés sur la NMF considérés dans nos travaux, et rappelle brièvement leur description, la section du document où ils sont discutés, et nos publications s'y rapportant.

Désignation	Description	Section	Publication
IS-NMF/MU	NMF minimisant la divergence IS par des règles multiplicatives, sans contrainte	III.4	[Févotte *et al.*, 2009]
(2→ 0)-NMF/MU	NMF avec approche tempérée entre $\beta = 2$ et $\beta = 0$, sans contrainte	IV.3	[Bertin *et al.*, 2009b]
(10→ 0)-NMF/MU	NMF avec approche tempérée entre $\beta = 10$ et $\beta = 0$, sans contrainte	IV.3	[Bertin *et al.*, 2009b]
HEUC-NMF/MU	NMF avec contrainte harmonique, règles multiplicatives, minimisation de la distance euclidienne pondérée	V.3	[Vincent *et al.*, 2008]
H-NMF/MU	NMF avec contrainte harmonique, règles multiplicatives, minimisation de la divergence IS	V.3	
S-NMF/EM	Algorithme SAGE avec contrainte de régularité temporelle	VII.2	[Févotte *et al.*, 2009]
HS-NMF/MU	NMF avec contraintes d'harmonicité et de régularité temporelle, règles multiplicatives	VII.3.4	[Bertin *et al.*, 2009a]
HS-NMF/EM	Algorithme SAGE avec contraintes d'harmonicité et de régularité temporelle	VII.3.3	[Bertin *et al.*, 2010]
PHS-NMF/MU	NMF multiplicative avec contraintes d'harmonicité et de régularité temporelle et composantes libres	VII.4	

TABLE IX.2 – Algorithmes originaux testés.

IX.4.3 Implantation et paramètres

Les algorithmes de référence sont exécutés dans la version de leurs auteurs respectifs, via le logiciel SONIC [2] [Marolt, 2004] et sous forme de scripts Matlab, aimablement fournis par leurs auteurs.

Les algorithmes ont été implantés par l'auteur à l'occasion de cette thèse [3]. L'ordre K est fixé à 88, sauf pour PHS-NMF pour lequel $K = 100$ (88 atomes harmoniques et 12 atomes libres). HS-NMF/EM et S-NMF/EM sont initialisés par 10 itérations de H-NMF/MU et IS-NMF/MU respectivement. Les seuils de détection A_{dB} sont réglés manuellement, en maximisant la F-mesure moyenne sur chaque jeu de test, algorithme par algorithme (leur valeur est reportée dans les tables X.1 et X.2, page 144). La durée minimale d'une note transcrite est de 50 ms. La largeur de bande des patrons P_{km} et le nombre maximum de bandes sont fixés manuellement (après avoir évalué leur valeur optimale sur une base d'apprentissage séparée). Pour la distance euclidienne pondérée (HEUC-NMF/MU), on prend $\Delta f = 3$ et $M_{max} = 6$; pour la divergence IS, $\Delta f = 3$ et $M_{max} = 10$. Le banc de filtres ERB comprend 257 bandes et les trames d'analyse sont séparées de 23 ms.

2. `http://lgm.fri.uni-lj.si/SONIC/`

3. Certains éléments de code pour H-NMF/MU et S-NMF/EM sont issus d'implantations fournies par Emmanuel Vincent et Cédric Févotte.

Chapitre X

Résultats expérimentaux

Résumé

Où l'on teste les algorithmes mis au point dans une tâche réelle de transcription musicale polyphonique, en évaluant leurs performances quantitativement et qualitativement, et en comparaison avec d'autres systèmes de l'état de l'art.

X.1 Introduction

Les propriétés de la NMF et des algorithmes inspirés de celle-ci étant établies et testées sur des données synthétiques, il reste donc à évaluer leurs performances en transcription musicale, en conditions réelles d'utilisation. Ce chapitre présente les résultats obtenus par les différents systèmes testés, en termes de performance quantifiée globale (section X.2), ainsi qu'en termes qualitatifs (section X.3). Enfin, à la section X.4, nous complétons ces résultats à grande échelle sur quelques évaluations de transcription multi-timbres, faisant intervenir d'autres instruments que le piano.

X.2 Résultats globaux

Les tables X.1 et X.2 présentent l'ensemble des scores de transcription moyen pour les sous-bases obtenues par synthèse logicielle et par enregistrement audio, respectivement.

Les intervalles de confiance à 95% pour les F-mesures de chaque algorithme pris séparément sont compris entre ±4 et ±7% autour des valeurs moyennes. Toutefois, la comparaison des F-mesures moyennes deux à deux [Welch, 1938] permet de vérifier que les intervalles de confiance sur la *différence* des F-mesures moyennes sont nettement asymétriques [1], ce qui permet les comparaisons.

Algorithme	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$	A_{dB}
Marolt'04	83.5	70.1	**75.8**	53.5	-
Virtanen'07	55.9	56.4	53.6	52.1	-22
Emiya'08	77.3	61.6	**67.7**	67.0	-
IS-NMF/MU	63.4	56.1	54.9	51.2	-62
$(2 \to 0)$-NMF	59.6	60.6	55.3	53.9	-57
$(10 \to 0)$-NMF	62.3	51.3	51.4	52.7	-59
H-NMF/MU	58.7	59.1	52.4	46.0	-33
HEUC-NMF/MU	60.7	60.0	58.4	54.8	-32
S-NMF/EM	62.4	43.3	49.5	50.7	-51
HS-NMF/EM	65.8	64.5	60.7	44.3	-38
HS-NMF/MU	78.5	62.6	67.0	46.4	-42
PHS-NMF/MU	77.6	65.4	**68.4**	45.0	-43

Table X.1 – Performance moyenne de transcription sur la base AkPnBcht.

Les algorithmes de NMF intégrant à la fois les contraintes d'harmonicité et de régularité temporelle (HS-NMF/MU, HS-NMF/EM et PHS-NMF/MU) fournissent les meilleures performances parmi tous les algorithmes basés sur la NMF. Leurs résultats sont comparables à celui de Emiya'08, ce qui place notre système au niveau d'un algorithme supervisé de l'état de l'art ; ils restent cependant inférieurs à ceux de Marolt'04, un système très perfectionné utilisant des bases d'apprentissage extérieures et spécifiquement réglé pour l'analyse de sons de piano, ce qui explique son excellente performance.

Les algorithmes tempérés $(2 \to 0)$-NMF et $(10 \to 0)$-NMF produisent des résultats comparables à l'algorithme standard non contraint et non tempéré IS-NMF ; on obtient un résultat légèrement meilleur lorsque $\beta_i = 2$ (zone de convexité, *cf.* section IV.3 page IV.3) et légèrement moins bon lorsque $\beta_i = 10$.

1. Quelques exemples : H-NMF/MU vs. HS-NMF/EM $[-0.2\%, 16\%]$, Marolt'04 vs. IS-NMF/MU $[19\%, 26\%]$.

Algorithme	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$	A_{dB}
Marolt'04	63.7	53.6	**58.0**	50.0	-
Virtanen'07	34.2	34.8	33.6	47.1	-21
Emiya'08	66.1	45.5	**52.9**	55.9	-
IS-NMF/MU	43.3	43.4	40.8	47.7	-60
$(2 \rightarrow 0)$-NMF	44.7	45.3	42.6	48.1	-56
$(10 \rightarrow 0)$-NMF	43.0	38.8	37.9	48.8	-60
H-NMF/MU	43.0	42.7	41.3	44.6	-30
HEUC-NMF/MU	38.7	37.4	36.1	50.0	-30
S-NMF/EM	46.2	32.0	36.6	45.6	-49
HS-NMF/EM	46.6	45.3	45.0	43.2	-32
HS-NMF/MU	50.6	42.7	**45.0**	44.5	-35
PHS-NMF/MU	52.4	43.9	44.7	44.4	-45

TABLE X.2 – Performance moyenne de transcription sur la base ENSTDkAm.

Sur l'ensemble des 60 morceaux de la base, nous n'avons observé aucun problème de convergence de ces algorithmes, malgré la preuve qu'ils ne font pas décroître de manière monotone la divergence IS.

La contrainte de régularité temporelle utilisée seule semble nuire à la qualité de la transcription, qu'elle soit implantée par un algorithme multiplicatif (Virtanen'07) ou dans un cadre bayésien (S-NMF). Par ailleurs, la contrainte d'harmonicité améliore les résultats, mais pas autant que lorsque les deux contraintes sont associées.

Les versions multiplicatives des algorithmes contraints (HS-NMF/MU et PHS-NMF/MU) produisent de meilleurs résultats que la version EM, tout en restant comparables. Nous n'observons pas de problèmes de convergence pour HS-NMF/MU. En revanche, il faut signaler que les résultats de PHS-NMF/MU sont obtenus en abaissant le pas η à 0.4, faute de quoi certaines exécutions conduisent à une divergence du critère vers $+\infty$.

Aucune tendance claire ne se dégage quant à la mesure du taux de recouvrement, bien que l'on ait pu s'attendre à ce que la contrainte de régularité temporelle améliore la détection des *offsets* et conduise à un meilleur $\mathcal{MOR}$. Emiya'08 atteint la meilleure valeur de cette mesure, ce qui est probablement à imputer au traitement temporel sophistiqué (pré-segmentation et post-traitement par HMM).

Nous pouvons enfin noter que les meilleures F-mesures sont obtenues en général en favorisant la précision sur le rappel. Ceci est par ailleurs perceptivement souhaitable, les notes ajoutées (« fausses notes ») étant perçues comme beaucoup plus gênantes que les notes oubliées par un auditeur humain [Daniel *et al.*, 2008].

X.3 Observations de détail

X.3.1 Convergence et minima locaux

Nous comparons l'évolution des critères lorsque HS-NMF/EM est initialisé au hasard, ou lorsqu'il est précédé de quelques itérations de H-NMF/MU, sur un exemple tiré de la base de test.

Bien que C^{MAP} chute brutalement pendant la phase d'initialisation (itérations 1 à 10), l'amorçage

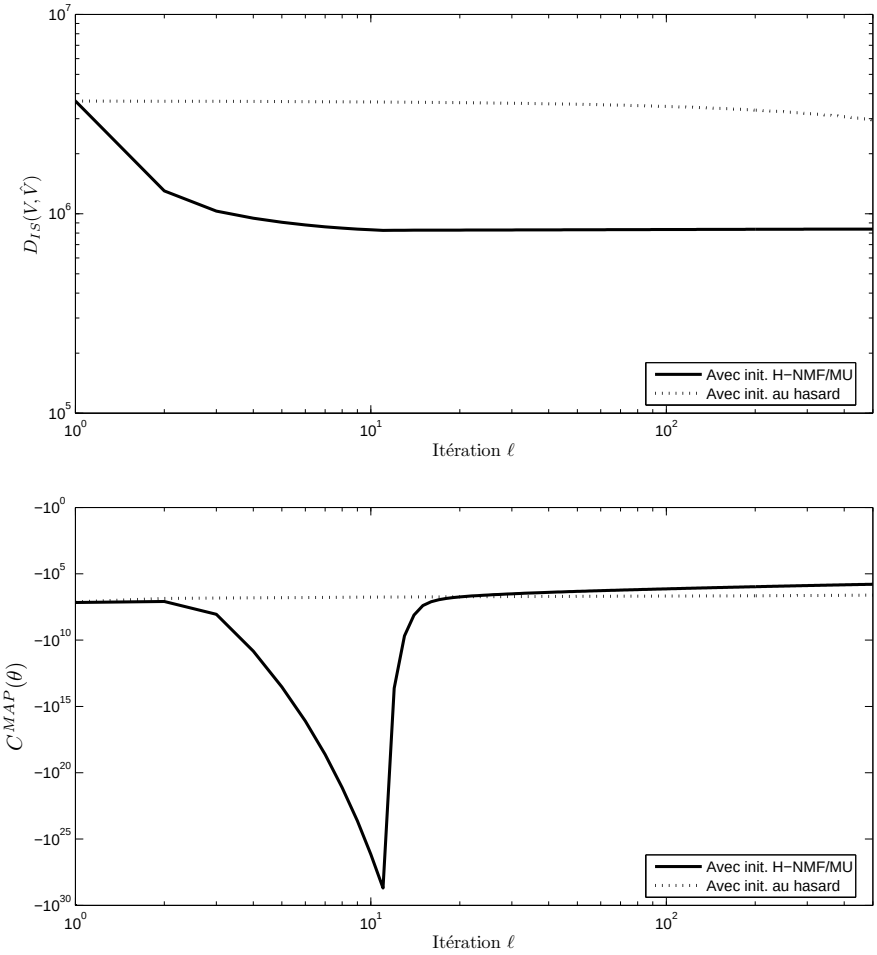


FIGURE X.1 – Évolution des critères D_{IS} (en haut) C^{MAP} (en bas) en fonction du nombre d'itérations.

multiplicatif permet à l'algorithme d'atteindre une valeur plus élevée du critère pour le même nombre total d'itérations, ainsi qu'une valeur plus faible de l'erreur de reconstruction D_{IS} (qui, rappelons-le, est égale à l'opposé de la log-vraisemblance à une constante près). Après quelques centaines d'itérations, l'erreur de reconstruction évolue très peu, tandis que la contribution liée à l'a priori temporel augmente lentement, ce qui ne change que très marginalement la performance de transcription. Plus important, sur l'extrait analysé ici (une pièce de 30 secondes issue de ENSTDkAm), HS-NMF/EM avec amorçage multiplicatif atteint une performance de transcription relativement bonne ($\mathcal{F}$=54.5%), tandis que son homologue sans cette initialisation s'avère incapable de séparer les notes dans le même temps ($\mathcal{F}$=0% après 500 itérations). Une explication de ce phénomène, déjà évoqué dans les simulations du chapitre VIII, est l'importance relative prise par chacun des deux termes constituant le critère C^{MAP} : le premier but à atteindre par NMF est avant tout une bonne reconstruction du signal, la régularité temporelle étant un « bonus ». Si la contribution de l'a priori temporel prend le pas sur l'erreur de reconstruction, celle-ci sera médiocre. L'initialisation multiplicative non contrainte temporellement permet d'optimiser en premier lieu le terme lié à l'erreur de reconstruction, puis de se focaliser sur le raffinement que constitue la contrainte de régularité.

X.3.2 Harmonicité et composition du dictionnaire

Sur la figure X.2, nous affichons les bases $\mathbf{W}$ après convergence, les colonnes étant classées par pitch croissant. La NMF sans contrainte harmonique produit un dictionnaire qui possède une structure de type harmonique, mais bruitée, tandis que les bases produites par les algorithmes incluant la contrainte d'harmonicité sont plus propres. S-NMF aboutit à un dictionnaire nettement moins parcimonieux que la NMF non contrainte, ce qui est cohérent avec l'observation de [Pascual-Montano *et al.*, 2006] discutée à la section V.2.2 (page 82). Ceci pourrait expliquer sa performance médiocre.

Un autre résultat remarquable est la répartition des hauteurs dans le dictionnaire. La figure X.3 présente le nombre d'occurrences des notes en fonction de leur pitch MIDI, à la fois pour la pièce analysée, et pour les bases résultant de la NMF. Dans le cas de la NMF non contrainte, la base $\mathbf{W}$ est totalement libre, ce qui permet à la répartition des hauteurs dans le dictionnaire de suivre la même tendance que la répartition des notes dans le morceau original. On peut notamment remarquer des pics de l'histogramme aux pitch MIDI 50, 62 et 74, qui correspondent au cinquième degré (⋆dominante) de la ⋆tonalité du morceau analysé (une suite d'Albeniz en sol majeur, tonalité dont le cinquième degré est le ré). En théorie classique, ce degré a la réputation de donner un fort sentiment de tonalité ; c'est de plus un pivot usuel pour la ⋆modulation au ton voisin. Par conséquent, il est plus fréquemment joué que les autres notes de la gamme. La NMF standard a donc tendance à utiliser davantage de composantes pour représenter le plus fidèlement possible les notes les plus fréquentes (ce qui a un fort impact sur le coût), en négligeant éventuellement des notes plus rares comme les notes étrangères à la tonalité (comme elles sont peu jouées, ne pas les représenter est moins coûteux que mal représenter des notes fréquentes).

Remarquons également que certaines composantes des bases renvoyées par les algorithmes de NMF non contraints ne possèdent pas de structure harmonique, donc de pitch. Sur l'ensemble de la base, ces composantes de bruit sont en moyenne au nombre de 5 par base.

À l'inverse, les NMF incluant la contrainte d'harmonicité ont un nombre fixe de composantes par pitch, que celui-ci soit rare, fréquent ou même absent dans la pièce. Ceci garantit la représentation de toutes les notes, y compris celles qui ne sont jouées que rarement. Cela implique cependant un surcroît de calcul inutile pour les composantes correspondant à des notes absentes du morceau, et ne permet pas à la représentation de s'adapter à la tonalité en représentant plus finement les notes fréquentes.

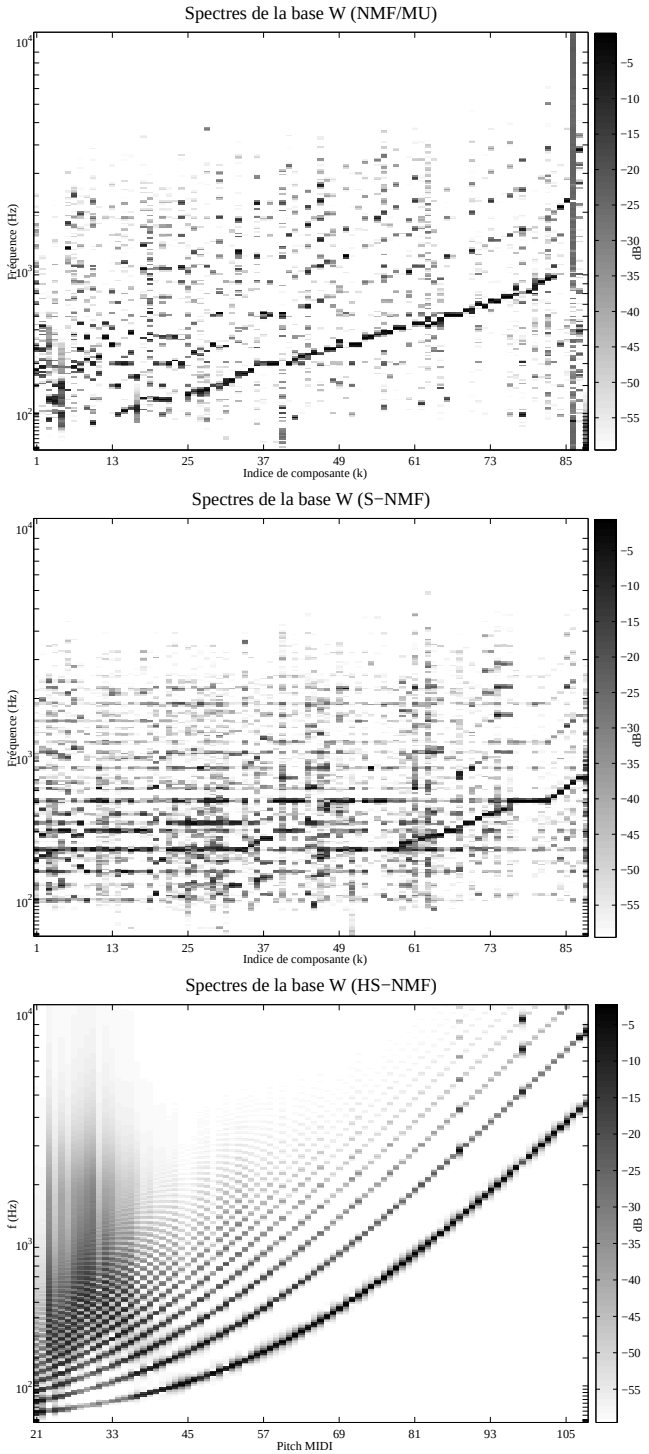


FIGURE X.2 – Exemples de bases **W** pour des algorithmes sans ou avec contrainte harmonique. Les colonnes sont triées par pitch associé croissant.

On pourrait envisager une exécution de l'algorithme en deux passes, la première servant à repérer les composantes inutiles, la seconde à calculer une nouvelle factorisation avec un ordre optimal.

X.3.3 Influence de l'ordre du modèle

Puisque nous prévoyons de séparer des notes de musique, le choix intuitif et commun dans le cas de la transcription de piano est $K = 88$, c'est-à-dire le nombre de touches sur un piano. Cependant, dans un même morceau de musique, il est rare d'entendre chacune de ces 88 notes, d'une part car les parties extrêmes de la tessiture sont peu utilisées par les compositeurs, d'autre part à cause des contraintes de tonalité (le $fa\sharp$ apparaîtra peu dans un morceau en *do* Majeur, par exemple). Le choix $K = 88$ peut ainsi être considéré comme une surestimation de ce qui pourrait être l'ordre « optimal », c'est-à-dire du nombre de hauteurs différentes effectivement présentes dans la pièce.

