THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE - GET-Télécom Paris

Spécialité :
Acoustique, Traitement du Signal, Informatique Appliqués à la Musique

Présentée par M. **Pierre Leveau**

Pour obtenir le grade de
Docteur de l'École Doctorale d'Informatique, de Télécommunications et d'Électronique de Paris

Sujet :

DÉCOMPOSITIONS PARCIMONIEUSES STRUCTURÉES : APPLICATION À LA REPRÉSENTATION OBJET DE LA MUSIQUE

MODÈLES DE SIGNAUX, ALGORITHMES ET APPLICATIONS

soutenue le 13 Novembre 2007

devant le jury composé de :

Pr. Philippe DEPALLE	Rapporteurs
Pr. Bruno TORRESANI	
Pr. Jean-Luc ZARADER	Examinateurs
Dr. Emmanuel VINCENT	
Dr. Jana EGGINK	
Pr. Gaël RICHARD	Directeurs de thèse
Dr. Laurent DAUDET	

Résumé

La quantité de musique numérisée disponible à la fois sur Internet et chez chaque utilisateur particulier a explosé depuis maintenant une dizaine d'années. Or, l'organisation et l'accessibilité de cette masse de données exigent que certaines informations soient disponibles, comme par exemple l'artiste, le nom de l'album, de la chanson, le style, le tempo, l'humeur ou d'autres attributs symboliques ou sémantiques.

Ainsi, l'indexation automatique de la musique est un domaine de recherche qui suscite un grand intérêt actuellement car il permet d'envisager une obtention automatique de ces annotations. Si certaines tâches sont maintenant traitées correctement pour certains types de musique, comme la classification par genre sur des musiques stéréotypiques, la reconnaissance d'instruments jouant en solo et l'extraction de tempo, d'autres données sont plus difficiles à extraire. Par exemple, la transcription automatique de signaux polyphoniques et la reconnaissance d'ensembles d'instruments sont encore limités à quelques cas particuliers.

Le but de l'étude n'est pas d'obtenir une transcription parfaite des signaux et une classification exacte de tous les instruments mis en jeu, mais plutôt de construire une représentation objet du signal, c'est-à-dire une représentation qui met en valeur certaines caractéristiques utiles des signaux en le représentant sous la forme d'objets sonores.

Afin de réaliser cette tâche, nous nous intéressons au domaine des représentations parcimonieuses de signaux. Ce domaine d'étude relativement récent traite de l'approximation des signaux par des formes d'ondes (*atomes*) appartenant à des dictionnaires. Les principaux sujets d'études sont la construction de dictionnaires adaptés aux signaux analysés, ainsi que la recherche d'algorithmes permettant de décomposer le signal de façon optimale et efficace.

Dans le travail présenté, des dictionnaires liés à des sources instrumentales ont été construits : nous définissons un atome Harmonique Spécifique à un Instrument comme une somme d'atomes de Gabor représentant les partiels d'une note, et dont les vecteurs d'amplitudes respectives appartiennent à un ensemble appris au préalable sur des sources annotées. Des variantes permettant de mieux modéliser les structures sortant du cadre strictement harmonique sont proposées : l'une permet de tenir compte des notes présentant des modulations de fréquence dans leurs partiels, et l'autre introduit un paramètre d'inharmonicité qui modélise la position des partiels pour les instruments légèrement inharmoniques comme le piano. Ces atomes peuvent être définis en stéréo grâce à l'ajout d'un paramètre de panoramique. Nous présentons également des *molécules*, groupements d'atomes permettant de modéliser les structures longues, comme les notes de musique entières.

Dans un second temps, nous présentons des algorithmes permettant d'extraire ces

atomes et molécules de façon efficace sur des signaux audio. Nous utilisons l'algorithme de Matching Pursuit, que nous adaptons afin d'extraire les structures définies précédemment. Les algorithmes permettant d'extraire les atomes font intervenir une optimisation des paramètres après une estimation grossière sur une grille. Les algorithmes moléculaires mettent en jeu des recherches de chemins, résolues par programmation dynamique.