Dans la NMF non contrainte, le choix d'un ordre K ne garantit en aucune manière que le dictionnaire contiendra K hauteurs différentes, comme nous l'avons observé sur l'histogramme de la figure X.3. De plus, certaines composantes $\mathbf{w}_k$ ne possèdent pas de hauteur (5 en moyenne pour IS-NMF/MU dans notre test). Puisque le dictionnaire s'adapte aux besoins de représentation et à la répartition des hauteurs dans le morceau analysé, en particulier en représentant finement les notes fréquentes, et comme le suggèrent les observations faites à la section IX.2.2 (page 130), la surestimation apparaît comme un bon choix.

À l'inverse, la contrainte d'harmonicité implique une plus grande rigidité du dictionnaire et le choix de l'ordre se pose en termes différents. [Vincent *et al.*, 2007] remarque que le doublement de la taille du dictionnaire (deux atomes par hauteur) n'influe pas sur le résultat de transcription de piano sur la base de test utilisée. Cependant, on pourrait imaginer qu'un tel choix (plusieurs composantes par pitch) serait utile si l'on cherchait à effectuer la transcription de morceaux multi-instruments.

Le choix de l'ordre du modèle est un problème peu évoqué dans la littérature. À notre connaissance, des techniques de choix de l'ordre du modèle n'ont été étudiés comme problèmes en soi que dans [Winther et Petersen, 2007,Cemgil, 2008,Tan et Févotte, 2009]. Ils proposent des méthodes bayésiennes variationnelles d'estimation de la dimensionalité optimale, mais seule la technique de [Tan et Févotte, 2009] est suffisamment peu coûteuse pour être implantable en pratique. Une autre approche, proposée dans [Marolt, 2009], consiste à permettre à l'ordre du modèle d'évoluer au cours de l'exécution de l'algorithme ; l'ordre optimal est ainsi obtenu a posteriori, conjointement avec la factorisation.

X.3.4 Régularité et dynamique des enveloppes

Les enveloppes temporelles h_k pour k associé à la note do_4, obtenues par IS-NMF/MU (non contrainte), H-NMF, S-NMF et HS-NMF sont représentées sur la figure X.4. Le pianoroll correspondant à la vérité-terrain (référence MIDI du morceau analysé) est également présenté. S-NMF et HS-NMF produisent effectivement des enveloppes plus lisses que les deux autres algorithmes, ce que l'on peut notamment remarquer aux instants où cette note est supposée être inactive. Dans les enveloppes produites par IS-NMF/MU et H-NMF/MU, on observe plusieurs pics superflus, par exemple pendant les 750 premières millisecondes du morceau, ou encore autour de t=10 s. On peut attribuer ces pics à l'activation simultanée de notes voisines (si_3 ou $do_4\sharp$, dont le spectre peut « contaminer » les notes adjacentes, en raison de la résolution fréquentielle limitée) ou de notes à la sous-octave (do_3) susceptibles d'activer la composante associée au do_4 qui contient toutes ses fréquences ; le spectre large-bande des transitoires de n'importe quelle autre note est également susceptible d'introduire du bruit

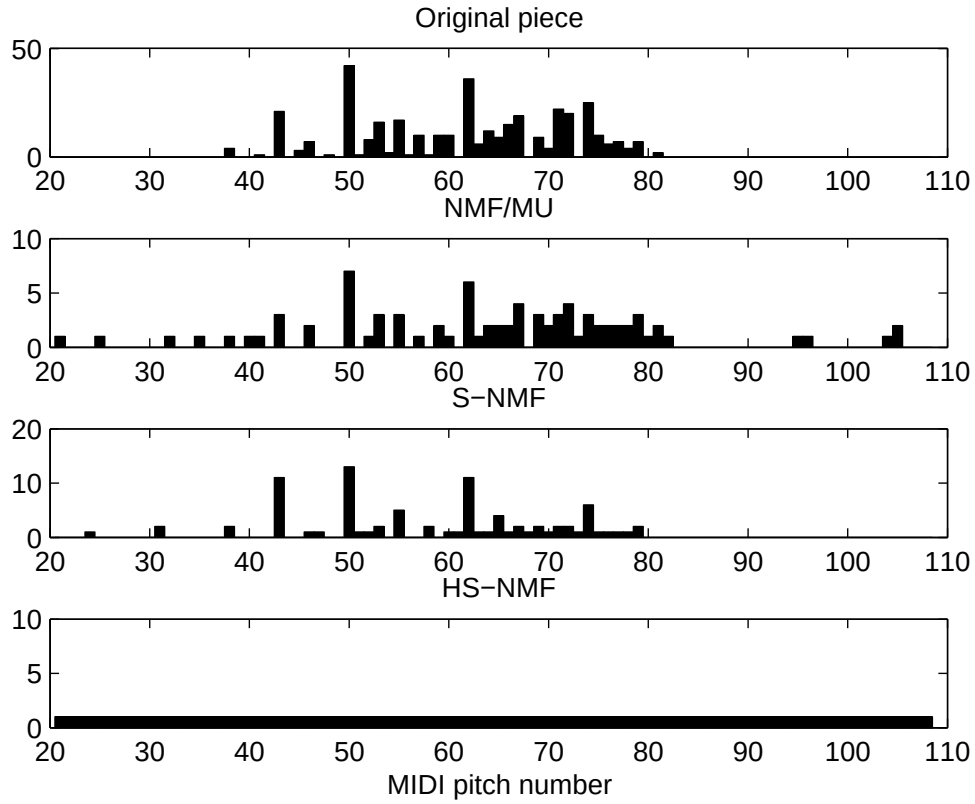


FIGURE X.3 – Histogramme du nombre d'occurrence des notes en fonction de leur pitch MIDI, dans la pièce originale et les bases calculées par NMF, pour trois différents algorithmes. Les composantes de **W** estimées comme non pitchées ne sont pas représentées.

dans une composante. Ces pics sont susceptibles de générer des insertions erronnées de notes dans la transcription.

L'amplitude de ces faux pics est réduite ou annulée par la contrainte de régularité. Par ailleurs, les pics de faible amplitude correspondant à des transitoires sont légèrement atténués par l'ajout de composantes libres, capables de les capturer donc d'éviter qu'ils ne polluent les activations des parties harmoniques du signal. Nous pouvons remarquer que la contrainte d'harmonicité semble défavoriser la régularité des enveloppes, *ceteris paribus*. Enfin, la dynamique de l'enveloppe est réduite par la contrainte de régularité, et ce d'autant plus que nous choisissons un α_k grand, ce qui est logique puisque la contrainte de régularité défavorise les forts changements d'amplitude, donc les attaques. Ceci pourrait conduire à une baisse de précision dans la détection des attaques, et en particulier à une plus grande sensibilité au choix du seuil de détection.

En ce qui concerne un éventuel effet sur l'estimation de la durée des notes (détection de l'extinction), nous aurions pu attendre un impact positif de la contrainte de régularité. On ne constate cependant pas, sur les tables X.1 et X.2, d'augmentation significative du $\mathcal{MOR}$, ce qui est décevant.

X.3.5 Robustesse du seuil de détection

Les scores des tables X.1 et X.2 sont obtenus en optimisant manuellement le seuil de détection A_{dB}, de manière à maximiser la F-mesure moyenne sur la sous-base analysée. Ceci nous donne bien sûr des résultats optimistes, permettant davantage d'évaluer le potentiel des méthodes que leur performance en conditions réalistes. L'apprentissage de ce seuil sur une base extérieure forcerait à renoncer à la non supervision des approches. Cependant, on peut espérer atteindre ces performances grâce à un post-traitement plus robuste dans le domaine temporel.

Si l'on fait varier le seuil de détection, nous pouvons afficher les courbes Précision-Rappel associées à chaque algorithme et ainsi obtenir une information supplémentaire sur leurs performances. La figure X.5 présente ces courbes pour IS-NMF/MU, H-NMF/MU, S-NMF/EM et HS-NMF/EM, le seuil variant de 0 à -100 dB. La courbe confirme la bonne performance de HS-NMF, qui atteint un meilleur compromis précision-rappel et semble plus robuste au choix du seuil. Les algorithmes multiplicatifs (H-NMF/MU et IS-NMF/MU) sont comparables autour de la F-mesure maximale (repérée par une étoile) mais H-NMF/MU est plus robuste au choix du seuil. S-NMF/EM produit les scores les plus faibles à tous les seuils.

Deux phénomènes relativement inhabituels pour de telles courbes sont à relever. D'une part, à très faible seuil, le rappel décroît, alors qu'on pourrait au contraire s'attendre à ce qu'il augmente puisque l'abaissement du seuil rend la détection d'une note plus probable. Ceci est vraisemblablement lié à la mauvaise détection de notes répétées ou s'enchaînant rapidement, l'abaissement du seuil tendant à les fusionner c'est-à-dire à ne détecter que l'attaque de la première note et l'extinction de la dernière. Ainsi, un rappel de 100% n'est jamais atteint. Á l'inverse, lorsque le seuil croît, des petites variations d'amplitude d'une même note peuvent être détectées comme des notes répétées, donnant lieu à une baisse de la précision, qu'on observe pour IS-NMF/MU (autour de $\mathcal{P} = 80$). Ceci illustre la faiblesse de cette technique de seuillage.

Les courbes de la figure X.5, tout comme les tables X.1 et X.2, sont obtenues en moyennant les scores de chaque morceau au sein de la sous-base de test. Il faut cependant noter une importante variabilité entre pièces, tant en terme de performance que de seuil optimal. À seuil fixé A_{dB}, l'écart-type de $\mathcal{F}$ est d'environ 12%, tous algorithmes confondus (de 9% pour Virtanen'07 à 16% pour HS-NMF/EM). Là encore, le choix d'un post-traitement par seuillage et la détermination de la valeur optimale de ce

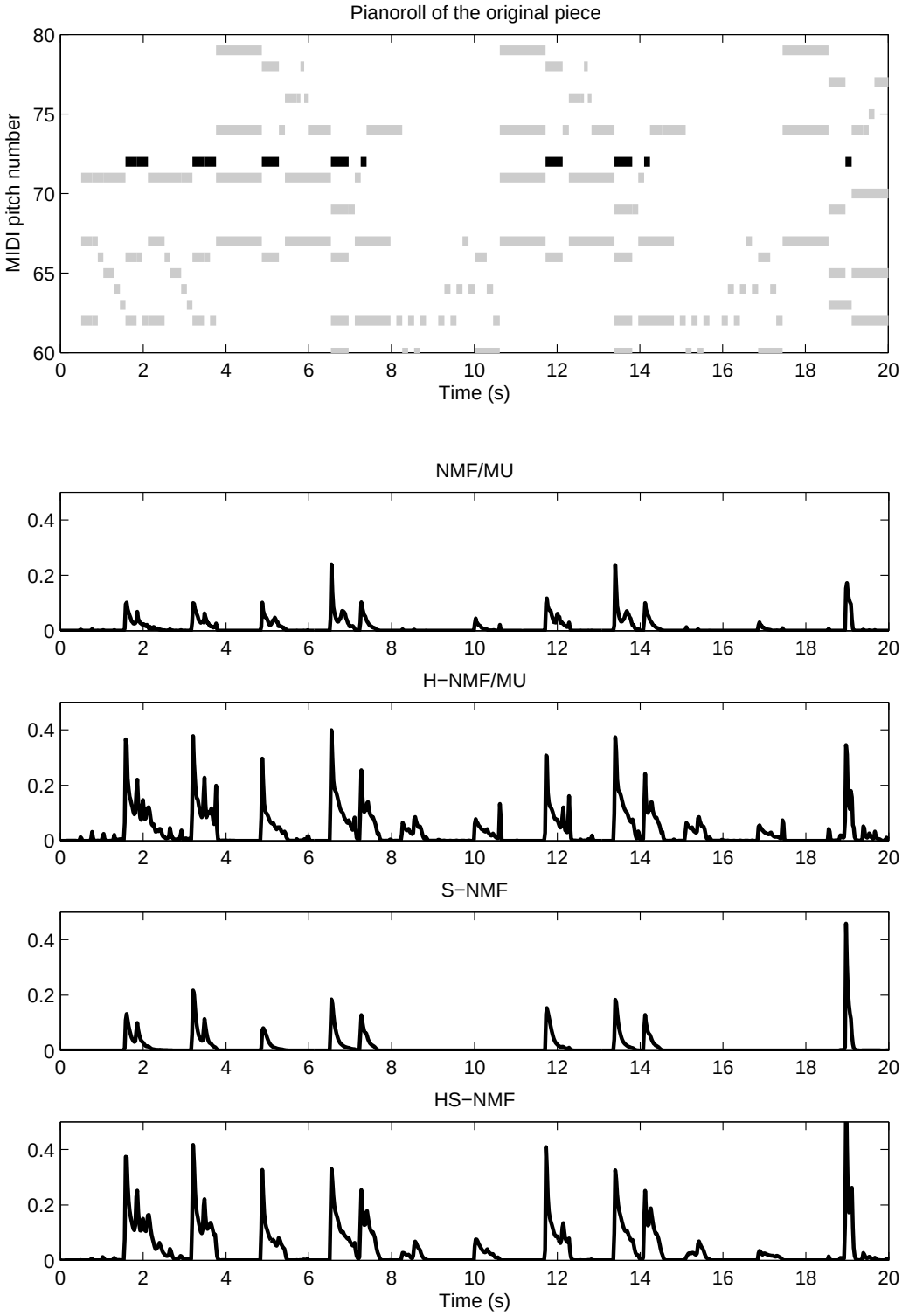


FIGURE X.4 – Activations temporelles de la note do_4 pour quatre algorithmes analysant le même extrait. Le pianoroll de l'extrait analysé est présenté sur la première figure, les occurrences du do_4 étant signalées en noir et les notes voisines en gris.

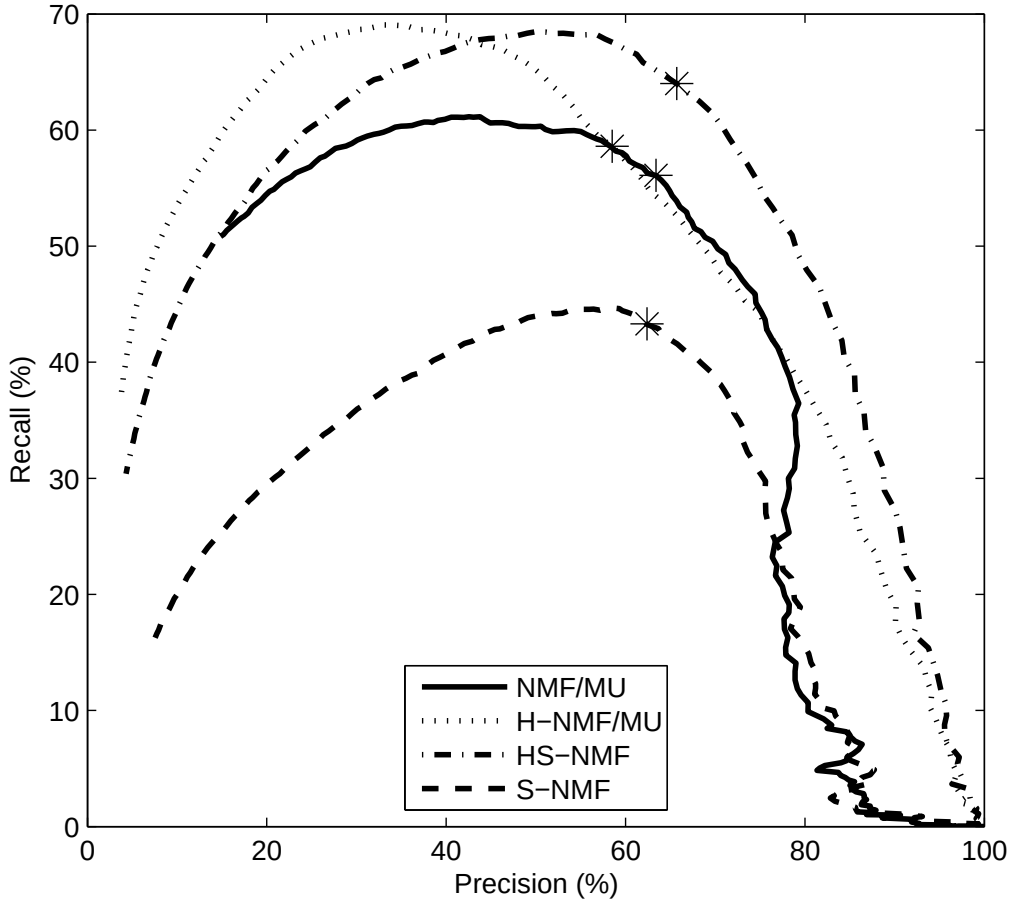


FIGURE X.5 – Courbe Précision-Rappel pour quatre algorithmes. Le seuil de détection varie de 0 à -100 dB par rapport au maximum de $\mathbf{H}$. Le couple $(\mathcal{P},\mathcal{R})$ réalisant le maximum de la $\mathcal{F}$-mesure est signalé par une étoile.

seuil sont à mettre en question.

X.3.6 Erreurs d'octave et précision de la détection d'attaque

Les scores globaux ne nous donnent pas d'information sur le type des erreurs faites, ce qui peut avoir son importance perceptive [Daniel *et al.*, 2008]. Deux erreurs typiques peuvent nous apporter une information supplémentaire sur les performances des algorithmes : les erreurs d'octave, très courantes en transcription, et les erreurs sur la détection de l'attaque. Pour évaluer ces erreurs, nous calculons la variation de précision par rapport aux évaluations précédentes, lorsqu'on considère comme correcte une note à l'octave ou la sous-octave de la vérité-terrain (ΔP_{octave}) ou une note dont l'attaque est à moins de 150 ms de la vérité-terrain (ΔP_{temps}). Nous résumons les résultats obtenus dans la table X.3.

Algorithme	ΔP_{octave}	ΔP_{temps}
Marolt'04	2 %	2.3 %
Virtanen'07	6 %	4.2 %
Emiya'08	3.1 %	3.9 %
IS-NMF/MU	4.9 %	2.8 %
$(2 \rightarrow 0)$-NMF	5.1%	3.1 %
$(10 \rightarrow 0)$-NMF	5.2 %	2.9 %
H-NMF/MU	4 %	2 %
HEUC-NMF/MU	3.8 %	1.5 %
S-NMF/EM	8.4 %	4.6 %
HS-NMF/EM	2.7 %	1.3 %
HS-NMF/MU	3.3 %	1.4 %
PHS-NMF/MU	2.9 %	1.7 %

TABLE X.3 – Pourcentage d'erreurs d'octaves et d'erreurs sur la détection de l'attaque.

L'usage simultané des contraintes d'harmonicité et de régularité temporelle permet d'amener le nombre d'erreurs d'octave au niveau de l'algorithme Emiya'08, dont la modélisation des fréquences fondamentales multiples est spécialement conçue pour les éviter. La contrainte de régularité temporelle, contrairement à ce qu'on pourrait craindre, n'a pas d'effet particulièrement néfaste sur la précision de la détection des attaques (le pourcentage de notes dont l'attaque est détectée entre +50 ms et +150 ms de la vérité-terrain n'est pas clairement plus élevé pour les algorithmes incluant cette contrainte).

X.3.7 Composantes libres

Nous visualisons sur la figure X.6 les éléments de la base correspondant aux composantes non contraintes, et les activations temporelles correspondantes, produites par PHS-NMF sur un des morceaux de la base AkPnBcht.

Ces composantes ne possèdent pas de hauteur et leurs activations sont très parcimonieuses, évoquant la représentation de phénomènes transitoires. En pratique, leurs amplitudes sont extrêmement faibles, comparées aux lignes associées à des composantes harmoniques, ce qui confirme notre intuition de la divergence IS. Pour cette raison, l'effet espéré de débruitage des activations temporelles des composantes harmoniques n'est que peu observé.

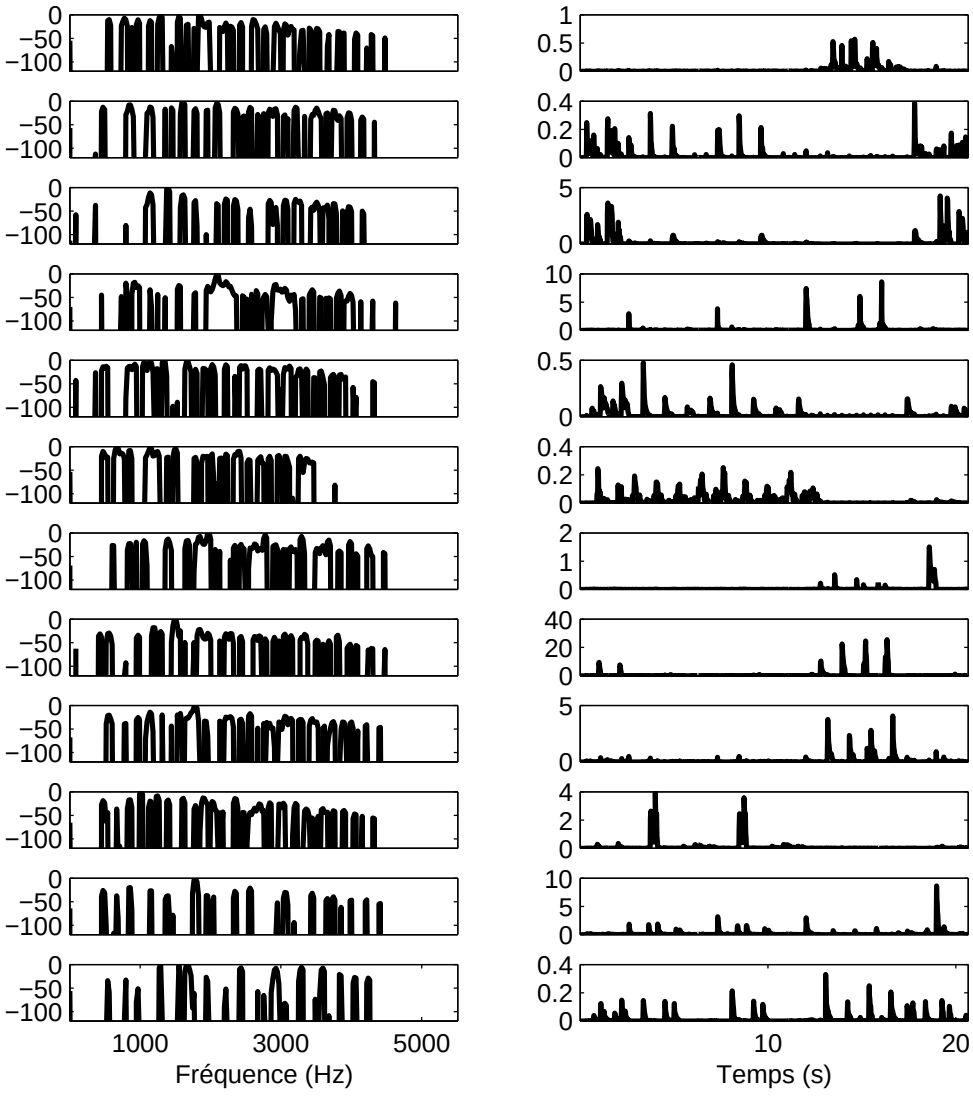


FIGURE X.6 – Partie de la base extraite par PHS-NMF correspondant aux composantes non contraintes, et activations associées.

X.4 Résultats complémentaires : données multi-instruments

Nous n'avons considéré jusqu'ici que la transcription de sons de piano, pour des raisons pratiques (disponibilité des bases de données annotées). Cependant, rien dans le modèle n'utilise les particularités du piano et les algorithmes proposés sont tout-à-fait susceptibles d'être opérants sur des sons issus d'autres instruments. Du reste, on trouve dans la littérature l'utilisation de la NMF pour transcrire des sons de batterie [Paulus et Virtanen, 2005], de flûte et de guitare [Wang et Plumbley, 2005], des mélanges instantanés issus de divers instruments [Virtanen, 2007], ou encore des cloches [Marolt, 2009].

On peut cependant s'attendre à des différences. D'une part, les sons de piano ne sont pas parfaitement harmoniques : il existe une déviation progressive, en allant vers l'aigu, des fréquences des partiels par rapport aux multiples exacts de la fréquence fondamentale. En revanche, les instruments à sons entretenus tels que les bois ou les cordes sont parfaitement harmoniques. Ceci pourrait améliorer les performances des NMF avec contraintes harmoniques. Inversement, les sons de piano ont des fréquences qui restent stables au cours de la note, alors que d'autres instruments sont susceptibles de produire des *vibrato qui ne peuvent pas, *a priori*, être représentés par l'approximation de NMF. Nous proposons ici une évaluation préliminaire à l'usage de la NMF pour la transcription d'instruments autres que le piano.

X.4.1 Cas particulier

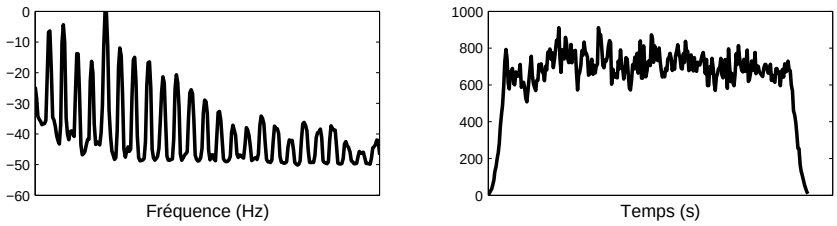
Pour se faire une première idée du type de phénomènes que nous pouvons rencontrer en analysant d'autres sons que le piano, on analyse une note vibrée de violon, par IS-NMF/MU avec $K = 1, 2, 4$.

Pour $K = 1$, la NMF parvient à représenter la note sous la forme d'un spectre moyenné, aux pics relativement élargis. L'enveloppe temporelle est très bruitée et peine à suivre la variation d'amplitude observée sur le spectrogramme de la figure I.3 (page 17). À $K = 2$, les deux composantes représentent les fréquences extrêmes du vibrato, et les enveloppes temporelles oscillent en opposition de phase. Suivant la profondeur du vibrato, cette représentation sera transcrite soit sous forme de trille (si les fréquences des composantes sont éloignées d'un demi-ton ou plus), soit sous la forme d'une seule note de fréquence constante ; dans les deux cas, la sémantique du vibrato est perdue. À $K = 4$, la représentation se raffine encore, avec en particulier l'apparition d'une composante ne représentant que les premiers partiels de la note, et s'activant surtout au cours de sa deuxième moitié. Ainsi, l'évolution non-linéaire du spectre de la note au cours de sa durée, qui ne se limite pas à la simple multiplication par un gain global d'une trame à l'autre, est capturée d'une manière peut-être efficace en transcription, mais insatisfaisante du point de vue sémantique.

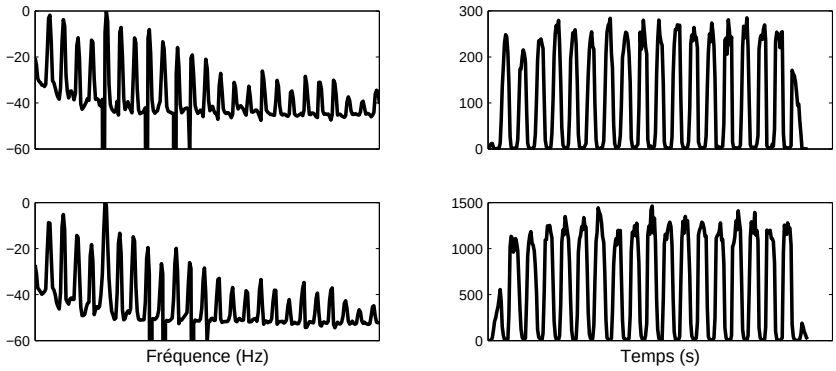
X.4.2 Musique de chambre

Pour compléter notre étude, nous transcrivons avec les algorithmes précédents un morceau issu de la base de développement de la compétition internationale MIREX 2007 [2]. Cette pièce polyphonique de 54 secondes fait intervenir cinq instruments à vent (hautbois, cor anglais, flûte, clarinette et basson) et comportant 1025 notes. Bien que le modèle de NMF ne soit pas, a priori, adapté à des notes dont les fréquences sont susceptibles de varier entre le début et la fin de la note, il est cependant plus générique que Marolt'04 et Emiya'08, spécifiquement conçus pour le piano. La table X.4 présente les résultats de ces transcriptions.

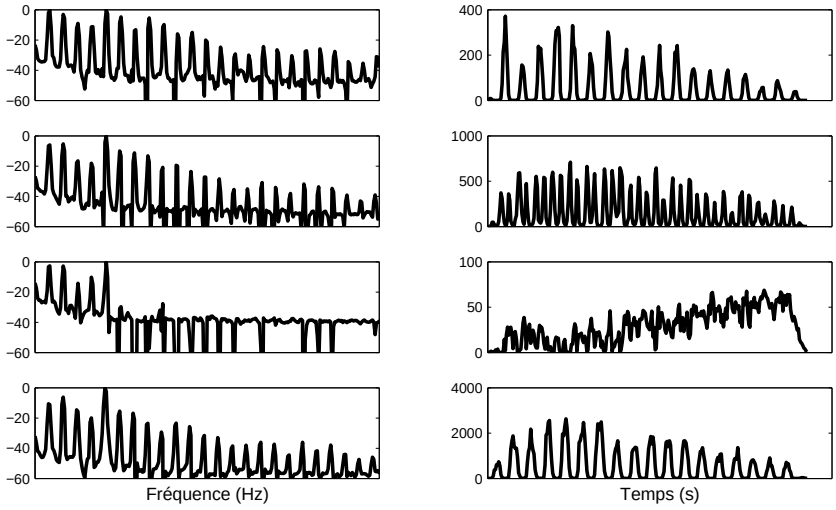
2. La pièce et une annotation des fréquences fondamentales présentes par pas de 10 ms sont fournies par Mert Bay. La référence MIDI a été produite semi-automatiquement et corrigée manuellement pour les besoins de cette thèse.



(a) $K = 1$



(b) $K = 2$



(c) $K = 4$

FIGURE X.7 – Analyse d'une note vibrée de violon par IS-NMF/MU, $K = 1, 2, 4$.

Algorithme	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$
Marolt'04	63.7	44.7	52.5	57.7
Virtanen'07	28.4	23.2	25.6	56.3
Emiya'08	15.5	57.2	24.4	52.0
IS-NMF/MU	40.9	59.9	48.6	62.1
$(2 \to 0)$-NMF	49.4	45.0	47.1	60.7
$(10 \to 0)$-NMF	38.4	59.3	46.6	61.2
H-NMF/MU	49.7	48.3	**49.2**	63.0
HEUC-NMF/MU	48.3	49.8	**49.1**	64.0
S-NMF/EM	48.1	32.6	38.9	49.7
HS-NMF/EM	42.5	36.6	39.3	54.0
HS-NMF/MU	43.5	37.2	40.1	53.1
PHS-NMF/MU	50.7	41.2	45.4	57.3

TABLE X.4 – Performance de transcription sur le jeu de données multi-instruments.

La plupart des algorithmes NMF parviennent à des performances intéressantes, comparables à celles obtenues sur la base ENSTDkAm, et presque équivalentes à celles de Marolt'04, qui, bien que conçu pour le piano, reste le plus efficace. En revanche, Emiya'08, spécifiquement conçu pour le piano, est mis en difficulté. La contrainte de régularité temporelle semble ici nuire aux résultats. On peut imaginer plusieurs explications : le choix du paramètre α_k (les durées de notes et les formes d'ondes des instruments à vent étant très différentes de celles d'un piano), ou l'impossibilité de représenter une note s'éteignant brutalement, qui découle du choix de la loi inverse-Gamma.

Ces résultats mettent en lumière les limites des méthodes fondées sur la NMF pour traiter des données multi-instruments et en particulier lorsqu'elles font intervenir des instruments dont le spectre n'est pas suffisamment stable dans le temps. Plusieurs obstacles apparaissent : d'une part, la sémantique de la représentation mi-niveau produite, car même si la performance de transcription est raisonnable, certains phénomènes musicaux y perdent leur sens ; deuxièmement, la représentation par une seule composante de notes de même hauteur mais produites par différents instruments conduit à un spectre « moyenné » peu fidèle à la réalité du signal, dès que les instruments en jeu possèdent des spectres très différents les uns des autres à une hauteur donnée ; enfin, l'augmentation de la taille du dictionnaire à plusieurs composantes par pitch risque de faire sortir le système du principe de « réduction de rang » à l'œuvre dans la NMF.

Conclusions et perspectives

Arrivée au terme de cette étude, nous pouvons en tirer les conclusions et en particulier, examiner dans quelle mesure nous pouvons désormais répondre aux questions soulevées à l'issue du chapitre II. Nous verrons ensuite vers quelles directions ces travaux pourraient se prolonger, et quelles réponses nous pouvons imaginer aux questions laissées en suspens ou soulevées au cours de l'étude.

Bilan de la thèse

Dans la première partie de ce document, nous avons présenté la définition du problème de transcription musicale en général et défini notre périmètre d'étude, la transcription des hauteurs et des durées des *notes* dans les musiques pour lesquelles cette notion est pertinente. Le passage en revue de l'état de l'art, des approches les plus anciennes aux plus récentes, nous a permis d'énoncer et de positionner nos problématiques tant du point de vue méthodologique (la factorisation en matrices non-négatives en tant que technique et outil) qu'applicatif (difficultés et enjeux de la transcription musicale).

Nous avons consacré la deuxième partie de la thèse à l'étude de la NMF en tant qu'outil, dans son cadre classique, comme problème d'algèbre linéaire déterministe. Nous avons étudié les questions de l'existence et de l'unicité des solutions, recensé et discuté les principales fonctions de coût, les principaux algorithmes à disposition, leurs propriétés de convergence et leur initialisation. L'étude théorique et expérimentale nous a amenée à considérer particulièrement la divergence d'Itakura-Saito, jamais utilisée jusqu'ici en contexte musical. Nous avons par ailleurs porté une attention particulière au comportement des algorithmes vis-à-vis des minima locaux, proposé et examiné différentes stratégies pour les éviter, et conclu de cette étude la nécessité de recourir à des contraintes supplémentaires. En particulier, nous avons proposé une contrainte d'harmonicité de la base de décomposition, que nous avons formalisée puis intégrée aux algorithmes existants.

La troisième partie de la thèse est consacrée au lien entre l'approche NMF et le cadre probabiliste, et à l'exploitation de ce cadre pour une approche théorique et algorithmique nouvelle au problème de NMF contraint. Après un bref panorama établissant le contexte en la matière, nous avons concentré notre étude sur un modèle de « Somme de Gaussiennes », établi des théorèmes d'équivalence entre les modèles déterministe et probabiliste, et dérivé de cette équivalence des algorithmes originaux d'estimation des paramètres de la NMF contrainte. Cette partie inclut une étude expérimentale du comportement et de la consistance de ces algorithmes, préliminaire à l'application en transcription musicale.

Cette application est le cœur de notre quatrième partie, consacrée à la description de l'architecture générale du système original de transcription mis au point pendant la thèse, et à l'évaluation de ses différentes instanciations, mettant en œuvre les algorithmes précédemment présentés. Les résultats expérimentaux, obtenus à partir d'un matériel varié (données de piano réelles et synthétiques, musique de chambre multi-instruments) et comparés à plusieurs systèmes de référence fondés sur des approches

radicalement différentes, plus informées, ont confirmé la pertinence des approches proposées. L'observation minutieuse des résultats a également permis de mettre en lumière les forces et faiblesses de l'approche et d'envisager des pistes d'amélioration.

Réponses aux problématiques

Comment fonctionne la NMF, et pourquoi fonctionne-t-elle ? Quelles sont ses propriétés ?

Au chapitre III, nous avons examiné le problème standard de NMF sous l'angle de l'état de l'art et notamment de ses propriétés théoriques (existence et non-unicité des solutions), des fonctions de coût sous-jacentes et des algorithmes existants. Si la contrainte de non-négativité et la réduction de rang semblent, au premier abord, être les ressorts qui permettent de dégager des données une forme de représentation sémantique, il demeure de nombreux problèmes. Le choix d'un coût adapté (la divergence d'Itakura-Saito, dont nous avons montré l'intérêt par des arguments théoriques et expérimentaux) améliore la représentation, mais renforce le problème des minima locaux, ce que nous avons pu vérifier expérimentalement. De plus, les algorithmes multiplicatifs sont heuristiques, et les éléments de preuves de convergence sont rares et limités.

Les algorithmes de NMF sont-ils efficaces ? Peut-on les améliorer ?

Au chapitre IV, nous avons proposé une première évaluation des algorithmes multiplicatifs pour les trois coûts usuels, en particulier sous les angles de la vitesse de convergence et des minima locaux. Les fonctions de coût euclidienne, de Kullback-Leibler et d'Itakura-Saito souffrent de minima locaux que les algorithmes multiplicatifs ne parviennent pas à éviter. Un choix d'initialisation autre que le pur hasard ne résout pas le problème. Nous avons développé un algorithme de type tempéré pour essayer de résoudre ce problème. Si les résultats à petite échelle sont satisfaisants (meilleur valeur finale du critère, meilleur performance de transcription sur un exemple), le passage à l'échelle met en lumière la faiblesse de l'approche qui ne consiste qu'à poursuivre un hypothétique minimum global ; en effet, on n'observe pas de corrélation évidente entre la valeur du critère minimisé et la performance de transcription. Cela suggère que la seule contrainte de non-négativité n'est pas suffisante pour l'analyse de sons musicaux.

L'approche totalement aveugle est-elle suffisante ? Quel est le degré minimum de connaissances à injecter dans le système ? Quelles connaissances et sous quelle forme ?

Si le caractère aveugle de la NMF nous a séduite au début de cette étude (chapitre I), les constats précédents, tout comme l'état de l'art dressé au chapitre II nous convainquent de la nécessité d'introduire des connaissances supplémentaires pour améliorer la qualité de la séparation. Au chapitre V, nous avons eu recours à la littérature pour examiner et discuter différentes contraintes ajoutées au problème de NMF standard pour améliorer la forme des solutions. Cela nous a amenée à proposer un modèle tenant compte de l'harmonicité des sons musicaux. Ce modèle est directement intégré dans un algorithme multiplicatif original, d'une part, puis converti dans un cadre probabiliste, au chapitre VII. L'intérêt du passage au paradigme bayésien est multiple : il offre un cadre théorique solide, et la possibilité d'introduire d'autres contraintes sous la forme de distributions a priori sur les paramètres. De plus, on a montré au chapitre VI une équivalence formelle entre la NMF par minimisation de la

divergence d'Itakura-Saito, d'une part, et l'estimation du maximum de vraisemblance dans le modèle que nous proposons, d'autre part. Ceci apporte un argument supplémentaire au choix de la divergence d'Itakura-Saito comme mesure de dissimilarité entre spectres musicaux. Finalement, au terme de cette étude, nous proposons plusieurs algorithmes, de type EM ou en équivalent multiplicatif, réalisant la factorisation en matrices non-négatives du spectrogramme de puissance sous deux contraintes : l'harmonicité des spectres de la base, et la régularité temporelle des enveloppes. Dans la première évaluation du chapitre VIII, nous avons observé leurs bonnes propriétés en terme de convergence et de robustesse.

L'intégration d'information sous forme de contraintes est ici un choix à confronter à d'autres options, notamment l'utilisation de données extérieures sous la forme de bases d'apprentissage. En comparant nos algorithmes à deux méthodes de l'état de l'art utilisant de telles bases, nous avons pu constater l'efficacité de l'approche par contrainte, qui parvient à atteindre des résultats comparables à partir des seules données analysées. De plus, l'économie de données extérieures offre une plus grande capacité de généralisation à la méthode : ainsi, nous avons pu constater, sur un morceau joué par des instruments à vent, que les algorithmes de NMF incluant la contrainte d'harmonicité restaient relativement efficaces, tandis que l'un des algorithmes de référence, spécialement conçu pour le piano, échouait à fournir une transcription satisfaisante du morceau. Il faut cependant noter que nous n'avons pu conserver un schéma totalement aveugle : les patrons harmoniques inclus dans le modèle sont fixés à partir de connaissances psycho-acoustiques, qui sont d'une certaine manière issues d'autres données.

Comment convertir la factorisation en réelle transcription ? S'agit-il d'une approche efficace ?

La NMF, contrainte ou non, fournit une représentation de mi-niveau du signal, intermédiaire entre le son brut et une transcription symbolique de haut-niveau. Si la pertinence de la représentation peut sauter aux yeux sur des exemples simples, les potentialités d'un système de transcription complet basé sur la NMF restaient à prouver. C'est ce que nous avons fait en proposant un système de transcription « WAV vers MIDI » au chapitre IX, et en évaluant ses performances dans une tâche réaliste de transcription musicale au chapitre X. Nos résultats établissent la compétitivité de ces algorithmes avec l'état de l'art, bien qu'ils souffrent de certains défauts que nous avons discutés.

Perspectives

Les expériences de transcription en conditions réelles ont mis en lumière plusieurs raisons pour lesquelles les systèmes basés sur la NMF souffrent de faiblesses à corriger. En particulier, le post-traitement que nous avons proposé pour la détection des débuts et fins de notes montre de nombreux défauts, en partie car il implique le choix d'un seuil de détection auquel l'algorithme le précédant n'est pas forcément robuste. Un traitement plus raffiné, prenant en compte des phénomènes courants en musique telles que les notes répétées ou les trilles, pourrait améliorer la performance et la robustesse au choix d'un seuil.

Par ailleurs, partant d'un système purement aveugle, nous avons dû injecter de plus en plus de connaissances dans le système, diminuant probablement ses capacités d'adaptation au signal. Certains de ces paramètres, comme l'enveloppe spectrale de sous-bande définissant les patrons harmoniques à bande étroite P_{km} ou le paramètre de forme α des lois inverse-Gamma modélisant les enveloppes temporelles, sont fixés arbitrairement. Il faudrait envisager de les apprendre, soit sur une base extérieure, soit sur les données elles-mêmes. Des expériences préliminaires sur l'apprentissage des enveloppes tem-

porelles à l'aide d'une base extérieure de notes isolées n'ont cependant pas été concluantes. Des travaux ont été poursuivis par nos collaborateurs dans cette direction [Vincent *et al.*, 2010]. Dans le cas du paramètre de forme, des calculs informels suggèrent que son intégration au jeu de paramètres à estimer nécessiterait le recours à des techniques d'inférence bayésienne plus sophistiquées que celles déployées ici, telles que les méthodes de Monte-Carlo (MCMC).