Enfin, nous montrons en quoi les modèles de signaux développés ainsi que les algorithmes permettent d'obtenir des représentations utiles pour l'indexation de la musique. Nous évaluons leur efficacité pour l'estimation de hauteur de note et la classification d'instruments de musique sur des solos, pour lesquels les résultats sont à la hauteur d'algorithmes de l'état de l'art. Le problème de l'identification d'ensembles d'instruments est également abordé en mono et en stéréo. Un codeur à extrêmement bas débit (1 à 4 kbs) est aussi implémenté, avec des résultats préliminaires encourageants.

Abstract

The amount of digital music available both on the Internet and by each listener has considerably raised for about ten years. The organization and the accessibillity of this amount of data demand that additional informations are available, such as artist, album and song names, musical genre, tempo, mood or other symbolic or semantic attributes.

Automatic music indexing has thus become a challenging research area. If some tasks are now correctly handled for certain types of music, such as automatic genre classification for stereotypical music, music instrument recoginition on solo performance and tempo extraction, others are more difficult to perform. For example, automatic transcription of polyphonic signals and instrument ensemble recognition are still limited to some particular cases.

The goal of our study is not to obain a perfect transcription of the signals and an exact classification of all the instruments involved, but rather to build an object representation of the signal, that exhibits some useful features of the music signals by representing it as sound objects.

To achieve this goal, we will employ sparse representations of the signal. This recent research area handles the approximation of signals by waveforms (*atoms*) belonging to dictionaries. The main topics are the building of dictionaries that are adapted to the analyzed signals, and the design of algorithms allowing to decompose the signal in an optimal and efficient way.

In the presented work, dictionaries linked to instrumental sources have been built : we define a Instrument-Specific Harmonic atom as a sum of Gabor atoms representing the note partials, and whose amplitude vectors belong to an ensemble learnt of annotated sources. Some variants of these atoms have been defined to better model the structures outside the strict harmonicity : one takes the frequency modulations into account, another one introduces an inharmonicity parameter that models the partial positions for the slightly inharmonic instruments like piano. These atoms can be defined in stereo signals with an additionnal panpot parameter. We also introduce *molecules*, atom subsets that models long structures like entire music notes.

Then, we present algorithms that extract these atoms and molecules in an efficient way from audio signals. We will use the Matching Pursuit algorithm, that we adapt for the extraction of the aforementioned signal structures. The algorithms used to extract the atoms involve an optimization of their parameters after a coarse estimate on a grid. The molecular algorithms are based on path search, resolved with dynamic programming.

Finally, we show how the signal models and the developed algorithms yield useful representations for music indexing. We evaluate their efficiency for pitch estimation and music instrument recognition on solo phrases, for which results are as high as state-of-

the-art algorithms. The identification of instruments ensembles has also been addressed in monophonic and stereophonic signals. A extremely low rate coder (1 à 4 kbs) has also been implemented, with encouraging preliminary results.

Remerciements

Tout d'abord, je remercie les membres du jury d'avoir accepté de juger ma thèse, et en particulier Philippe Depalle et Bruno Torrésani d'avoir évalué mon document de thèse.

Je tiens à remercier particulièrement mes deux directeurs de thèse, Laurent Daudet et Gaël Richard. Ils m'ont en effet énormément apporté dans cette longue aventure qu'est la thèse. Tout d'abord, leurs compétences et leur culture scientifique m'ont permi d'aborder mon travail en toute sérénité : en cas de doute, je savais que je pouvais benéficier de leurs bons conseils. Leur bonne humeur constante a également rendu toutes nos intéractions très agréables, et a contribué au déroulement agréable de ces trois années. J'aimerais aussi les remercier de m'avoir laissé une certaine liberté en ce qui concerne l'orientation de mon sujet. Enfin, je leur suis très reconnaissant de m'avoir permis de rencontrer d'autres chercheurs extrêmement intéressants et compétents. Alors que je m'attendais à un travail relativement solitaire de trois ans, cette période a été, grâce à eux, riche en rencontres et en collaborations. Tout cela m'a appris qu'il est souvent plus efficace de discuter quelques minutes avec un autre chercheur que de travailler seul pendant plusieurs semaines.