Le cadre probabiliste étudié ici ouvre, dans ces directions, de nombreuses perspectives. Il supporte une forme de hiérarchisation, permettant par exemple de poser sur les paramètres des lois a priori elles-mêmes paramétrées par des quantités que l'on peut apprendre de la même manière. Dans cette optique, une possibilité d'améliorer le traitement temporel des enveloppes h_k serait de déployer un modèle hiérarchique à états (attaque, entretien, silence par exemple), qui constituerait une loi a priori pour les paramètres du modèle ; l'estimation des probabilités de transition entre états et le décodage pourraient être faits en même temps que l'estimation des paramètres du premier niveau de modélisation.

De plus, le cadre SAGE déployé permet de traiter chaque composante de manière différente. On pourrait ainsi choisir des lois a priori différenciées suivant le type de composante : une loi de régularité temporelle pour les parties stationnaires du signal, comme cela est fait dans notre étude, et une loi par exemple plus parcimonieuse (telle qu'une loi laplacienne) pour les composantes représentant les transitoires.

Toutefois, toutes ces idées ne permettent pas de s'affranchir d'une limitation intrinsèque et fondamentale du modèle de NMF : l'hypothèse que le spectre d'une note au cours du temps n'est que la répétition de la même enveloppe spectrale à un gain global près est limitante, et interdit la représentation de phénomènes courants en musique, tels que le vibrato ou les non-linéarités lors de changements de nuance. Pour pallier ce problème, on a vu apparaître ces dernières années des modèles convolutifs de NMF, permettant la représentation des variations de contenu spectral d'une note au cours de son évolution temporelle [Virtanen, 2004, Smaragdis, 2004, Smaragdis, 2007, Ozerov et Févotte, 2009]. Ces méthodes s'avèrent très puissantes et on peut imaginer, en leur ajoutant des contraintes similaires à celles proposées ici, atteindre de très bonnes performances de transcription, sur des classes bien plus vastes de signaux musicaux.

Bibliographie

Bibliographie de l'auteur

— Articles de revues —

[Bertin *et al.*, 2010] N. Bertin, R. Badeau, et E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. on Audio, Speech and Language Processing*, 18(3) :538–549, février 2010.

[Févotte *et al.*, 2009] C. Févotte, N. Bertin, et J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3) :793–830, mars 2009.

[Vincent *et al.*, 2010] E. Vincent, N. Bertin, et R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. on Audio, Speech and Language Processing*, 18(3) :528–537, février 2010.

— Articles de conférences —

[Bertin et Badeau, 2008] N. Bertin et R. Badeau. Initialization, distances and local minima in audio applications of the non-negative matrix factorization. Dans *Proc. of Acoustics'08, JASA*, 2008. `http://perso.telecom-paristech.fr/~nbertin/publis/acoustics08.pdf`.

[Bertin *et al.*, 2007] N. Bertin, R. Badeau, et G. Richard. Blind signal decompositions for automatic transcription of polyphonic music : NMF and K-SVD on the benchmark. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 1, pages 65–68, Honolulu, Hawaii, USA, 15-20 avril 2007.

[Bertin *et al.*, 2009a] N. Bertin, R. Badeau, et E. Vincent. Fast Bayesian NMF algorithms for enforcing harmonicity and smoothness in polyphonic music transcription. Dans *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, New Paltz, NY, 18-21 Octobre 2009.

[Bertin *et al.*, 2009b] N. Bertin, C. Févotte, et R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, Taipei, R.O.C., 18-24 avril 2009.

[Emiya *et al.*, 2009] V. Emiya, N. Bertin, B. David, et R. Badeau. MAPS : a piano database for multipitch estimation and automatic transcription. Dans *10th International Symposium on Music Information Retrieval (ISMIR 2009)*, 2009. Soumis.

[Vincent *et al.*, 2007] E. Vincent, N. Bertin, et R. Badeau. Two nonnegative matrix factorization methods for polyphonic pitch transcription. Dans *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, University of Vienna, Austria, 23-30 septembre 2007.

[Vincent *et al.*, 2008] E. Vincent, N. Bertin, et R. Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 109–112, Las Vegas, Nevada, USA, 30 Mars - 4 avril 2008.

Bibliographie du document

[Abdallah et Plumbley, 2004] S.A. Abdallah et M.D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. Dans *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 318–325, Barcelone, Espagne, 10-14 octobre 2004.

[Abdallah et Plumbley, 2006] S.A. Abdallah et M.D. Plumbley. Unsupervised analysis of polyphonic music using sparse coding. *IEEE Transactions on Neural Networks*, 17 :179–196, janvier 2006.

[Aharon *et al.*, 2005] M. Aharon, M. Elad, et A. Bruckstein. K-SVD and its non-negative variant for dictionary design. Dans *Proc. of the SPIE conference wavelets*, pages 327–339, San Diego, USA, juillet 2005.

[Aharon *et al.*, 2006] M. Aharon, M. Elad, A. Bruckstein, et Y. Katz. K-SVD : An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54 :4311–4322, novembre 2006.

[Alonso-Arevalo, 2006] M.A. Alonso-Arevalo. *Extraction d'information rythmique à partir d'enregistrements musicaux*. Thèse de doctorat, Institut Télécom, Télécom ParisTech, 2006.

[Anttila *et al.*, 1995] P. Anttila, P. Paatero, U. Tapper, et O. Järvinen. Source identification of bulk wet deposition in Finland by positive matrix factorization. *Atmospheric Environ.*, 29 :1705–1718, mai 1995.

[Auger et Flandrin, 1995] F. Auger et P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Proc.*, 43 :1068–1089, mai 1995.

[Bello *et al.*, 2006] J.P. Bello, L. Daudet, et M.B. Sandler. Automatic piano transcription using frequency and time-domain information. *IEEE Trans. on Audio, Speech and Language Processing*, 14 :2242–2251, novembre 2006.

[Bello *et al.*, 2000] J. Bello, G. Monti, et M. Sandler. Techniques for automatic music transcription. Dans *Proc. of International Conference on Music Information Retrieval (ISMIR'00)*, Plymouth, MA, USA, 23-25 octobre 2000.

[Benaroya *et al.*, 2006] L. Benaroya, R. Blouet, C. Févotte, et I. Cohen. Single sensor source separation using multiple-window STFT representation. Dans *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, 12-14 septembre 2006.

[Benaroya *et al.*, 2003] L. Benaroya, L. McDonagh, R. Gribonval, et F. Bimbot. Non negative sparse representation for Wiener based source separation with a single sensor. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pages 613Ű–616, Hong Kong, Chine, 6-10 avril 2003.

[Berry et Browne, 2005] M.W Berry et M. Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11 :249–264, octobre 2005.

[Berry *et al.*, 2007] M.W Berry, M. Browne, A.N Langville, V.P. Pauca, et R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52(1) :155–173, septembre 2007.

[Blumensath et Davies, 2006] T. Blumensath et M. Davies. Sparse and shift-invariant representations of music. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1) :50–57, janvier 2006.

[Bregman, 1994] A. Bregman. *Auditory Scene Analysis. The perceptual organization of sounds*. The MIT Press, septembre 1994.

[Brown, 1991] J. Brown. Calculation of a constant q spectral transform. *J. Acoust. Soc. Am.*, 89(1) :425–434, 1991.

[Brown, 1992] J. Brown. Musical fundamental frequency tracking using a pattern recognition method. *J. Acous. Soc. Amer.*, 92(3) :1394–1402, 1992.

[Cardoso, 1998] J.F. Cardoso. Blind signal separation : statistical principles. Dans *Proc. IEEE. Special issue on blind source separation*, volume 9, pages 2009–2025, 1998.

[Cemgil, 2004] A.T. Cemgil. *Bayesian Music Transcription*. Thèse de doctorat, Radboud University Nijmegen, Pays-Bas, 2004.

[Cemgil *et al.*, 2006] A.T. Cemgil, H.J. Kappen, et D. Barber. A generative model for music transcription. *IEEE Trans. on Audio, Speech and Language Processing*, 14(2) :679–694, Mars 2006.

[Cemgil, 2008] A. T. Cemgil. Bayesian inference in non-negative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2008. Article n° 785152.

[Chaigne et Kergomard, 2008] A. Chaigne et J. Kergomard. *Acoustique des instruments de musique*. Echelles. Belin, 2008.

[Chailley et Challan, 1951] J. Chailley et H. Challan. *Théorie de la musique*, volume 2. Leduc, 1951.

[Champagne et Tremblay, 2003] F. Champagne et G. Tremblay. Application de la distance d'édition à la correction de dictées musicales. 2003. Disponible à l'adresse `www.info2.uqam.ca/~tremblay/Seminaire-CADiM.ppt`.

[Chen *et al.*, 2006] Zhe Chen, Andrzej Cichocki, et Tomasz M. Rutkowski. Constrained nonnegative matrix factorization method for EEG analysis in early detection of Alzheimer's disease. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, volume 5, pages 893–896, Toulouse, France, 14-19 mai 2006.

[Chien et Jeng, 2002] Y.R. Chien et S.K. Jeng. An automatic transcription system with octave detection. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, volume 2, pages 1865–1868, Orlando, FL, USA, 13-17 mai 2002.

[Choi, 2008] Seungjin Choi. Algorithms for orthogonal nonnegative matrix factorization. Dans *Proc. of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008*, pages 1828–1832, Hong-Kong, Chine, 1e-6 juin 2008.

[Cichocki *et al.*, 2006] A. Cichocki, R. Zdunek, et S. Amari. Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. Dans *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39, Charleston, SC, USA, 5-8 mars 2006.

[Cichocki *et al.*, 2008] A. Cichocki, R. Zdunek, et S. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, pages 142–145, janvier 2008.

[Cichocki *et al.*, 2009] A. Cichocki, R. Zdunek, A.-H. Phan, et S. Amari. *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis*. John Wiley, novembre 2009.

[Cohen et Gannot, 2007] I. Cohen et S. Gannot. Spectral enhancement methods. Dans *Springer handbook of speech processing*. J. Benesty and Y. Huang (Eds.), New York, Springer, 2007.

[Cont, 2006] Arshia Cont. Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. Dans *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, Toulouse, France, 14-17 mai 2006.

[Cont *et al.*, 2007] A. Cont, S. Dubnov, et D. Wessel. Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. Dans *Proc. Int. Conf. Digital Audio Effects (DAFx)*, pages 85–92, Bordeaux, France, 10-15 septembre 2007.

[Cullin, 2008] O. Cullin. L'histoire paradoxale de la notation musicale. *Sciences Humaines*, Juin-Juillet-Août 2008.

[Dablemont, 2008] P.-A. Dablemont. Petite histoire de la notation musicale. Site internet, 2008. `http://pierre-arnaud-dablemont.com/blog`.

[Daniel *et al.*, 2008] A. Daniel, V. Emiya, et B. David. Perceptually-based evaluation of the errors usually made when automatically transcribing music. Dans *Proc. International Symposium on Music Information Retrieval (ISMIR'08)*, Philadelphie, PA, USA, septembre 2008.

[Daudet et Torrésani, 2006] L. Daudet et B. Torrésani. Sparse adaptive representations for musical signals. Dans *Signal Processing Methods for Music Transcription*. Springer, 2006.

[Davy *et al.*, 2006] M. Davy, S. Godsill, et J. Idier. Bayesian analysis of polyphonic western tonal music. *J. Acoust. Soc. Am.*, 119(4) :2498–2517, avril 2006.

[de Cheveigné, 2003] A. de Cheveigné. Separation of concurrent harmonic sounds : Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acous. Soc. Amer.*, 93(6) :3271–3290, juin 2003.

[de Cheveigné et Kawahara, 2002] A. de Cheveigné et H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4) :1917–1930, avril 2002.

[Demars, 2004] C. Demars. Représentations bidimensionnelles d'un signal de parole : éléments de monographie. Disponible à l'adresse `http://www.limsi.fr/Individu/chrd/rapport2004.pdf.gz`, mai 2004.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, et D.B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[Dhillon et Modha, 2001] I.S. Dhillon et D.S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1) :143–175, janvier 2001.

[Dhillon et Sra, 2005] I.S. Dhillon et S. Sra. Generalized nonnegative matrix approximations with bregman divergences. *Advances in Neural Information Processing Systems*, 18 :283–290, 2005.

[Ding *et al.*, 2006] C. Ding, T. Li, et W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing : Equivalence, chi-square statistic, and a hybrid method. Dans *Proc. of the National Conference on Artificial Intelligence*, volume 21, Boston, MA, USA, 16-20 juillet 2006.

[Ding *et al.*, 2008] C. Ding, T. Li, et W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52(8) :3913–3927, April 2008.

[Dixon, 2000] S. Dixon. On the computer recognition of solo piano music. Dans *Proc. Australasian Computer Music Conference*, pages 31–37, Brisbane, Australie, 5-8 juillet 2000.

[Donoho et Stodden, 2003] D. Donoho et V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts ? Dans *Advances in Neural Information Processing Systems*, volume 16, 2003.

[Doval et Rodet, 1991] B. Doval et X. Rodet. Estimation of fundamental frequency of musical sound signals. Dans *Proc. International Conference on Audio, Speech and Signal Processing (ICASSP'91)*, volume 5, pages 3657–3660, Toronto, Canada, 14-17 avril 1991.

[Drakakis *et al.*, 2008] K. Drakakis, S. Rickard, R. de Fréin, et A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum*, 3(37–40) :1853–1870, 2008.

[Duan *et al.*, 2008] Z. Duan, Y. Zhang, C. Zhang, et Z. Shi. Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Trans. Audio, Speech and Language Processing*, 16(4) :766–778, avril 2008.

[Durrieu *et al.*, 2008] J.L. Durrieu, G. Richard, et B. David. Singer melody extraction in polyphonic signals using source separation methods. *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 169–172, 31 mars - 4 avril 2008.

[Eggert et Körner, 2004] J. Eggert et E. Körner. Sparse coding and NMF. Dans *IEEE International Joint Conference on Neural Networks*, volume 4, pages 2529–2533, Budapest, Hongrie, 25-29 juillet 2004.

[Eguchi et Kano, 2001] S. Eguchi et Y. Kano. Robustifying maximum likelihood estimation. Rapport Technique, Tokyo Institute of Statistical Mathematics, 2001. Disponible à l'adresse `http://www.ism.ac.jp/~eguchi/pdf/RobustifyMLE.pdf`.

[Emiya, 2008] V. Emiya. *Transcription automatique de la musique de piano.* Thèse de doctorat, Institut Télécom, Télécom ParisTech, 2008.

[Emiya *et al.*, 2007] V. Emiya, R. Badeau, et B. David. Multipitch estimation of inharmonic sounds in colored noise. Dans *Proc. 10th International Conference on Digital Audio Effects (DAFx)*, Bordeaux, France, 10-15 septembre 2007.

[Emiya *et al.*, 2008] V. Emiya, R. Badeau, et B. David. Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. Dans *Proc. Eur. Conf. Sig. Proces. (EUSIPCO)*, Lausanne, Suisse, 25-29 août 2008.

[Feder et Weinstein, 1988] M. Feder et E. Weinstein. Parameter estimation of superimposed signals using the em algorithm. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(4) :477–489, avril 1988.

[Fessler et Hero, 1994] J. A. Fessler et A. O. Hero. Space-alternating generalized expectation maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10) :2664–2677, Octobre 1994.

[Finesso et Spreij, 2004] L. Finesso et P. Spreij. Approximate nonnegative matrix factorization via alternating minimization. Dans *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems*, Louvain, Belgique, 5-9 juillet 2004.

[Fletcher et Rossing, 1998] N.H. Fletcher et T.D. Rossing. *The Physics of Musical Instruments.* Springer, 1998.

[Févotte, 2007] C. Févotte. Bayesian audio source separation. Dans S. Makino, T.-W. Lee, et H. Sawada, éditeurs, *Blind speech separation*, pages 305–335. Springer, 2007.

[Févotte et Cemgil, 2009] C. Févotte et A. T. Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. Dans *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, Écosse, 24-28 août 2009.

[Gaussier et Goutte, 2005] E. Gaussier et C Goutte. Relation between PLSA and NMF and implications. Dans *Proc. 28th ACM International Conference on Research and Development of Information Retrieval (SIGIR)*, pages 601–602, Salvador, Brazil, 15-19 août 2005.

[Gillet et Richard, 2006] O. Gillet et G. Richard. ENST-drums : an extensive audio-visual database for drum signals processing. Dans *Proc. International Symposium on Music Information Retrieval (ISMIR'06)*, Victoria, Canada, 8-12 octobre 2006.

[Godsill, 2002] S. Godsill. Bayesian harmonic models for musical pitch estimation and analysis. Dans *Proc. IEEE International Conference on Audio, Speech and Signal Processing (ICASSP'02)*, volume 2, pages 1769–1772, Orlando, FL, USA, 13-17 mai 2002.

[Godsmark et Brown, 1999] D. Godsmark et G.J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27(3) :351–366, avril 1999.

[Gonzales et Zhang, 2005] E. F. Gonzales et Y. Zhang. Accelerating the lee-seung algorithm for non-negative matrix factorization. Rapport Technique, Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA, 2005.

[Goto *et al.*, 2002] M. Goto, H. Hashiguchi, T. Nishimura, et R. Oka. RWC music database : Popular, classical, and jazz music databases. Dans *Proc. of the 3rd International Symposium on Music Information Retrieval (ISMIR 2002)*, pages 287–288, Paris, France, 13-17 Octobre 2002.

[Gray *et al.*, 1980] R. M. Gray, A. Buzo, A. H. Gray, et Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4) :367–376, août 1980.

[Gribonval et Bacry, 2003] R. Gribonval et E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions in Signal Processing*, 51(1) :101–111, janvier 2003.

[Gribonval et Lesage, 2006] R. Gribonval et S. Lesage. A survey of sparse component analysis for source separation : principles, perspectives, and new challenges. Dans *Proc. 14th European Symposium on Artificial Neural Networks (ESANN'06)*, pages 323–330, Bruges, Belgique, 26-28 avril 2006.

[Guillamet *et al.*, 2001] D. Guillamet, B. Schiele, et J. Vitrià. Color histogram classification using NMF. Rapport Technique 057, Centre de Visió Per Computador, Université autonome de Barcelone, Octobre 2001.

[Guillamet et Vitri, 2002] David Guillamet et Jordi Vitri. Determining a suitable metric when using non-negative matrix factorization. Dans *Proc. International Conference on Pattern Recognition*, volume 2, pages 128–131, Québec, Canada, 11-15 août 2002.

[Hainsworth, 2001] S.W. Hainsworth. Analysis of musical audio for polyphonic transcription. Rapport Technique, University of Cambridge, UK, août 2001.

[Hainsworth, 2004] S. W. Hainsworth. *Techniques for the Automated Analysis of Musical Audio.* Thèse de doctorat, Université de Cambridge, UK, septembre 2004.

[Hartmann, 1998] W.M. Hartmann. *Signals, sounds and sensation.* American Institute of Physics, New York, 1998.

[Hess, 1983] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, New York, NY, USA, 1983.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic analysis. Dans *Proc. Uncertainty in Artificial Intelligence, UAI*, pages 289–296, Stockholm, Norvège, 30 juillet - 1e août 1999.

[Hopke, 2000] J. Hopke. A guide to positive matrix factorization. Rapport Technique, Materials from EPA Workshop on UNMIX and PMF as applied to PM2.5, 2000.

[Hoyer, 2002] P. Hoyer. Non-negative sparse coding. Dans *Proc. IEEE 12th Workshop on Neural Networks for Signal Processing*, pages 557–565, Martigny, Suisse, 4-6 septembre 2002.

[Hoyer, 2004] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5 :1457–1469, décembre 2004.

[Itakura et Saito, 1968] F. Itakura et S. Saito. Analysis synthesis telephony based on the maximum likelihood method. Dans *Proc. 6th International Congress on Acoustics*, pages C17–C20, Tokyo, Japon, août 1968.

[Jeter et Pye, 1982] M.W. Jeter et W.C. Pye. Some nonnegative matrices without nonnegative rank factorizations. *Industrial Mathematics*, 32 :37–41, 1982.

[Junnto et Paatero, 1994] S. Junnto et P. Paatero. Analysis of daily precipitation data by positive matrix factorization. *Environmetrics*, 5(2) :127–144, juin 1994.

[Kameoka, 2007] H. Kameoka. *Statistical Approach to Multipitch Analysis*. Thèse de doctorat, Université de Tokyo, 2007.

[Kameoka *et al.*, 2005] H. Kameoka, T. Nishimoto, et S. Sagayama. Harmonic temporal-structured clustering via deterministic annealing em algorithm for audio feature extraction. Dans *Proc. International Symposium on Music Information Retrieval (ISMIR)*, Londres, UK, 11-15 septembre 2005.

[Kameoka *et al.*, 2007] H. Kameoka, T. Nishimoto, et S. Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. Audio, Speech and Language Processing*, 15(3) :982–994, 2007.

[Karjalainen et Tolonen, 1999] M. Karjalainen et T. Tolonen. Separation of speech signals using iterative multi-pitch analysis and prediction. Dans *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2187–2190, Budapest, 5-9 septembre 1999.

[Kashino *et al.*, 1998] K. Kashino, K. Nakadai, T. Kinoshita, et H. Tanaka. Application of the Bayesian probability network to music scene analysis. Dans *Computational Auditory Scene Analysis*, pages 115–137. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 1998.

[Kim et Choi, 2006] M. Kim et S. Choi. Monaural music source separation : Nonnegativity, sparseness, and shift-invariance. Dans *Proc. 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, pages 617–624, Charleston, SC, USA, 5-8 mars 2006.

[Kim *et al.*, 2005] S-.P. Kim, Y.N. Rao, D. Ergodmus, J.C. Sanchez, M.A.L. Nicolelis, et J.C. Principe. Determining patterns in neural activity for reaching movements using nonnegative matrix factorization. *EURASIP Journal on Applied Sig. Proc., Special Issue of Trends in Brain Computer Interfaces*, 2005(19) :3113–3121, 2005.

[Kim et Choi, 2007] Y.-D. Kim et S. Choi. A method of initialization for nonnegative matrix factorization. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volume 2, pages 537–540, Honolulu, USA, 15-20 Avril 2007.

[Klapuri, 2001a] A.P. Klapuri. Means of integrating audio content analysis algorithms. Dans *Proc. 110th convention of the audio engineering society (AES)*, Amsterdam, Pays-Bas, 12-15 mai 2001.

[Klapuri, 2001b] A.P. Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 5, pages 3381–3384, Salt Lake City, UT, USA, mai 2001.

[Klapuri, 2003] A.P. Klapuri. Automatic transcription of music. Dans *Proceedings of the Stockholm Music Acoustics Conference (SMAC)*, volume II, pages 587–590, Stockholm, Norvège, 6-9 août 2003.

[Klapuri, 2005] A.P. Klapuri. A perceptually motivated multiple-f0 estimation method. Dans *Proc. IEEE Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, New Paltz, NY, USA, 16-19 octobre 2005.

[Klapuri, 2008] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. Audio, Speech and Language Processing*, 16(2) :255–266, février 2008.

[Klapuri et Davy, 2006] A.P. Klapuri et M. Davy. *Signal Processing Methods for Music Transcription*. Springer, 2006.

[Klingenberg *et al.*, 2009] B. Klingenberg, J. Curry, et A. Dougherty. Non-negative matrix factorization : Ill-posedness and a geometric algorithmstar. *Pattern Recognition*, 42(5) :918–928, mai 2009.

[Kompass, 2007] R. Kompass. A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, 19(3) :780–791, mars 2007.

[Kuhn et Tucker, 1951] H. W. Kuhn et A. W. Tucker. Nonlinear programming. Dans *Proc. 2nd Berkeley Symposium*, pages 481–492, Berkeley, CA, USA, 31 juillet - 12 août 1951.

[Laroche, 1995] J. Laroche. *Traitement des Signaux Audio-Fréquences*. École Nationale Supérieure des Télécommunications, 1995. Polycopié de cours.

[Laurberg *et al.*, 2008] H. Laurberg, M. Christensen, M.D. Plumbley, L.K. Hansen, et S.H. Jensen. Theorems on positive data : On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008, 2008. Article n°764206.

[Lee et Seung, 1999] D.D. Lee et H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 :788–791, octobre 1999.

[Lee et Seung, 2001] D.D. Lee et H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13 :556–562, 2001.

[Leipp, 1971] E. Leipp. *Acoustique et musique*. Masson, 1971.

[Lesser *et al.*, 1993] V. Lesser, H. Nawab, I. Gallastegi, et F. Klassner. IPUS : an architecture for integrated signal processing and signal interpretation in complex environments. Dans *Proc. AAAI*, pages 249–255, Raleigh, NC, USA, 22-24 octobre 1993.

[Leveau, 2007] P. Leveau. *Décompositions parcimonieuses structurées : application à la représentation objet de la musique*. Thèse de doctorat, Université Pierre et Marie Curie, 2007.

[Leveau *et al.*, 2008] P. Leveau, E. Vincent, G. Richard, et L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(1) :116–128, janvier 2008.

[Li *et al.*, 2001] S.Z. Li, X. Hou, H. Zhang, et Q. Cheng. Learning spatially localized, parts-based representation. Dans *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 207–212, Hawaii, USA, 11-13 décembre 2001.

[Lin, 2007a] C.-J. Lin. On the convergence of multiplicative update algorithms for non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6) :1589–1596, novembre 2007.

[Lin, 2007b] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10) :2756–2779, octobre 2007.

[Lin, 2008] C.-J. Lin. Nonnegative matrix factorization based on alternating nonnegativity- constrained least squares and the active set method. *SIAM Journal in Matrix Analysis and Applications*, 30(2) :713–730, juillet 2008.

[Lissandre, 1990] M. Lissandre. *Maîtriser SADT*. Armand Colin, 1990.

[Liu et Yuan, 2008] Weixiang Liu et Kehong Yuan. Sparse p-norm nonnegative matrix factorization for clustering gene expression data. *International Journal of Data Mining and Bioinformatics*, 2(3) :236–249, juillet 2008.

[Liu et Zheng, 2004] Weixiang Liu et Nanning Zheng. Non-negative matrix factorization based methods for object recognition. *Pattern Recognition Letter*, 25(8) :893–897, mars 2004.

[Liu *et al.*, 2004] W. Liu, N. Zheng, et X. Li. Nonnegative matrix factorization for EEG signal classification. Dans *International Symposium on Neural Networks*, volume II, pages 470–475, Dalian, Chine, 19-21 août 2004.

[MacQueen, 1967] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. Dans *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, USA, 27 décembre - 7 janvier 1967.

[Mallat et Zhang, 1993] S. Mallat et Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12) :3397–3415, décembre 1993.

[Marolt, 2004] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Trans. on Multimedia*, 6(3) :439–449, Juin 2004.

[Marolt, 2009] M. Marolt. Non-negative matrix factorization with selective sparsity constraints for transcription of bell chiming recordings. Dans *Proc. of the 6th Sound and Music Computing Conference (SMC)*, Porto, Portugal, 23-25 juillet 2009.

[Martin, 1996a] K. D. Martin Martin. Automatic transcription of simple polyphonic music : Robust front end processing. Dans *3rd Joint Meeting of the Acoustical Societies of America and Japan*, Honolulu, HI, USA, 23-28 décembre 1996.

[Martin, 1996b] K. D. Martin Martin. A blackboard system for automatic transcription of simple polyphonic music. Rapport Technique TR.385, Media Laboratory, MIT, 1996.

[Meddis et Hewitt, 1991] R. Meddis et M.J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. *J. Acous. Soc. Amer.*, 89(6) :2866–2882, juin 1991.

[Minc, 1988] H. Minc. *Nonnegative matrices*. John Wiley and sons, New York, 1988.

[Mongeau et Sankoff, 1990] M. Mongeau et D. Sankoff. Comparison of musical sequences. *Computer and the Humanities*, 24(3) :161–175, juin 1990.

[Monti et Sandler, 2002] G. Monti et M. Sandler. Automatic polyphonic piano note extraction using fuzzy logic in a blackboard system. Dans *Proc. 5th International Conference on Digital Audio Effects (DAFx)*, pages 39–44, Hambourg, Allemagne, 26-28 septembre 2002.

[Moorer, 1975] J.A. Moorer. *On the segmentation and analysis of continuous musical sound by digital computer*. Thèse de doctorat, CCRMA, Université de Stanford, 1975.

[Murray et Kreutz-Delgado, 2004] J.F. Murray et K. Kreutz-Delgado. Sparse image coding using learned overcomplete dictionaries. Dans *Proc. 14th IEEE Workshop on Machine Learning for Signal Processing*, pages 579–588, São Luís, Brésil, 29 septembre - 1eoctobre 2004.

[Niedermayer, 2008] B. Niedermayer. Non-negative matrix division for the automatic transcription of polyphonic music. Dans *Proc. International Symposium on Music Information Retrieval (ISMIR)*, pages 544–545, Philadelphie, PA, USA, 14-18 septembre 2008.

[Nii, 1986a] H. Penny Nii. Blackboard systems, part one : The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2) :38–53, juin-juillet 1986.

[Nii, 1986b] H. Penny Nii. Blackboard systems, part two : Blackboard application systems, blackboard systems from a knowledge engineering perspective. *AI Magazine*, 7(3) :82–106, août 1986.

[Noll, 1967] A. M. Noll. Cepstrum pitch determination. *J. Acous. Soc. Amer.*, 41(2) :293–309, février 1967.

[Oja et Plumbley, 2003] E. Oja et M. Plumbley. Blind separation of positive sources using non-negative PCA. Dans *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, pages 11–16, Nara, Japon, 1e-4 avril 2003.

[Ortiz-Berenguer, 2002] L.I. Ortiz-Berenguer. *Identificatión automática de acordes musicales*. Thèse de doctorat, Universidad Politécnica de Madrid, 2002.

[Ortiz-Berenguer *et al.*, 2004] L.I. Ortiz-Berenguer, F.J. Casajús-Quirós, M. Torres-Guijarro, et J.A. Beracoechea. Piano transcription using pattern recognition : aspects on parameter extraction. Dans *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, pages 212–216, Naples, Italie, 5-8 octobre 2004.

[Ozerov et Févotte, 2009] A. Ozerov et C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. Audio, Speech and Language Processing*, 2009. À paraître.

[Ozerov *et al.*, 2007] A. Ozerov, P. Philippe, F. Bimbot, et R. Gribonval. Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(5) :1564–1578, juillet 2007.

[Paatero, 1997] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chronometrics and Intelligent Laboratory Systems*, 37(1) :23–35, mai 1997.

[Paatero et Tapper, 1994] P. Paatero et U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2) :111–126, juin 1994.

[Pascual-Montano *et al.*, 2006] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, et R.D Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(3) :403–415, mars 2006.

[Paterson *et al.*, 1999] K.G. Paterson, J.L. Sagady, D.L. Hooper, S.B. Bertman, M.A. Carroll, et P.B. Shepson. Analysis of air quality data using positive matrix factorization. *Environ. Sci. Technol.*, 33(4) :635–641, février 1999.

[Patterson *et al.*, 1991] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, et M. Allerhand. Complex sounds and auditory images. Dans *9th International Symposium on Hearing : Auditory Physiology and Perception*, pages 429–446, Carcens, France, 9-14 juin 1991.

[Paulus et Virtanen, 2005] Jouni Paulus et Tuomas Virtanen. Drum transcription with non-negative spectrogram factorisation. Dans *Proc. 13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turquie, 4-8 septembre 2005.

[Peeters, 2006] G. Peeters. Music pitch representation by periodicity measures based on combined temporal and spectral representations. Dans *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Toulouse, France, 14-19 mai 2006.

[Pielemeier *et al.*, 1996] W.J. Pielemeier, G.H. Wakefield, et M.H. Simoni. Time-frequency analysis of musical signals. *Proc. of the IEEE*, 9(84) :1216–1230, septembre 1996.

[Piszczalski, 1986] M. Piszczalski. *A computational model of music transcription.* Thèse de doctorat, University of Michigan, Ann Arbor, 1986.

[Piszczalski et Galler, 1979] M. Piszczalski et B. Galler. Predicting musical pitch from component frequency ratios. *J. Acoust. Soc. Am.*, 66(3) :710–720, septembre 1979.

[Plumbley, 2002] M.D. Plumbley. Conditions for non-negative independent component analysis. *IEEE Signal Processing Letters*, 9 :177–180, juin 2002.

[Plumbley, 2003] M.D. Plumbley. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, 14(3) :534–543, mars 2003.

[Plumbley, 2004] M.D. Plumbley. A "nonnegative PCA" algorithm for independent component analysis. *IEEE Trans. on Neural Networks*, 15(1) :66–76, janvier 2004.

[Plumbley *et al.*, 2002] M.D. Plumbley, S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti, et M.B. Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6) :603–627, septembre 2002.

[Plumbley *et al.*, 2006] M. D. Plumbley, S. A. Abdallah, T. Blumensath, et M. E. Davies. Sparse representations of polyphonic music. *Signal Processing*, 86(3) :417–431, mars 2006.

[Poliner et Ellis, 2007] G.E Poliner et D.P.W. Ellis. A discriminative model for polyphonic piano transcription. *Eurasip Journal of Applied Signal Processing (special issue on Music Signal Processing)*, 2007(1) :154–162, janvier 2007.

[Polissar *et al.*, 1998] A.V. Polissar, P.K. Hopke, W.C. Malm, et J.F. Sisler. Atmospheric aerosol over Alaska. 2. Elemental composition and sources. *Journal of Geophysical Research*, 103(D15) :19045–19057, 1998.

[Rabiner, 1977] L. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Trans. Acoustics, Speech and Signal Processing*, 25(1) :24–33, février 1977.

[Rabiner *et al.*, 1976] L. Rabiner, M. Cheng, A. Rosenberg, et C. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Trans. Acoustics, Speech and Signal Processing*, 24(5) :399–418, octobre 1976.

[Raczyński *et al.*, 2007] S.A. Raczyński, N. Ono, et S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. Dans *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, Vienne, Autriche, 23-27 septembre 2007.

[Raphael, 2002] C. Raphael. Automatic transcription of piano music. Dans *Proc. Internationam Conference on Music Information Retrieval (ISMIR)*, Paris, France, 13-17 octobre 2002.

[Rigouste, 2006] L. Rigouste. *Méthodes probabilistes pour l'analyse exploratoire de données textuelles.* Thèse de doctorat, Télécom ParisTech, 2006.

[Robert, 2007] Christian P. Robert. *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics).* Springer Verlag, New York, 2e édition édition, Juin 2007.

[Ross *et al.*, 1974] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, et H. Manley. Average magnitude difference function pitch extractor. *IEEE Trans. Acoustics, Speech and Signal Processing*, 22(5) :353–362, 16-19 octobre 1974.

[Rossi, 1998] L. Rossi. *Identification de sons polyphoniques de piano.* Thèse de doctorat, Université de Corse, 1998.

[Ryynänen et Klapuri, 2005] M. Ryynänen et A. Klapuri. Polyphonic music transcription using note event modeling. Dans *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 319–322, New Paltz, New York, USA, 16-19 octobre 2005.

[Ryynänen et Klapuri, 2008] M. Ryynänen et A. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3) :72–86, printemps 2008.

[Schmidt et Laurberg, 2008] M.N. Schmidt et H. Laurberg. Nonnegative matrix factorization with gaussian process priors. *Computational Intelligence and Neuroscience*, 2008, janvier 2008. Article n°361705.

[Schroeder, 1968] M. R. Schroeder. Period histogram and product spectrum : New methods for fundamental-frequency measurement. *J. Acous. Soc. Amer.*, 43(4) :829–834, avril 1968.

[Segbroeck et hamme, 2009] Maarten Van Segbroeck et Hugo Van hamme. Unsupervised learning of time-frequency patches as a noise-robust representation of speech. *Speech Communication*, Á paraître, 2009.

[Shahnaz *et al.*, 2006] F. Shahnaz, M.W. W. Berry, V.P. Pauca, et R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Journal on Information Processing and Management*, 42(2) :373–386, mars 2006.

[Shang, 2008] L. Shang. Non-negative sparse coding shrinkage for image denoising using normal inverse gaussian density model. *Image and Vision Computing*, 26(8) :1137–1147, août 2008.

[Shashanka *et al.*, 2008a] M. Shashanka, B. Raj, et P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008, 2008. Article n°947438.

[Shashanka *et al.*, 2008b] M. Shashanka, B. Raj, et P. Smaragdis. Sparse overcomplete latent variable decomposition of counts data. *Advances in neural information processing systems*, 20 :1313–1320, 2008.

[Smaragdis, 2004] Paris Smaragdis. Non-negative matrix factor deconvolution ; extraction of multiple sound sources from monophonic inputs. Dans *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA'04)*, pages 494–499, Grenade, Espagne, 22-24 septembre 2004.

[Smaragdis, 2007] P. Smaragdis. Convolutive speech bases and their application to speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1) :1–12, janvier 2007.

[Smaragdis et Brown, 2003] P. Smaragdis et J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. Dans *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, pages 177–180, New Paltz, New York, USA, 19-22 octobre 2003.

[Sterian *et al.*, 1999] A. Sterian, M. Simoni, et G.H. Wakefield. Model based musical transcription. Dans *Proc. International Computer Music Conference (ICMC)*, Beijing, Chine, 22-27 octobre 1999.

[Tan et Févotte, 2009] V.Y.F. Tan et C. Févotte. Automatic relevance determination in nonnegative matrix factorization. Dans *Proc. Atelier Structure et parcimonie pour la représentation adaptative de signaux (SPARS'09*, Saint Malo, France, 6-9 avril 2009.

[Theis *et al.*, 2005] F.J. Theis, K. Stadlthanner, et T. Tanaka. First results on uniqueness of sparse non-negative matrix factorization. Dans *Proc. EUSIPCO 2005*, Antalya, Turkey, 4-8 septembre 2005.

[Thomas, 1974] L. Thomas. Solution to problem 73-14, rank factorizations of nonnegative matrices. *SIAM Review*, 16(3) :393–394, 1974.

[Tolonen et Karjalainen, 2000] T. Tolonen et M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Trans. on Speech and Audio Processing*, 8(6) :708–716, 2000.

[Tropp, 2003] J.A Tropp. Literature survey : Non-negative matrix factorization. Rapport Technique, EE381K-14 Multidimensional Digital Signal Processing - Spring 2003 Projects, The University of Texas at Austin, 2003.

[van de Par *et al.*, 2002] S. van de Par, A. Kohlrausch, G. Charestan, et R. Heusdens. A new psycho-acoustical masking model for audio coding applications. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1805–1808, 2002.

[van Rijsbergen, 1979] C.J. van Rijsbergen. *Information retrieval*. Butterworths, London, UK, 2nd édition, 1979.

[Viitaniemi *et al.*, 2003] T. Viitaniemi, A. Klapuri, et A. Eronen. A probabilistic model for the transcription of single-voice melodies. Dans *Proc. of Finnish Signal Processing Symposium*, Tampere, Finlande, 19 mai 2003.

[Vincent, 2006] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1) :91–98, janvier 2006.

[Vincent et Plumbley, 2005] E. Vincent et M.D. Plumbley. Predominant-F0 estimation using Bayesian harmonic waveform models. Dans *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, Londres, UK, 11-15 septembre 2005.

[Vincent et Plumbley, 2007] E. Vincent et M.D. Plumbley. Low bitrate object coding of musical audio using bayesian harmonic models. *IEEE Trans. on Audio, Speech and Language Processing*, 15(4) :1273–1282, mai 2007.

[Vinokourov, 2002] A. Vinokourov. Why non-negative matrix factorization works well for text information retrieval. Disponible à l'adresse `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.4936&rep=rep1&type=ps`, 2002.

[Virtanen, 2004] T. Virtanen. Separation of sound sources by convolutive sparse coding. Dans *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Corée, 3 Octobre 2004. Article n°55.

[Virtanen, 2006] T. Virtanen. *Sound Source Separation in Monaural Music Signals*. Thèse de doctorat, Tampere University of Technology, Finlande, November 2006.

[Virtanen, 2007] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3) :1066–1074, mar 2007.

[Virtanen *et al.*, 2008] T. Virtanen, A. T. Cemgil, et S. Godsill. Bayesian extensions to nonnegative matrix factorisation for audio signal modelling. Dans *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 1825–1828, Las Vegas, NV, USA, 30 mars - 4 avril 2008.

[Virtanen et Klapuri, 2002] T. Virtanen et A. Klapuri. Separation of harmonic sounds using linear models for the overtone series. Dans *Proc. IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), volume 2, pages 1757–1760, Orlando, FL, USA, 13-17 mai 2002.

[Walmsley *et al.*, 1999] P. Walmsley, S. Godsill, et P. Rayner. Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters. Dans *Proc. IEEE Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, New Paltz, NY, USA, 12-20 octobre 1999.

[Wang et Plumbley, 2005] B. Wang et M.D. Plumbley. Musical audio stream separation by non-negative matrix factorization. Dans *Proc. of the DMRN Summer Conference*, Glasgow, UK, 23-24 juillet 2005.

[Wang *et al.*, 2006] F.-Y. Wang, C.-Y. Chi, T.-H. Chan, et Y. Wang. Blind separation of positive dependent sources by non-negative least-correlated component analysis. Dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, pages 73–78, Maynooth, Irlande, 6-8 Septembre 2006.

[Wang *et al.*, 2009] Fa-Yu Wang, Chong-Yung Chi, Tsung-Han Chan, et Yue Wang. Non-negative least-correlated component analysis for separation of dependent sources by volume maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), mars 2009.

[Wang et Zou, 2008] W. Wang et X. Zou. Non-negative matrix factorization based on projected nonlinear conjugate gradient algorithm. Dans *Proc. ICA Research Network International Workshop (ICARN 2008)*, pages 5–8, Liverpool, UK, 25-26 septembre 2008.

[Wang *et al.*, 2004] Y. Wang, Y. Jia, C. Hu, et M. Turk. Fisher non-negative matrix factorization for learning local features. Dans *Asian Conference on Computer Vision*, Jeju, Corée, 24-27 janvier 2004.

[Welch, 1938] B.L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3–4) :350–362, 1938.

[Wiener, 1949] N. Wiener. *Extrapolation, interpolation and smoothing of stationary time series*. MIT Press, 1949.