Parmi les chercheurs avec qui j'ai pu travailler, je tiens particulièrement à remercier Slim Essid. Il m'a initié aux techniques de classification automatique dès la fin de mon stage de DEA, et les discussions que j'ai pu avoir avec lui sur la reconnaissance des instruments et l'indexation multimedia en général ont été très enrichissantes. Et surtout, j'ai pu réutiliser la base de données de sons d'instruments qu'il a patiemment construit pendant son travail de thèse, ce qui m'a été d'un grand secours et m'a apporté un gain de temps considérable. Je remercie également Rémi Gribonval qui a eu la gentillesse de me consacrer du temps alors que je débutais dans le domaine des représentations parcimonieuses et que mon *Matching Pursuit* était fort impur. Grâce à lui, même si mes algorithmes ne sont pas encore transcendantaux, j'ai progressé sur la voie de la pureté, et je lui en suis très reconnaissant. J'ai ensuite pu collaborer avec Emmnanuel Vincent, que je remercie de faire partie de mon jury. Les discussions que nous avons eues ont contribué de façon majeure à l'élaboration des algorithmes présentés dans ce document. Sa précision et sa rigueur ont été extrêmement précieuses pour la rédaction des articles qui ont été issus de notre collaboration. Je remercie également les autres personnes avec qui j'ai eu le plaisir de travailler, notamment Sacha Krstulovic̀ à l'Irisa, les stagiaires Hadrien Cousin, Grégory Cornuz et Adrien Daniel, David Sodoyer, post-doctorant à l'Institut Jean Le Rond D'Alembert, Emmanuel Ravelli, doctorant dans le même laboratoire, Nancy Bertin et Valentin Emiya à Télécom Paris. Je me dois aussi d'adresser un remerciement spécial à tous les autres stagiaires, doctorants, personnels qui ont contribué aux ambiances très agréables dans mes lieux de travail et parfois en dehors.

La thèse s'est déroulée dans deux endroits exceptionnels, à l'Institut Jean Le Rond d'Alembert (équipe LAM) et au département TSI de Télécom Paris. Ces deux environnements de travail ont énormément contribué au travail que j'ai effectué, j'en remercie donc

les directeurs Jean-Dominique Polack et Yves Grenier de m'y avoir accueilli.

Je remercie enfin mes parents pour leur soutien sans faille.

Table des matières

Table des figures

Liste des tableaux

.

Chapitre 1

Introduction

1.1 Contexte technologique

1.1.1 Multimedia

La quantité de données numériques stockées dans le monde est en constante augmentation, du fait de la production abondante de contenu numérique, mais aussi de la conversion en numérique des contenus analogiques déjà existants. Pouvoir accéder rapidement à un document de cette masse considérable de données pose un grand nombre de problèmes techniques et théoriques. Une condition préalable pour répondre à cette exigence est que les programmes effectuant les recherches "comprennent" les documents. Il faut alors leur ajouter des données annexes, appelées *métadonnées*. Ces métadonnées peuvent être de nature très variée : elle peuvent aussi bien décrire le contexte du document (données de production) que le contenu lui-même du document en fournissant des informations linguistiques, symboliques ou numériques. Elles peuvent être produites essentiellement de deux manières : soit elles sont annotées par des utilisateurs au moment de la production, de la diffusion, de l'utilisation ou de l'archivage, soit elles sont extraites automatiquement par des algorithmes d'*indexation automatique*.

Si l'on s'intéresse à l'indexation automatique des documents multimedia, les techniques sont encore émergentes. Par exemple, les moteurs de recherche les plus populaires actuellement recherchent dans des données déjà annotées sémantiquement, à l'exception des textes où ce problème est contourné en exploitant entre autres la fréquence des mots qu'ils contiennent.

1.1.2 Le cas de la musique

En particulier, les supports de la musique se sont largement numérisés. Si le disque compact a déjà consisté en une numérisation des enregistrements (les signaux sont numérisés avec une fréquence d'échantillonnage de 44100 Hz et une quantification sur 16 bits avant gravage), l'avènement du format de compression MPEG 1 Layer III (dit MP3) et de ses successeurs a permis le stockage et la mise en réseau d'une grande partie des données musicales, ainsi qu'une accessibilité quasi-instantanée au contenu pourvu qu'on dispose de son adresse.

Actuellement, la gestion de ces données musicales est principalement faite à partir de données de production comme l'interprète, le titre de la chanson, de l'album... Cependant, l'extraction d'autres informations est désirable : la connaissance et la synchro-

nisation des tempos et de l'harmonie permettraient d'effectuer des mixages automatisés, l'émotion véhiculée des chansons (*mood*) de générer des listes de lecture adaptée à l'humeur de l'auditeur, les similarités et dissimilarités entre chansons de proposer à l'auditeur des musiques en conformité avec ses goûts...

De nombreuses entreprises ou projets s'intéressent à l'organisation de l'information musicale. Parmi les projets les plus aboutis, on peut notamment citer MusicBrainz (identification des morceaux par signature spectrale), Last.fm (similarités entre chansons calculées à partir des habitudes d'écoutes et d'annotations linguistiques manuelles), Pandora (similarités calculées à partir de caractéristiques sémantiques annotées par des experts)...

La gestion de ces informations intéresse également une communauté de chercheurs depuis un certain nombre d'années, qui se définit comme la communauté de *Music Information Retrieval*. Elle aborde des thèmes pluridisciplinaires comme l'extraction automatique d'information sémantique, la classification, les comportements des utilisateurs et la gestion des librairies musicales. Les domaines de connaissances mis en jeu sont relativement variés : la musicologie, le traitement du signal, l'intelligence artificielle (reconnaissance de forme, modélisation statistique), les sciences de l'information (organisation des méta-données), etc.

1.2 Contexte scientifique

Des avancées récentes dans différents domaines scientifiques permettent maintenant d'aborder le sujet de l'extraction de métadonnées linguistiques ou symboliques à partir de données audionumériques brutes. Pour réaliser une indexation efficace des données multimédia, il faut s'intéresser à de nombreux domaines de connaissance. On a besoin en effet de connaître les paramètres symboliques, linguistiques ou numériques pertinents de description du contenu pour son traitement automatique, les méthodes permettant d'extraire de tels paramètres à partir de mesures sur les données du contenu, l'influence des conditions de création du contenu sur les données (par exemple, pour la musique, avoir une bonne connaissance des sons produits par les instruments) etc...

Parmi ces nouvelles connaissances, on peut mentionner tout d'abord les avancées en sciences cognitives concernant la perception des sons. En effet, les processus mis en jeu entre le signal acoustique arrivant à l'oreille et la représentation sémantique qu'il induit chez l'auditeur sont de mieux en mieux compris. La psychoacoustique permet de modéliser les premières étapes transformant le son en signal interprété par le cerveau, et ainsi d'en connaître les principales composantes. Les neurosciences et la psycholinguistique permettent également de faire des hypothèses sur les processus mis en jeu au niveau des intéractions entre la perception du signal et les représentations sémantiques de l'humain. On parle alors de processus *bottom-up* (signal vers sens) et *top-down* (sens vers signal).

Des progrès sont également faits en intelligence artificielle, domaine qui vise à fournir aux machines des aptitudes similaires à celles des humains. Il s'agit dans ce domaine de définir des algorithmes permettant aux machines de réaliser des tâches effectuées correctement par des humains à l'aide de processus cognitifs. Ainsi, de nombreux algorithmes permettent maintenant de réaliser des tâches comme la classification automatique de phénomènes visuels ou sonores, dans des contextes relativement fermés. On peut par exemple mentionner la reconnaissance automatique de la parole.

Enfin, on peut mentionner les progrès en optimisation numérique et traitement du signal. Il est maintenant commun de faire intervenir des modèles statistiques des phénomènes observés et des connaissances afin d'extraire des informations à partir de mesures physiques.

1.3 Typologie succinte des tâches en indexation de signal musical

L'indexation du signal musical audio vise à réaliser les types de tâches suivants à partir des signaux audio :

- **La classification** : ce type de tâche consiste à attribuer une ou plusieurs classes, linguistiques ou symboliques, au signal analysé. On peut chercher à identifier le genre musical, les instruments mis en jeu, la tonalité, l'artiste exécutant la pièce, etc.
- **La segmentation** : cette tâche consiste à découper le signal musical en segments temporels pertinents. Les segments intéressants à extraire sont par exemple les couplets et les refrains, la parties chantées ou non chantées, les mesures, les notes d'une performance solo etc.
- **L'estimation de paramètres numériques** : l'estimation du tempo est un exemple de ce type de tâche.
- **La génération de représentations symboliques** : il s'agit ici d'obtenir une représentation du son en éléments distincts. On peut mentionner par exemple la transcription automatique de musique, qui *grosso modo* vise à obtenir une représentation symbolique de type MIDI[1] du son.
- **La mise en relation** : la similarité timbrale et la similarité mélodique sont des exemples de relations qui présentent des intérêts applicatifs reconnus.