[Wild *et al.*, 2004] S. Wild, J. Curry, et A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11) :2217–2232, novembre 2004.

[Winther et Petersen, 2007] O. Winther et K. B. Petersen. Bayesian independent component analysis : Variational methods and non-negative decompositions. *Digital Signal Processing*, 17(5) :858–872, septembre 2007.

[Wise *et al.*, 1976] J. Wise, J. Caprio, et T. Parks. Maximum likelihood pitch estimation. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(5) :418–423, octobre 1976.

[Xue *et al.*, 2008] Yun Xue, Chong Sze Tong, Ying Chen, et Wen-Sheng Chen. Clustering-based initialization for non-negative matrix factorization. *Applied Mathematics and Computation*, 205(2) :525–536, novembre 2008.

[Young *et al.*, 2006] S.S. Young, P. Fogel, et D. Hawkins. Clustering scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14 :11–13, 2006.

[Zdunek et Cichocki, 2007] R. Zdunek et A. Cichocki. Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8) :1904–1916, août 2007.

[Zdunek et Cichocki, 2008] R. Zdunek et A. Cichocki. Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems. *Computational Intelligence and Neuroscience*, 2008, 2008. Article nř939567.

[Zhang et Fang, 2007] Ye Zhang et Yong Fang. A NMF algorithm for blind separation of uncorrelated signals. Dans *Proc. of the 2007 International Conference on Wavelet Analysis and Pattern Recognition*, pages 999–1003, Beijing, Chine, 2-4 novembre 2007.

[Zheng *et al.*, 2007] Z. Zheng, J. Yang, et Y. Zhu. Initialization enhancer for non-negative matrix factorization. *Engineering Applications of Artificial Intelligence*, 20(1) :101–110, février 2007.

[Zwicker et Fastl, 1999] E. Zwicker et H. Fastl. *Psychoacoustics : Facts and Models, 2nd Edition.* Springer, Heidelberg, 1999.

Cinquième partie

Annexes

Annexe A

Article

Dans cette annexe, nous reproduisons l'article [Bertin *et al.*, 2007], publié au début de cette thèse et dont le contenu n'a pas été développé dans le corps de ce document afin d'éviter de le surcharger.

BLIND SIGNAL DECOMPOSITIONS FOR AUTOMATIC TRANSCRIPTION OF POLYPHONIC MUSIC: NMF AND K-SVD ON THE BENCHMARK

Nancy Bertin, Roland Badeau, Gaël Richard

GET-Télécom Paris (ENST) - Signal and Image Processing Department
46 rue Barrault - 75634 PARIS Cedex 13, FRANCE
nbertin@enst.fr

ABSTRACT

This paper investigates on the behavior of two blind signal decomposition algorithms, non negative matrix factorization (NMF) and non negative K-SVD (NKSVD), in a polyphonic music transcription task. State-of-the-art transcription systems are based on a frame-by-frame, low-level approach; blind systems could be an alternative to them. Two raw but effective audio-to-MIDI systems are proposed and evaluated. Performances are similar, but in favor of NMF, which is more robust to initialization, choice of the order and computationnally less costly.

Index Terms— Automatic transcription, polyphonic music, non negative matrix factorization, K-SVD.

1. INTRODUCTION

Automatic music transcription consists in deriving a symbolic representation (*e.g.* a MIDI-like file) of the music from an audio file. Transcribing monodic music is henceforth a well understood problem; but the case of polyphonic music remains a largely open question.

To address this issue, most of the proposed approaches rely on prior knowledge (*e.g.* signal models [1] or supervised learning [2, 3]) and/or frame-by-frame low-level analysis. The main weakness of this kind of methods is their low capacity to adapt to signals that do not comply with the model. In order to avoid this drawback, a more recent set of approaches consists in using as few hypotheses as possible about the audio content and trying to separate the notes blindly. Among those techniques we find: sparse coding [4], non-negative matrix factorization (first introduced for image processing in [5]), blind source separation [6] (*e.g.* independent component analysis), and their variants [7]. They rely on few and weak hypotheses, and show promising results in polyphonic music transcription.

The work presented here investigates further on the efficiency of this kind of approach for a full audio-to-MIDI transcription. In particular, two algorithms are studied: non-negative matrix factorization (NMF), proposed in [8], and the non-negative variant of the k-means singular value decompositions algorithm (NKSVD), successfully applied to image processing in [9]. If both provide an exploitable decomposition, they behave differently with respect to design choices like initialization, length of the analyzed piece or order of the model, which will be discussed here as well.

The two studied algorithms are briefly presented in section 2 and their implementation in an effective transcription system is described in section 3. We then present in section 4 carried on experiments, and their results in section 5. Conclusions and directions for future work are proposed in section 6.

The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference fot automatic annotation and retrieval of multimedia content - K-SPACE, and by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales - DESAM. The authors also wish to thank Dr. Juan Bello and Dr. Laurent Daudet for their sharing audio and MIDI data.

2. BLIND SIGNAL DECOMPOSITIONS

The two algorithms studied in this paper are briefly described below since more details can be found in [10] for NMF and in [9] for NKSVD. Although they have their own specificities, both algorithms rely on common hypotheses and principles.

2.1. Common framework

Both algorithms consider the magnitude spectrogram of the data as a linear combination of r elementary spectra, or atoms, at each time step; determining a decomposition consists in finding the basis of elementary atoms, and the decomposition of the data on this basis. The magnitude spectrum is obviously not additive, but this is a relevant approximation in many cases.

Basically, let us consider a time-frequency representation V of a musical excerpt. V is in $\mathbb{R}_+^{m\times n}$ where m is the number of frequency bins and n the number of time frames. We search for two matrices $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ such that:

$$V \approx WH \tag{1}$$

The approximation is to be understood as a minimization of a "distance" (which has to be defined) between the original V and its reconstruction WH. The specific property exploited here is the non-negativity of all these matrices: they only have zero or positive coefficients. Columns of W are seen as frequency-domain atoms, and lines of H are the temporal activities of each of these atoms in the observed signal. At each time frame j, the spectrum V_j is thus expressed as a linear combination of several atoms, the coefficients of the combination being given by the j-th column of H. The recovered atoms are interpreted as *notes* and the matrix H as *temporal activities*.

2.2. Non-negative matrix factorization (NMF)

In Non-Negative Matrix Factorization, the non-negativity of the matrices involved is the only constraint used to process the decomposition. The approximation comes from the constraint $r < min(n, m)$, so that the factorization is also a rank reduction. Considering a long

enough music piece, containing a certain number of musical events, the most natural (and, we hope, only) way to represent the signal on a reduced-sized basis should be to have a basis of *notes*.

The factorization is processed by iterative minimization of a cost function (Frobenius distance or I-divergence, see [10]) and leading to a local minimum; non-negativity is guaranteed by multiplicative update rules at each iteration.

2.3. Non-negative K-SVD (NKSVD)

K-SVD and its non-negative variant, implemented here, is a sparse-coding-like algorithm, developed for image coding and denoising purpose. In typical algorithms, sparse coding (determination of the decomposition of a signal in a given basis) and dictionary design (determination of the basis in which signals will be decomposed) are often conducted separately. K-SVD was proposed, as a generalization of the k-means algorithm, in order to simultaneously get (and in an unsupervised way) both basis and decomposition.

Sparsity is the property of having few non-zero entries. In music, the intuitive interpretation is that among all possible notes (the 88 keys of the piano for instance), only a few of them can be played simultaneously by a human musician. It is used in addition to the non-negativity constraint to compute the decomposition. In this model, W is considered as a dictionary, and is generally searched overcomplete, that is $r >> m$. Two quantities have to be minimized during the calculation: the reconstruction error (taken as the Frobenius norm of $V - WH$) and the l^0-norm of H, which conveys the notion of sparsity. This is performed iteratively, with two steps at each iteration: pursuit (W being fixed, find the best and sparsest H) and dictionary update (refine W, using singular value decompositions (SVD)).

2.4. Main differences

The main difference between the two algorithms is a matter of "philosophy", in particular as far as the order of the model is concerned. NMF aims at reducing rank, in order to let emerge a meaningful representation; whereas NKSVD, which is in first intention a coding algorithm, aims only at finding an economical representation (through sparsity), with an overcomplete dictionary. Another difference is related to complexity: NKSVD is much more costly than NMF (about 10 times more CPU time), because of the performed singular value decompositions and because the dictionary is supposed to be overcomplete. Eventually, NKSVD allows to control the degree of sparsity of the decomposition; in NMF, sparsity comes as an uncontrolled side effect.

3. THE AUTOMATIC TRANSCRIPTION SYSTEM

The audio-to-MIDI system consists of three steps: processing of a time-frequency representation, its factorization by one of the previously presented algorithms, post-processing of the factors to get a MIDI representation.

3.1. Pre-processing

The short-time Fourier transform of the signal is computed, using a 64 ms Hanning window[1] (2822 samples at 44.1 kHz) with a 50% overlap, and its module is taken to get the non-negative matrix V. There are 4096 frequency bins, negative frequencies being then discarded, leading to 2048-line matrices.

[1]This is obviously a long window, made necessary by frequency resolution considerations; the choice of the Fourier transform is a well-known limit in this domain, and improvements are bound to find a smarter time-frequency representation, not searched here.

3.2. Factorization

Non-negative matrix factorization (NMF) and the non-negative variant of the K-SVD (NKSVD) are then performed iteratively on V until convergence is reached. Initialization is either random, as proposed in the original papers, or set to the spectrum of isolated real piano notes. We implement the most common version of NMF, described in [10]; in NKSVD, the pursuit stage is performed by Matching Pursuit, with a number of retained coefficients set to 10 (*i.e.* at each frame, at most 10 atoms are active simultaneously).

3.3. Post-processing

The post-processing step consists in interpreting the factors W and H as respectively, pitched atoms and temporal activities. Each column of W is considered as a note spectrum. Its pitch is estimated by the maximum of the sum of the log-spectra (rather than the spectral product, in order to avoid numerical errors). Lines of H associated with atoms of the same pitch are summed. We then determine notes onsets and offsets by thresholding the lines of H: the atom j is turned on at time k when h_{jk} exceeds a threshold and turned off when it is below this threshold. The threshold is fixed empirically as the sum of the mean and standard deviation of the line. Velocity was not treated, and arbitrarily set to a constant value. Finally, components without identified pitch and too short events are discarded. A note event is thus described by a pitch, an onset time and a duration.

This post-processing provides a raw, yet useful transcription, for we can listen to the result (by re-synthesis from the MIDI) and compare it to a reference. As our purpose was to focus on the previous (factorization) step, we chose coarse methods for this last part, being aware of their weakness and the necessity to refine them in the future.

4. EXPERIMENTS

4.1. Database

Six pieces from the classical piano repertoire, described in [11], are used for tests. They were recorded on a Disklavier mechanically playing the MIDI file in input. Thus, we have a MIDI reference for each test piece, allowing quantitative comparison with the processed transcription. Each piece was then re-synthetized from the MIDI in order to compare the system performances on real and synthetic audio. In addition, the system was also tested on *La Campanella* by Liszt from the RWC database [12], to evaluate the system in more realistic conditions.

4.2. Parameters

Each piece, in real and synthetized versions, was then systematically analyzed with different values of the model order r. Initialization was either random or fixed to real piano spectra from the RWC database [12].

4.3. Performance evaluation

Music transcription system performance evaluation is a challenging issue, which has not yet reached a consensus in the community.

Possible metrics are overviewed in [13]: frame-level metrics (precision/recall and scores agregated from them), note-level metrics, note onset detection error rates. In order to avoid penalizing the whole system because of the coarseness of the back-end (especially temporal) and the slight misalignments in the reference (due to technical constraints of MIDI acquisition), we chose to measure the note onset detection error, with a tolerance of one window (64ms) before and after the reference onset time.

True positive (TP) is the number of correctly detected notes, false positive (FP) is the number of wrong notes detected and false negative (FN) the number of missing notes. *Precision rate* is the ratio $TP/(TP+FP)$, *recall rate* is the ratio $TP/(TP+FN)$, and *overall accuracy* is defined according to [13] as the ratio $TP/(TP+FP+FN)$.

5. RESULTS

5.1. Preliminary tests

As [8] suggests that notes have to be played alone at least once in the piece in order to get a proper separation, we analyzed this simple example:

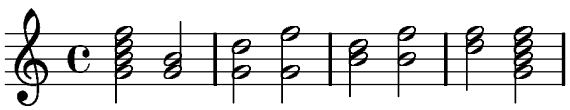


Fig. 1. *A simple example.*

This excerpt was generated by a MIDI synthetizer (with a piano timbre) and analyzed by NMF, with the number of components set to $r = 4$ (4 different notes are heard). Figure 2 shows the visualization of the result: the 4 columns of W on the left, the corresponding lines of H on the right. This way, each left-right pair of graphs shows the frequency content of one component and its temporal occurrences in the analyzed piece.

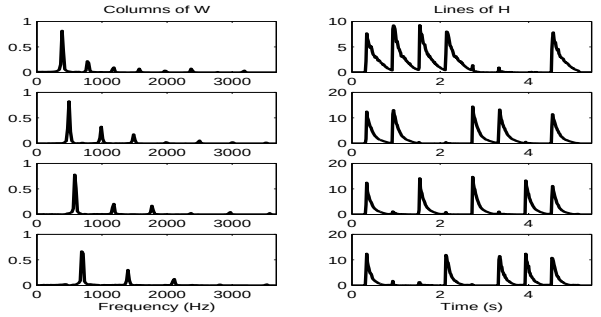


Fig. 2. *NMF result of the analysis of Figure 1.*

This case is rather ideal: the synthetic sounds make two notes of the same pitch strictly identical, and every possible combination of 2 notes among 4 is played. However, we can also notice that no note is being played alone, yet the algorithm succeeds in separating them. This suggests that NMF can efficiently separate a polyphony even if the notes are never isolated, provided enough various combinations are heard.

5.2. Main experiments: general observations

Though the pieces are rather difficult to transcribe, performed transcriptions are good enough to easily recognize the piece when the audio signal is synthetized from the transcribed MIDI file. Most errors are typical of the difficulty of the task: octave-related pitch errors (substitution, insertion or deletion of a note having the same chroma[2] as the target), note detection errors (notes late or too short, or spurious notes, depending on the choice of the threshold), bad representation of low-pitched notes, missing notes in chords of 4 notes or more. Performances follow the same trends when analyzing either synthetic or real sounds. This suggests that it would not be unreasonable to test transcription algorithms on databases containing mainly synthetic sounds.

Table 1 shows the mean results of the analysis of the six pieces by NMF and NKSVD, in their synthetic and real versions. r was set to 88.

Table 1. Transcription results (r=88).

	Algorithm	NMF		NKSVD	
	Initialization	no	yes	no	yes
Synth. audio	Precision (%)	52.4	51.4	36.7	44.9
	Recall (%)	49.3	54.4	35.9	40.2
	Accuracy (%)	34.5	36.1	22.4	25.6
Real audio	Precision (%)	51.5	45.5	47.0	47.8
	Recall (%)	55.1	56.1	38.1	41.2
	Accuracy (%)	36.4	33.6	27.2	29.2

As a reference, for similar analyzed data and identical metrics, accuracies from 30% to 60% are obtained in [13]. Transcription of the Liszt piece showed similar trends, but with lower accuracy, certainly due to the use of the pedal and the fact that the piece is played by a real musician. On this piece, a significative difference was observed between NKSVD with and without structured initialization, in favour of the former.

Original and transcription audio examples are available at `http://www.enst.fr/~nbertin/icassp2007`.

5.3. Parameters influence

Besides the previous general remarks, we can make some additional observations about the behaviors of the system with respect to different parameters and design choices.

5.3.1. Order of the model

Following the original goal of the algorithms, a natural choice for the order r of the model would be: the theoretical r for NMF (*i.e.* the number of different pitches in the MIDI reference), and a largely overestimated r for NKSVD-based transcription. We chose however the same order values for both to keep them comparable. Results are relatively stable with regard to the chosen r, as shown on figure 3. NMF is slightly more sensitive to it, which was rather unexpected.

At any order r, we frequently observe that several atoms have the same pitch, corresponding to the most frequent notes appearing in the piece: it is a better strategy for the algorithms to represent more accurately notes frequently occurring, and the variability between notes of the same pitch increases the need for several atoms to represent each note. This suggests that overestimation of the order is preferable. This is confirmed by figure 3.

[2] *Chroma* is the modulus-12 pitch.

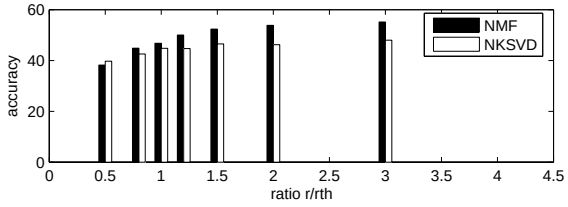


Fig. 3. *Accuracy w.r.t. the ratio between r and the number of different pitches in the reference MIDI file (denoted by rth).*

5.3.2. Duration of the transcribed piece

Classical, frame-by-frame approaches are indeed not sensitive to the length of the analyzed piece. On the contrary, we expect different performances of NMF and NKSVD-based transcriptions with respect to this length. Indeed, preliminary tests showed that the variety of chords (notes belonging to different chords) could help a lot to separate notes. We analyzed the first 30 seconds of each test piece separately, and compared the transcription performance with the results we get for the same 30 seconds analyzed within the whole piece. Accuracy is between 5% and 8% higher for the whole piece analysis. It is however difficult to claim without precaution that performance is all the better as the piece is long. The piece must remain somewhat homogeneous along time to benefit from notes redundancy.

5.3.3. Initialization

As both algorithms converge to local minima, the initialization of W and H may influence the results. We compared the results of a random initialization of W *vs.* a chosen one. Columns of W were then initialized with Fourier magnitude spectra of real piano notes. As we are not supposed to know neither the number of pitches in the piece, nor their values, we fixed $r = 88$ and initialized W with the 88 notes of the piano.

As shown on table 1, NMF and NKSVD did not show a particular sensitivity to initialization (except that convergence was reached in twice less iterations). Random and determined initialization lead to very similar results, except for NKSVD analysis of the Liszt piece (accuracy is 8% higher with non-random initialization). This observations could be explained by the match or mismatch between initialization spectra and the actual signals.

6. CONCLUSIONS AND FUTURE WORK

The goal of this preliminary work was to assess the potential of two promising approaches for music transcription. The first conclusion is that blind signal decomposition methods may be an alternative to frame-by-frame approach to build efficient transcription systems. Clearly, such methods provide an interesting mid-level representation that could lead, with an efficient post-processing of the decomposition, to a very accurate transcription.

The comparison between NMF and NKSVD does not highlight a clear superiority of one of them upon the other. NMF seems preferable for its lower computational cost. NKSVD with initialization was however better on the only real music test piece which suggests further investigation on a larger database.

There are a lot of remaining questions and possible improvements. First, an efficient model order estimation method is needed. Second, the problem of complexity and computational cost remains the main handicap of NKSVD. The choice of the pursuit algorithm and of the desired degree of sparsity are other questions to raise. Both methods still need a better back-end (pitch detection in atoms and onset detection in temporal envelopes). For both, the time-frequency representation remains unsatisfactory (because of the well known resolution trade-off of Fourier transform). Eventually, the limitation of the methods to stable spectral profiles (along one note) is a strong constraint, unrealistic in music (for instance in notes played *vibrato*). This has to be overtaken, for instance by taking into account the temporal evolution of the note spectrum.

7. REFERENCES

[1] A.P. Klapuri, "Automatic transcription of music," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC)*, Aug. 2003, vol. II, pp. 587–590.

[2] J. Bello, G. Monti, and M. Sandler, "Techniques for automatic music transcription," in *Proceedings of International Conference on Music Information Retrieval (ISMIR'00)*, Oct. 2000.

[3] M.D. Plumbley, S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti, and M.B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics and Systems*, vol. 33, no. 6, pp. 603–627, Sept. 2002.

[4] S.A. Abdallah and M.D. Plumbley, "Unsupervised analysis of polyphonic music using sparse coding," *IEEE Transactions on Neural Networks*, vol. 17, pp. 179–196, Jan. 2006.

[5] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[6] J.F. Cardoso, "Blind signal separation : statistical principles," in *Proc. IEEE. Special issue on blind source separation*, 1998, vol. 9, pp. 2009–2025.

[7] M.D. Plumbley, "Algorithms for nonnegative independant component analysis," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 534–543, 2003.

[8] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003, pp. 177–180.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Proceedings of the SPIE conference wavelets*, July 2005, vol. 5914, p. 591411.

[10] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[11] J.P. Bello, L. Daudet, and M.B. Sandler, "Automatic piano transcription using frequency and time-domain information.," *IEEE Transactions on Speech and Audio Processing*, Accepted for publication, 2006.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, October 2002, pp. 287–288.

[13] G.E Poliner and D.P.W. Ellis, "A discriminative model for polyphonic piano transcription," *Eurasip Journal of Applied Signal Processing (special issue on Music Signal Processing)*, 2006, to appear.

Annexe B

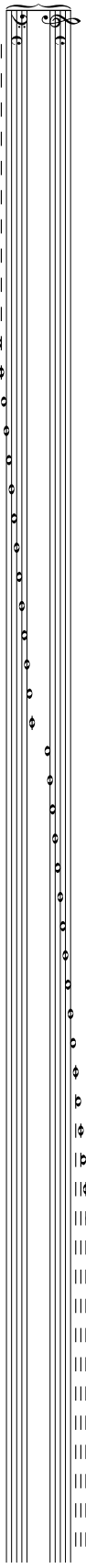
Échelles de hauteur

Le tableau B.1 donne la correspondance entre les notes (nom et numéro d'octave conventionnel), les fréquences fondamentales selon un tempérament égal (le diapason étant fixé à 440 Hz) et codes MIDI associés (entre 0 et 127). Seules les notes de la tessiture standard du piano sont représentées. Les formules de conversion sont les suivantes :

$$p_{MIDI} = \left[12 \log_2 \frac{f_0}{440}\right] + 69 \tag{B.1}$$

$$f_0 = 2^{\frac{p_{MIDI}-69}{12}} \times 440 \tag{B.2}$$

$$n^o\ d'octave = \left\lfloor \frac{p_{MIDI} - 60}{12} + 4 \right\rfloor \tag{B.3}$$



21	22	la_0			27.500	29.135	
23		si_0			30.868		
24	25	do_1			32.703	34.648	
26	27	$ré_1$			36.708	38.891	
28		mi_1			41.203		
29	30	fa_1			43.654	46.249	
31	30	sol_1			48.999	51.913	
33	34	la_1			55.000	58.270	
35		si_1			61.735		
36	37	do_2			65.406	69.296	
38	39	$ré_2$			73.416	77.782	
40		mi_2			82.407		
41	42	fa_2			87.307	92.499	
43	44	sol_2			97.999	103.83	
45	46	la_2			110.00	116.54	
47		si_2			123.47		
48	49	do_3			130.81	138.59	
50	51	$ré_3$			146.83	155.56	
52		mi_3			164.81		
53	54	fa_3			174.61	185.00	
55	56	sol_3			196.00	207.65	
57	58	la_3			220.00	233.08	
59		si_3			246.94		
60	61	do_4			**261.63**	277.18	
62	63	$ré_4$			293.67	311.13	
64		mi_4			329.63		
65	66	fa_4			349.23	369.99	
67	68	sol_4			392.00	415.30	
69	70	la_4			**440.00**	466.16	
71		si_4			493.88		
72	73	do_5			523.25	554.37	
74	75	$ré_5$			587.33	622.25	
76		mi_5			659.26		
77	78	fa_5			698.46	739.99	
79	80	sol_5			783.99	830.61	
81	82	la_5			880.00	932.33	
83		si_5			987.77		
84	85	do_6			1046.5	1108.7	
86	87	$ré_6$			1174.7	1244.5	
88		mi_6			1318.5		
89	90	fa_6			1396.9	1480.0	
91	92	sol_6			1568.0	1661.2	
93	94	la_6			1760.0	1864.7	
95		si_6			1975.5		
96	97	do_7			2093.0	2217.5	
98	99	$ré_7$			2349.3	2489.0	
100		mi_7			2637.0		
101	102	fa_7			2793.0	2960.0	
103	104	sol_7			3136.0	3322.4	
105	106	la_7			3520.0	3729.3	
107		si_7			3951.1		
108		do_8			4186.0		

TABLE B.1 – Correspondances entre échelles de hauteur.

Overview

Non-negative matrix factorization (NMF) is a powerful, unsupervised decomposition technique allowing the representation of two-dimensional non-negative data as a linear combination of meaningful elements in a basis.

NMF has been widely and successfully used to process audio signals, including various tasks such as monaural sound source separation [Virtanen, 2007], audio stream separation [Wang et Plumbley, 2005], audio-to-score alignment [Cont, 2006], drum transcription [Paulus et Virtanen, 2005]. In particular, it has been efficiently used to separate notes in polyphonic music [Smaragdis et Brown, 2003,Bertin *et al.*, 2007] and transcribe it in a symbolic format such as MIDI. In this case, a time-frequency representation of the signal is factored as the product between a basis (or dictionary) of pseudo-spectra and a matrix (decomposition) of time-varying gains. When obtained from harmonic instruments sounds, the basis is shown to partially retain harmonic components, with a pitched structure, that can be interpreted as musical notes, while the decomposition gives information about the onset and offset times of the associated notes.

Meaningful is here a key word : we expect the basis to be formed of interpretable elements, exhibiting certain semantics. The non-negativity constraint is a first step towards this interpretability, compared to other well-known techniques such as Singular Value Decomposition (SVD). For instance, the basis learnt by NMF from an image database is expected to contain meaningful images (the so-called "part-based representation" [Lee et Seung, 1999]). This interpretability is often observed in practice, which is certainly one of the reasons for NMF's popularity ; but it is not always as satisfying as expected (see, for instance, facial images in [Li *et al.*, 2001], that are expected to retain facial parts like eyes, nose, mouth, but do not exactly). As some other desirable characteristics of the decomposition, it is more observed as a welcome side-effect, than enforced and controlled.

To alleviate this lack of control on the decomposition properties, most authors have proposed constrained variants of NMF, ensuring and enhancing those side-effects of baseline NMF : sparsity, spatial localization, temporal continuity for instance. The typical approach for such constrained variants is to add a penalty term to the usual cost function (reconstruction error) and minimize their sum, see *e.g.* [Li *et al.*, 2001, Zhang et Fang, 2007, Virtanen, 2007].

On the other hand, several authors have imported the idea of a non-negative constraint in other frameworks than NMF, in particular statistical framework. We can cite non-negative variants of Independent Component Analysis (ICA) [Plumbley, 2003] and non-negative sparse coding [Abdallah et Plumbley, 2004]. The Bayesian framework offers both a strong theoretical framework, and the possibility to manage constraints through models and priors.

This thesis is divided in four parts. The first part introduces the music transcription task and draws the related state-of-the-art. The second part is devoted to the deterministic approach to NMF and its variants, whereas the third part addresses it in a probabilistic paradigm. The fourth and last part is

applicative, and proposes an experimental evaluation of the algorithms proposed in previous parts.

Part One : *Automatic music transcription*

Automatic music transcription consists in deriving a symbolic representation (*e.g.* a MIDI-like file) of the music from an audio file. Transcribing monodic music is henceforth a well understood problem ; but the case of polyphonic music remains a largely open question. In **chapter I**, we present this task in its various forms and the historical approaches to solve it. In **chapter II**, we discuss the state-of-the-art, introduce non-negative matrix factorization (NMF) as a music transcription tool and raise our problematics.

Most of the proposed approaches in automatic transcription rely on prior knowledge (*e.g.* signal models [Klapuri, 2003] or supervised learning [Bello *et al.*, 2000, Plumbley *et al.*, 2002]) and/or frame-by-frame low-level analysis. The main weakness of this kind of methods is their low capacity to adapt to signals that do not comply with the model, and the necessity to develop refined post-procesing to integrate in time the multiple fundamental frequency information. In order to avoid this drawback, a more recent set of approaches consists in using as few hypotheses as possible about the audio content and trying to separate the notes blindly. Among those techniques we find : sparse coding [Abdallah et Plumbley, 2006], non-negative matrix factorization (first introduced for image processing in [Lee et Seung, 1999]), blind source separation [Cardoso, 1998] (*e.g.* independent component analysis), and their variants [Plumbley, 2003]. They rely on few and weak hypotheses, and show promising results in polyphonic music transcription.

As far as NMF is concerned, it is simply defined as the problem of finding a factorization of a given matrix $\mathbf{V}$ of dimensions $F \times N$ and whose entries are non-negative real numbers :

$$\mathbf{V} \approx \mathbf{WH} = \hat{\mathbf{V}} \tag{B.4}$$

where $\mathbf{W}$ and $\mathbf{H}$ are matrices of dimensions $F \times K$ and $K \times N$ respectively, with all coefficients being non negative, and where the operator $\approx$ is an « approximation » that has to be defined. The order K of the model is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimension. In typical audio applications, the matrix $\mathbf{V}$ is chosen as a time-frequency representation (*e.g.* magnitude or power spectrogram), f denoting the frequency bin and n the time frame. Note that the factorization is in general only approximate, so that the terms “approximate nonnegative matrix factorization” or “nonnegative matrix approximation" also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, let us mention the problems of learning parts of faces and semantic features of text [Lee et Seung, 1999], polyphonic music transcription [Smaragdis et Brown, 2003], object characterization by reflectance spectra analysis [Berry *et al.*, 2007], portfolio diversification [Drakakis *et al.*, 2008], as well as Scotch whiskies clustering [Young *et al.*, 2006].

To illustrate the interest of NMF for musical signal, let us take a simple polyphonic example, whose score is represented on figure B.1.

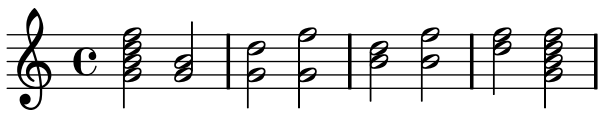


FIGURE B.1 – A simple polyphonic example.

This excerpt contains 8 musical events (chords), but only 4 notes. If we wish to represent its spectrogram $\mathbf{V}$ by a product between a dictionary $\mathbf{W}$ of 4 atoms and temporal envelopes $\mathbf{H}$, we can reasonably expect that the factorization will retain 4 notes in the dictionary. Each spectrum of a chord will then be represented as a linear combination of the spectra of the notes making the chord.

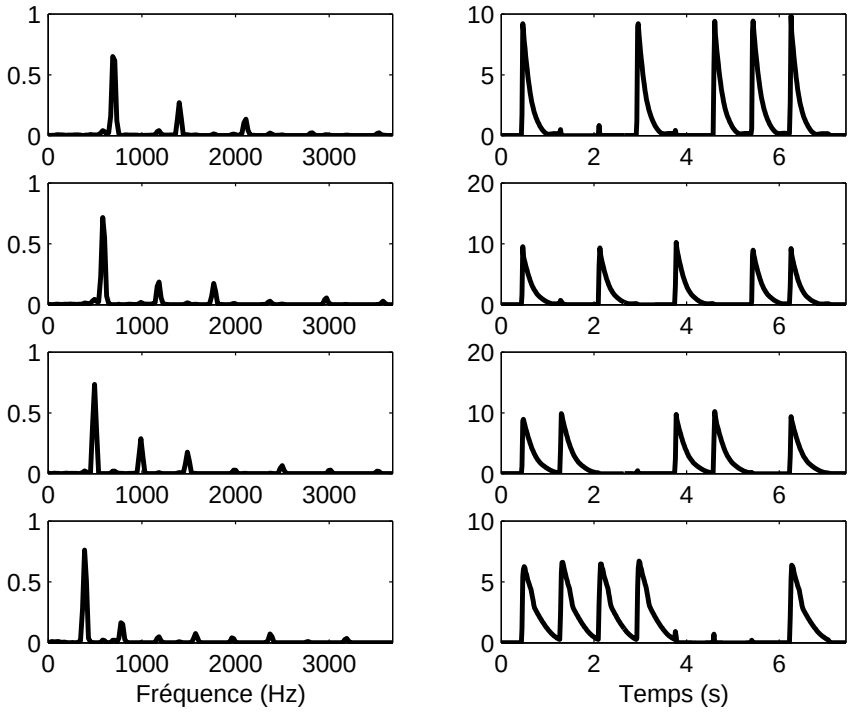


FIGURE B.2 – Factorization of the sequence on figure B.1

Part Two : *Deterministic approach to NMF*

The factorization (B.4) is generally obtained by minimizing a cost function defined by

$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \tag{B.5}$$

where $d(a|b)$ is a function of two scalar variables. d is typically non-negative and takes value zero if and only if (iff) $a = b$.

In **chapter III**, we examine usual cost functions for the NMF problem, and their properties. We also discuss the questions of existence, unicity, local minima and convergence of typical algorithms.

The most popular cost functions for NMF are the Euclidean (EUC) distance and the generalized Kullback-Leibler (KL) divergence, which were particularly popularized (as NMF itself) by Lee and Seung, see, *e.g.*, [Lee et Seung, 1999]. They described multiplicative update rules under which $D(\mathbf{V}|\mathbf{WH})$ is shown to be non-increasing, while ensuring non-negativity of $\mathbf{W}$ and $\mathbf{H}$. The update

rules are obtained by using a simple heuristics, which can be seen as a gradient descent algorithm with an appropriate choice of the descent step. By expressing the gradient of the cost function ∇D as the difference of two positive terms $\nabla^+ D$ and $\nabla^- D$, the cost function is shown (in particular cases) or observed to be nonincreasing under the rules :

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\nabla^-_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})}{\nabla^+_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\nabla^-_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})}{\nabla^+_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})} \end{cases} \tag{B.6}$$

For some choices of d, like EUC or KL, monotonicity of the criterion under these rules can be proven [Lee et Seung, 1999], but in the general case, these updates do not guarantee any convergence.

Local minima

The β-divergence introduced by Eguchi and Kano in [Eguchi et Kano, 2001] is defined as

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)} \left(x^\beta + (\beta-1)\, y^\beta - \beta\, x\, y^{\beta-1} \right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \log \frac{x}{y} + (y-x) & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{B.7}$$

As observed in [Cichocki *et al.*, 2006], the IS divergence is a limit case of the β-divergence. [Eguchi et Kano, 2001] assume $\beta > 1$, but the definition domain can very well be extended to $\beta \in \mathbb{R}$. The β-divergence is shown to be continuous in β by using the identity $\lim_{\beta \to 0} (x^\beta - y^\beta)/\beta = \log(x/y)$. EUC distance is obtained for $\beta = 2$, so that the β-divergence is inclusive for our three choices of NMF costs : EUC, KL and IS.

We studied $d_\beta(x|y)$ as a function of y (remembering that x acts as data). Its first and second-order derivatives write

$$\nabla_y d_\beta(x|y) = y^{\beta-2}(y-x), \tag{B.8}$$

$$\nabla^2_y d_\beta(x|y) = y^{\beta-3}\left((\beta-1)y + (2-\beta)x\right). \tag{B.9}$$

The next properties follow :

- $d_\beta(x|y)$ has a single minimum in $y = x$ and increases with $|y - x|$. This justifies its relevance as a measure of fit.
- $d_\beta(x|0)$ is finite iff $\beta \geq 1$.
- $d_\beta(x|y)$ is convex on $\mathbb{R}_+$ iff $1 \leq \beta \leq 2$.

The cost function $D_\beta(V|WH)$ is not convex in general wrt the pair (W, H), even if the cost $d_\beta(x|y)$ is convex (wrt y). However, when $d_\beta(x|y)$ is convex, D_β is at least convex as a function of W (resp. H) with fixed V and H (resp. W), because it is expressed as a sum of convex functions composed with linear functionals (see Eq. (III.3)).

Computing the gradient $\nabla_H D_\beta(V|WH)$ (resp. $\nabla_W D_\beta(V|WH)$) using Eq. (B.8), and multiplying H (resp. W) at previous iteration by the ratio of the negative and positive parts of the gradient, we

obtain the following alternate multiplicative algorithm [Cichocki *et al.*, 2006] :

$$H \leftarrow H \otimes \frac{W^T(V \otimes (WH)^{\cdot[\beta-2]})}{W^T((WH)^{\cdot[\beta-1]})} \tag{B.10}$$

$$W \leftarrow W \otimes \frac{(V \otimes (WH)^{\cdot[\beta-2]})H}{((WH)^{\cdot[\beta-1]})H^T} \tag{B.11}$$

where $\otimes$ and $\oslash$ denote (Hadamard) entrywise product and division respectively, the fraction is also entrywise and $A^{\cdot[n]}$ denotes the matrix with entries $[A]_{ij}^n$. For $\beta = 1, 2$, we obtain Lee and Seung's original algorithm. Using convexity of $d_\beta(x|y)$, monotonicity of the criterion under the latter rules can be shown for $1 \leq \beta \leq 2$ [Kompass, 2007]. In other cases, this monotonicity was observed in practice, though not proven.

As we observed in practice that IS-NMF is more prone to local minima [Févotte *et al.*, 2009, Bertin et Badeau, 2008], we used this study to describe a tempering scheme that favors convergence of IS-NMF to global minima. It simply consists of using β as a temperature parameter, which is set to a value between 1 and 2 in the first iterations (where the cost $D_\beta(V|WH)$ is at least convex wrt to either W or H) and gradually decrease it to the target cost, i.e, IS in our case, obtained for $\beta = 0$. As such, we simply apply update rules (B.10) and (B.11), with β being a function of the iteration number. More precisely, we use the template described by Fig. ; β takes value β_i during ℓ_i iterations, then starts to decrease following a cosine during ℓ_d more iterations, until it finally reaches its target value, to which it remains fixed during the last ℓ_e iterations.

The tempering approach is experimentally evaluated in **chapter IV**, as well as effects of different initializations. Our conclusion is that pursuing an hypothetic global minimum by a means or another is not enough to enhance the meaning and relevance of the produced representation, which lead us, in **chapter V**, to consider the idea of adding more constraints to the problem.

Constrained approaches

In standard NMF, the only constraint is the elementwise non-negativity of all matrices. All other properties of the decomposition, as satisfying as it is, come as uncontrolled side-effects and in a way, the fact that the decomposition retains certain semantics of the original signal, performs separation or provides meaningful and interpretable components is just "good news". It sounds thus natural to try to improve this potential by adding explicit constraints to the factorization problem, in order to enhance and control desired properties.

Then, several constraints have been introduced to get NMF solutions that better fit certain expectancies. Among other proposed constraints, we can cite sparsity [Hoyer, 2004], spatial localization [Li *et al.*, 2001], least correlation between sources [Zhang et Fang, 2007] or temporal continuity [Virtanen, 2007, Chen *et al.*, 2006]. Those constraints are examined and discussed in chapter V.

The commun point between those algorithms, whichever constraint is considered, is the "penalty term approach". Rather than minimizing only a reconstruction error term D_r (EUC or KL, typically), the minimized cost function includes a term D_c that quantifies the desired property. The constrained NMF problem is then expressed as :

$$\min_{\mathbf{W},\mathbf{H}} D_r(\mathbf{V}|\mathbf{WH}) + \lambda D_c(\mathbf{V}|\mathbf{WH})$$

where λ is a weight parameter.

Sparsity	$\sum_{k=1}^{K} \frac{1}{\sqrt{N}-1} \left(\sqrt{N} - \sum_{n=1}^{N} \lvert h_{kn} \rvert / \sqrt{\sum_{n=1}^{N} h_{kn}^2} \right)$	[Hoyer, 2004]
Spatial localization	$\lambda_1 \sum_{k=1}^{K} \sum_{k'=1}^{K} \left[\mathbf{W}^T \mathbf{W} \right]_{kk'} - \lambda_2 \sum_{k=1}^{K} \left[\mathbf{H}\mathbf{H}^T \right]_{kk}$	[Li *et al.*, 2001]
Least correlation	$\sum_{k=1}^{K} \log \left[\mathbf{H}\mathbf{H}^T \right]_{kk} - \log \lvert \mathbf{H}\mathbf{H}^T \rvert$	[Zhang et Fang, 2007]
Temporal continuity	$\sum_{k=1}^{K} \sum_{n=1}^{N} \lvert h_{kn} - h_{k(n-1)} \rvert^2$	[Virtanen, 2007]

TABLE B.2 – Some state-of-the-art constraints D_c in NMF problem.

Table B.2 gives a few examples of literature penalty terms. Temporal smoothness is one of these examples. In standard NMF and most of its variants, time frames are considered as independent, non-related observations, which is obviously not true for real-world sounds and in particular for music. In the case of musical notes, the main part of the note (the sustain and decay parts, after the attack) possesses a slowly time-varying spectrum. When expressed as the product between a template spectrum $\mathbf{w}_k$ and a time-varying gain h_k, according to NMF formulation, it is equivalent to saying that the row h_k is smooth, or, in other words, that the coefficient h_{kn} is not that different from $h_{k(n-1)}$.

[Virtanen, 2007] and [Chen *et al.*, 2006] thus introduce penalty terms in the NMF cost function to take into account this temporal continuity. In [Virtanen, 2007] the term is directly linked to the differences $h_{kn} - h_{k(n-1)}$, while [Chen *et al.*, 2006] variant relies on a ratio between short-time and long-time variance of h_k. Those terms are shown to favor smoothness in lines of $\mathbf{H}$. Another possible approach is the statistical approach from [Févotte *et al.*, 2009]. Temporal continuity is favored through putting an appropriate prior on $\mathbf{H}$. This solution will be exposed with more details and adapted to our case ilater in this overview.

It is interesting to notice that non-smoothness may also be an objective (see for instance [Pascual-Montano *et al.*, 2006]), depending on the data and the application. [Pascual-Montano *et al.*, 2006] points out that smoothness of one of the NMF factors (*i.e.* $\mathbf{W}$ or $\mathbf{H}$) may enhance sparsity of the other one, thus establishing a link between those two popular constraints. On the other hand, [Virtanen, 2007] combines sparsity and temporal continuity constraints on $\mathbf{H}$, but concludes to the non-efficiency of the sparsity constraint in his particular case.

The penalty approach has several drawbacks. First, a criterion quantifying the desired property must be found. Second, no general proof of convergence is available for the update scheme (B.6). Moreover, the parameter λ has to be chosen empirically. These reasons motivated our approach for harmonicity constraint in current work.

Musical notes, excluding transients, are pseudo-periodic. Their spectra are then comb-alike, with regularly spaced frequency peaks. As we wish to use NMF to separate musical notes in a polyphonic recording, we expect that elements in the basis $\mathbf{W}$ are as near as possible from a harmonic distribution. This property is yet not easily quantified by a penalty term.

In [Vincent *et al.*, 2008], we rather proposed an alternative model to baseline NMF problem, enforcing the basis harmonicity. We impose the basis components to be expressed as the linear combination of narrow-band harmonic spectra (patterns), which are arbitrarily fixed :

$$w_{fk} = \sum_{m=1}^{M} e_{mk} P_{km}(f) \tag{B.12}$$

For a given component number k, all the patterns $\mathbf{P}_{km}$ share the same pitch (fundamental frequency

$f_0)$; they are defined by summation of the spectra of a few adjacent individual partials at harmonic frequencies of f_0, scaled by the spectral shape of subband k. This spectral envelope is chosen according to perceptual modelling [Vincent *et al.*, 2008]. Figure V.1 illustrates the patterns for one note and the corresponding atom $\mathbf{w}_k$.

Coefficients e_{mk} are learned by NMF as well as the decomposition $\mathbf{H}$. Update rules are obtained by minimizing the same cost function as in baseline NMF, except that it is minimized with respect to (wrt) $\mathbf{E}$ and $\mathbf{H}$ rather than $\mathbf{W}$ and $\mathbf{H}$.

Part Three : *Probabilistic approach to NMF*

Another way to induce properties in the NMF is to switch to a statistical framework and introduce adequate prior distributions, which is done in **chapter VI**. Let us consider the following model, proposed in [Benaroya *et al.*, 2003, Benaroya *et al.*, 2006] : $\forall n = 1, ..., N$,

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{kn} \in \mathbb{C}^F \tag{B.13}$$

where latent variables $\mathbf{c}_{kn}$ are independent and follow a multivariate Gaussian distribution

$$\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\mathrm{diag}(\mathbf{w}_k)) \tag{B.14}$$

In [Févotte *et al.*, 2009], the estimation of the parameter $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, in a maximum likelihood (ML) sense is shown to be equivalent to solving the NMF problem $\mathbf{V} \approx \mathbf{WH}$, when observing $\mathbf{V} = (|x_{fn}|^2)_{fn}$ and choosing the underlying cost function d as the Itakura-Saito divergence :

$$d_{IS}(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1 \tag{B.15}$$

The direct usage of formulation (B.12) in the model (B.13) is possible, but leads to computational issues. An equivalent model is obtained by assuming :

$$\mathbf{x}_n = \sum_{k=1}^{K}\sum_{m=1}^{M} \mathbf{d}_{kmn} \tag{B.16}$$

with

$$\begin{aligned} \mathbf{x}_n &\in \mathbb{C}^F \\ \mathbf{d}_{kmn} &\sim \mathcal{N}(0, h_{kn}e_{mk}\mathrm{diag}(\mathbf{P}_{km})) \\ \mathbf{P}_{km} &= [P_{km}(1)...P_{km}(F)]^T \end{aligned}$$

Assuming the equality $\mathbf{c}_{kn} = \sum_m \mathbf{d}_{kmn}$ and the independence of $\mathbf{d}_{kmn}$, we can verify that $\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\sum_m e_{mk}\mathrm{diag}(\mathbf{P}_{km}))$.

We choose here to use the Markov chain prior structure proposed in [Févotte *et al.*, 2009] :

$$p(h_k) = p(h_{k1})\prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)}) \tag{B.17}$$

where $p(h_{kn}|h_{k(n-1)})$ reaches its maximum at $h_{k(n-1)}$, thus favoring a slow variation of h_k in time. We proposed for instance the following choice :

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha_k, (\alpha_k + 1)h_{k(n-1)}) \tag{B.18}$$

where $\mathcal{IG}(u|\alpha, \beta)$ is the inverse-Gamma distribution with mode $\beta/(\alpha + 1)$ and the initial distribution $p(h_{k1})$ is Jeffrey's non-informative prior. Parameters α_k are here arbitrarily fixed, depending on the desired degree of smoothness (the higher α_k, the smoother h_k), but we could consider in future work the possibility to learn it as well. We do not put here any prior on $\mathbf{E}$.

From this model, an EM-like algorithm (HS-NMF/EM) is derived, as well as a multiplicative counterpart of it (HS-NMF/MU). Convergence properties, consistence of the algorithm and their sensitivity to a misestimation of certain parameters are experimentally studied in **chapter VIII**. We also propose an algorithm which produces a partially harmonic dictionary, including free (unconstrained) components (PHS-NMF/MU).

Part Four : *Application to music transcription*

Database and protocol

To evaluate and quantify transcription performance, we need a set of polyphonic music pieces with accurate MIDI references. The two most simple ways to get such data are either to record a MIDI instrument (the acquisition of audio and MIDI being simultaneous), or to synthetize sound from given MIDI files. For the sake of timbre realism and ease of acquisition, the piano is an instrument of choice : very high quality software synthetizers are available on sale, and an acoustic piano can be equipped to play mechanically, and produce a MIDI output, while retaining the timbre of a real instrument. In his thesis [Emiya, 2008], Valentin Emiya collected such a database. *MAPS* (MIDI-Aligned Piano Sounds) includes isolated notes, random and tonal chords, pieces from the piano repertoire, recordings on an upright DisKlavier and high quality software synthesis. From this very complete database, we excerpted two subsets to evaluate our algorithms : a synthetic subset, produced by Native Instruments' Akoustik Piano ("Bechstein Bach" preset, from samples recorded on a Bechstein D280 piano), and a real audio subset, recorded at Télécom ParisTech on a Yamaha Mark III (upright DisKlavier). Each subset is composed of 30 pieces of 30 seconds each (original pieces from *MAPS* were truncated).

NMF-based transcription systems

All NMF-based transcription systems used here follow the same workflow :

1. Computation of an adapted time-frequency representation of the signal, $\mathbf{V}$;
2. Factorization $\mathbf{V} \approx \mathbf{WH}$;
3. Attribution of a MIDI pitch to each basis spectrum $\mathbf{w}_k$ (either from original labelling of columns, in harmonically-constrained cases, or by performing a single-pitch estimation) ;
4. Onset/offset detection applied to each time envelope h_k.

In [Vincent *et al.*, 2007], it is observed that using a nonlinear frequency scale resulted in a representation of smaller size, with better temporal resolution in the higher frequency range, than usual Short-Time Fourier Transform (STFT), while preserving the subsequent transcription performance. We then pass the signal through a filterbank of 257 sinusoidally modulated Hanning windows with frequencies linearly spaced between 5 Hz and 10.8 kHz on the Equivalent Rectangular Bandwidth (ERB)

scale. We then split each subband into disjoint 23 ms time frames and compute the power within each frame.

Pitch estimation of basis spectra is superfluous in harmonically constrained NMF, since each basis component can be labelled from the beginning with the pitch of the patterns $\mathbf{P}_{km}$ used to initialize it. For non harmonically constrained NMF, pitch identification is performed on each column of $\mathbf{W}$ by the harmonic comb-based technique used in [Vincent *et al.*, 2008].

Note onsets and offsets are determined by a simple threshold-based detection, followed by a minimum-duration pruning, see [Vincent *et al.*, 2008]. The detection threshold is denoted by A_{dB} and expressed in dB under $\mathbf{H}$ maximum.

Results and comments

Table B.3 gives an excerpt of the results we obtained during this thesis. Full results and comments are available in **chapter X**.

Algorithm	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$	A_{dB}	Reference
Marolt'04	83.5	70.1	**75.8**	53.5	-	[Marolt, 2004]
Virtanen'07	55.9	56.4	53.6	52.1	-22	[Virtanen, 2007]
Emiya'08	77.3	61.6	**67.7**	67.0	-	[Emiya, 2008]
IS-NMF/MU	63.4	56.1	54.9	51.2	-62	[Févotte *et al.*, 2009]
$(2 \rightarrow 0)$-NMF	59.6	60.6	55.3	53.9	-57	[Bertin *et al.*, 2009b]
$(10 \rightarrow 0)$-NMF	62.3	51.3	51.4	52.7	-59	[Bertin *et al.*, 2009b]
H-NMF/MU	58.7	59.1	52.4	46.0	-33	[Bertin *et al.*, 2010]
HEUC-NMF/MU	60.7	60.0	58.4	54.8	-32	[Vincent *et al.*, 2008]
S-NMF/EM	62.4	43.3	49.5	50.7	-51	[Févotte *et al.*, 2009]
HS-NMF/EM	65.8	64.5	60.7	44.3	-38	[Bertin *et al.*, 2010]
HS-NMF/MU	78.5	62.6	67.0	46.4	-42	[Bertin *et al.*, 2009a]
PHS-NMF/MU	77.6	65.4	**68.4**	45.0	-43	-

TABLE B.3 – Average performance of tested algorithms on a synthetic piano database.

Among all NMF-based algorithms, algorithms with both harmonicity and smoothness contraints (HS-NMF/MU, HS-NMF/EM et PHS-NMF/MU) give the best performances. Results are comparable to Emiya'08, which places our approach at the state-of-the-art. They however remain lower than results from Marolt'04, a very performant system, especially tuned for piano music and which uses additionnal training databases. Algorithms with tempering ($(2 \rightarrow 0)$-NMF and $(10 \rightarrow 0)$-NMF) produce results comparable to the baseline IS-NMF/MU without tempering. Smoothness constraint seems detrimental to the transcription when used alone, being implemented either in a multiplicative (Virtanen'07) or EM-like (S-NMF) algorithm. Harmonicity constraint improves slightly the scores, but the best results are obtained when both are used together.

Conclusion

NMF-based methods remain here less performant than other finely tuned state-of-the-art methods, especially methods implying a training phase, the use of learning data and musicologically inspired

post-processing. However, NMF is totally data-driven, it requires no training and then adapts itself to the data while avoiding the risk of a mismatch between training and test data. It also provides a semantically meaningful mid-level representation of the data. Its potential here assessed is clear, letting the hope of very good performance with better tuning and improvements.

The temporal smoothness constraint does not bring all improvements we could expect, in particular in terms of robustness to the detection threshold and efficiency of the note duration estimation. However, it seems useful to compensate the tendency of NMF with harmonicity constraint to produce non-smooth decomposition, and lead therefore to a better transcription performance when both constraints are used. A limitation of our common NMF framework (NMF core algorithm plus detection threshold based post-processing) appears here, as a 100% recall rate is never reached, for any value of the threshold or any tested algorithm.

Using a statistical model relies of course on the fact that the ground truth actually follows this model. Performance obtained here let hope it is more or less the case, but adequation between the data and the model should be further investigated on. In particular, the choice of the shape parameter α of the inverse-Gamma prior put on temporal envelopes should be discussed, and its learning, as well as NMF factors are learnt, should be considered.

Possible improvements include a refinement of the temporal prior, which suits for modelling the sustain and decay parts of the note, but disfavour attacks and silences. An option to alleviate this mismatch between the model and the data could be the use of switching state models for the rows of $\mathbf{H}$, that would explicitly model the possibility for h_{kn} to vary quickly (attack) or to be strictly zero (absence of the note). Moreover, certain quantities that are fixed here should be trained or learnt, which could be another direction of improvement.

Glossaire musical

Ambitus Intervalle entre la note la plus basse et la note la plus haute d'une partition ou d'un morceau de musique.

Armure Ensemble d'altérations (dièses ou bémols) réunies à la clé, et s'appliquant à toutes les notes du morceau sans que l'altération soit rappelée à chacune des occurrences des notes concernées.

Consonant En harmonie classique, intervalle ou accord produisant une impression de stabilité, de détente et d'accomplissement.

Demi-ton Plus petit intervalle conjoint de la gamme diatonique, séparant par exemple le mi du fa. Les notes composant un demi-ton ont un rapport de fréquences de $2^{1/12}$ (dans la gamme tempérée).

Diapason Fréquence du la_4, qui fixe les autres notes. Par extension, objet produisant une note à cette fréquence.

Diatonique Se dit d'une échelle (gamme) musicale composée de 7 notes, réparties en 5 tons et 2 demi-tons.

Dissonant En harmonie classique, discordance d'un ensemble de sons produisant une impression d'instabilité et de tension, et nécessitant une résolution.

Dominante Cinquième degré de la gamme, formant un intervalle de quinte avec la tonique. Exemple : note *sol* dans la gamme de *do* Majeur.

Enharmonie Deux notes appelées différemment mais dont on considère qu'elles possèdent la même hauteur (ex : *do*♯ et *ré*♭).

Glissando Glissement d'une note à une autre, pendant lequel la fréquence fondamentale varie continument.

Harmonique Partiel dont la fréquence est multiple d'une fréquence fondamentale. Par extension, son constitué d'une fréquence fondamentale et d'un ensemble de partiels harmoniques.

Homorythmie Musique au cours de laquelle aucune note ne démarre tant que la précédente n'est pas terminée et toutes les attaques des notes sont simultanées.

Inharmonicité Propriété d'un son musical dont les composantes fréquentielles, que l'on appelle partiels ne sont pas strictement harmoniques, c'est-à-dire ne sont pas exactement à des fréquences multiples entiers du son fondamental.

Legato Phrasé musical caractérisé par des notes liées, sans silence entre elles.

Métrique Couple de nombres indiquant le type et la durée des temps dans chaque mesure de la partition, et déterminant le placement des barres de mesure.

Modulation Changement de tonalité en cours de morceau.

Monodique Qui ne contient qu'une note à la fois.

Partiel Composante du spectre d'un son, localisée en fréquence.

Polyphonie Combinaison de plusieurs voix indépendantes. Par extension, capacité d'un instrument de jouer plusieurs notes à la fois.

Portée En notation musicale traditionnelle, ensemble de cinq lignes horizontales parallèles permettant de représenter les figures de notes, de silence et diverses informations comme l'armure ou la clef.

Résolution Mouvement d'un accord dissonant vers un accord consonant, apaisant la tension musicale.

Sensible Septième degré de la gamme. Exemple : note *si* dans la gamme de *do* Majeur. Dans la théorie classique, cette note est souvent suivie de la tonique.

Staccato Phrasé dans lequel les notes doivent être exécutées avec des courtes suspensions entre elles.

Tablature Forme simplifiée de notation musicale dans laquelle les accords ne sont pas transcrits *in extenso* mais sous la forme d'un chiffrage décrivant la fondamentale et le type d'accord (majeur, mineur, septième...)

Tempo Vitesse d'exécution d'une œuvre, exprimée en noires par minute.

Tessiture Ensemble des notes qu'un chanteur ou un instrument peut produire, du grave à l'aigu. Synonyme : registre.

Ton Plus grand intervalle conjoint de la gamme diatonique, par exemple entre fa et sol. Les notes composant un ton ont un rapport de fréquences de $2^{2/12}$ (dans la gamme tempérée).

Tonalité Gamme diatonique dans lequel les différentes notes (appelées « degrés ») ont un usage et une sémantique différenciés, consacrés par la théorie classique. Exemple : dans la tonalité dite de « do Majeur », les notes *do* (tonique), fa (sous-dominante) et *sol* sont considérées comme des notes affirmant la sensation de tonalité, reposantes et conclusives, le *mi* (médiante) comme une note faisant ressentir la modalité (mineure ou majeure).

Tonique Premier degré de la gamme, qui donne son nom à la tonalité.

Tremolo Modulation périodique d'intensité d'une note.

Unisson Une même note jouée simultanément par plusieurs instruments.

Valeur Durée d'une note, quantifiée en fonction du tempo et de la mesure. Exemple : une blanche, une noire, une croche.

Vibrato Variation oscillante continue de la hauteur d'une note pendant tout ou partie de sa durée.

Crédits

Une thèse ne se prépare pas sans ressources et outils. Elle nécessiterait de longues années supplémentaires sans l'apport de nombreux contributeurs qui sans le savoir, lui évitent de réinventer l'eau chaude à chaque page. Nous tenons à porter au crédit de ces personnes l'important travail réalisé pour le bienfait de tous, dans une démarche humaniste, scientifique, altruiste. De licence GPL et Creative Commons en freeware et boîtes à outils simplement offertes, ces outils libres et gratuits facilitent au quotidien le travail du scientifique.

Outil rédactionnel incontournable dans le monde scientifique, **LaTeX** propulse ce mémoire. J'ai utilisé la distribution **MiKTeX** pour Windows et l'environnement de développement **TeXnicCenter**, à l'ergonomie précieuse.

En ce qui concerne l'obtention, la visualisation, la représentation de sons musicaux, il me faut citer : le logiciel de visualisation, édition et traitement de sons **Audacity**, le projet **Mutopia** qui offre une grande quantité de pièces de musique sous divers formats et notamment sous forme de partition, l'éditeur de partitions **Lilypond**, le site de sources musicales aux formats MIDI et .mp3 `http://piano-midi.de`.

La production de jolies figures au bon format n'est pas toujours une sinécure. Nombre d'images présentées dans ce mémoire ont été produites, converties ou retravaillées avec les outils : **ghostview**, **The Gimp**, **CoolPDFReader**, **WinFIG**.

Enfin, les simulations informatiques de cette thèse, outre bien sûr l'inévitable ®Matlab, ont été réalisées à l'aide d'un certain nombre de boîtes à outils mises à disposition par leurs auteurs : la boîte-à-outils « **MATLAB and MIDI** » de Ken Schutte, le **nmfpack** de Patrik Hoyer, la **TFTB** (*Time-Frequency ToolBox*) du GdR Isis et enfin **MPTK** (*Matching Pursuit ToolKit*), distribué par l'INRIA.

Si le logiciel libre est une fort belle chose, il n'en reste pas moins que le doctorant ne se nourrit pas exclusivement de science et d'eau fraîche. Cette recherche a été menée grâce au support financier de différentes institutions et projets que je remercie de leur soutien : le Ministère de l'Éducation Nationale, de l'Enseignement Supérieur et de la Recherche ; la Commission Européenne, via les contrats FP6-027026 (*Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content KSPACE*) ; le GIP ANR sous le contrat ANR-06-JCJC-0027-01 (Décomposition en Éléments Sonores et Applications Musicales DESAM) ; l'agence OSEO, via le projet Quaero.

Index

Les factorisations en matrices non-négatives.
Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.

La transcription automatique de la musique est l'opération qui consiste, partant du seul contenu audio, à produire une représentation symbolique (par exemple un fichier MIDI) d'un morceau de musique. Si la transcription de musique monodique est aujourd'hui bien maîtrisée, le cas de la musique polyphonique est en revanche un problème largement ouvert.

Pour s'atteler à cette tâche, certaines approches sont basées sur des connaissances a priori tels que des modèles de signaux, ou des bases de données d'apprentissage. La contrepartie de ce type de méthodes est leur faible capacité d'adaptation à des signaux s'éloignant trop du modèle ou des données initiales. Pour s'affranchir de cette contrainte, une nouvelle famille d'approches consiste à introduire le moins d'a priori possible sur l'audio de départ, et à tenter de séparer les notes jouées « à l'aveugle ». Parmi celles-ci, les représentations parcimonieuses, les techniques de factorisation en matrices non-négatives ou de séparation de sources font des hypothèses faibles et réduites sur les signaux. Elles ont montré des résultats prometteurs en transcription de musique polyphonique. L'objectif de cette thèse est d'obtenir une représentation orientée-objet du signal, exhibant clairement les structures sémantiques qui le composent, et qui serait un intermédiaire entre le signal et une représentation plus haut niveau. Une telle représentation aura l'avantage de simplifier les tâches d'indexation et de transcription automatique de musique. Pour extraire ces structures du signal, l'approche que nous visons repose sur des techniques d'analyse matricielle.

Les décompositions de matrices en valeurs propres et en valeurs singulières sont des techniques classiques d'algèbre linéaire utilisées dans un grand nombre d'applications de traitement du signal. Elles permettent de représenter efficacement les données observées en utilisant un nombre limité d'atomes élémentaires. Contrairement à d'autres techniques de représentations du signal, ces atomes ne sont pas recherchés au sein d'un dictionnaire pré-défini, mais sont extraits des données elles-mêmes. La factorisation en matrices non-négatives (NMF) est une technique analogue d'algèbre linéaire, qui réduit le rang tout en fournissant des atomes à valeurs exclusivement positives, donc plus facilement interprétables, et sémantiquement plus pertinentes pour la représentation de données elles-mêmes à valeurs positives. Alors que d'autres travaux se concentrent soit sur la mise au point de dictionnaires, soit sur la décomposition de signaux sur ces dictionnaires, la NMF fournit conjointement un dictionnaire extrait des données et la décomposition de ces mêmes données dans ce dictionnaire.

Ce mémoire est consacré à l'étude théorique et expérimentale détaillée de ces méthodes. Il poursuit plusieurs objectifs : l'amélioration des performances des systèmes de transcription qui les utilisent, de la pertinence sémantique des représentations mi-niveau produites, et du contrôle des propriétés théoriques et pratiques des algorithmes existants et originaux mis en œuvre au cours de la thèse.