1.4 Problématique

Les méthodes employées jusqu'à présent pour extraire de l'information à partir des signaux audio font généralement intervenir deux étapes. La première consiste à produire une nouvelle *représentation* du signal. Chaque élément de cette représentation possède des propriétés plus pertinentes pour des tâches données que le signal original. Plus formellement, les propriétés d'un élément sont représentées par une vecteur de paramètres. Parmi ces représentations, on peut citer par exemple les représentations temps-fréquence, dans lesquelles chaque élément porte une localisation temporelle, une fréquence et une intensité, ou alors plus généralement les représentations par des caractéristiques (*features*) calculées sur des portions de signal. La seconde étape fait intervenir le traitement aboutissant à l'extraction de l'information de haut-niveau. Par exemple, si l'on considère une tâche de classification, il s'agit au préalable de modéliser les classes à identifier dans l'espace des paramètres de la représentation. Puis, une fois le signal inconnu représenté dans le nouvel espace, des algorithmes sont mis en oeuvre afin d'effectuer l'indexation : comparaison à des modèles pour la classification, détection de changement pour la segmentation, comparaisons de représentations entre elles pour la similarité etc.

[1] *Musical Instrument Digital Interface*

Si les approches par calculs de caractéristiques sont utilisées dans la majeure partie des applications actuelles, leurs limitations apparaissent également : elles ne permettent pas de rendre compte de la nature additive de la musique. En effet, celle-ci est en général le résultat d'un mélange de sons produits au même moment. Des méthodes de séparation de sources efficaces permettraient de contourner cet obstacle, mais leurs résultats ne sont pas encore satisfaisants, où alors au prix d'une complexité calculatoire élevée. On peut alors réfléchir à une méthode intermédiaire, consistant à ne pas séparer les sources parfaitement mais néanmoins à produire une représentation où certaines des caractéristiques des sources mises en jeu peuvent être extraites.

Dans ce document, nous détaillerons une nouvelle approche permettant de représenter le signal musical en introduisant des connaissances sur les sources instrumentales qui peuvent le composer. Nous verrons ainsi en quoi de récentes avancées en traitement du signal, notamment dans le domaine des représentations parcimonieuses, permettent d'obtenir des représentations utiles pour les tâches d'indexation audio (reconnaissance des instruments, transcription, analyse harmonique). Les éléments de ces représentations pourront être assimilés à des objets sonores. Le cadre de l'étude sera celui des instruments produisant des sons harmoniques ou quasi-harmoniques, mais la méthodologie et l'outil théorique permettraient d'aborder des cadres plus génériques. Nous nous appuierons sur l'état de l'art de l'extraction d'informations symboliques ou sémantiques du signal audio, et présenterons les forces et faiblesses des techniques utilisées.

Chapitre 2

État de l'art

Dans cette partie, nous nous intéresserons aux différentes méthodes qui peuvent être mises en oeuvre afin d'extraire des données de haut-niveau à partir de signaux audio de musique. Nous verrons quelles sont les représentations intermédiaires du signal audio qui peuvent être utlisées pour réaliser ce type de tâche, ainsi que les différentes modélisations utilisées pour l'extraction des informations de haut-niveau proprement dites.

2.1 Algorithmes pour l'extraction d'informations haut-niveau en musique

Nous avons posé les bases de la problématique d'extraction de données haut-niveau (symboliques, sémantiques) à partir de données bas-niveau (signal audio). Les algorithmes qui traitent ces tâches mettent en jeu d'une part divers processus d'analyse du signal, et d'autre part des modélisations permettant, dans une certaine mesure, d'exploiter des connaissances a priori sur le signal analysé. Dans cette section, nous chercherons à mettre en valeur ces processus et leur articulation dans ces algorithmes. Nous parlerons tout d'abord des algorithmes *bag-of-frames* (ou sacs de trames), puis présenterons les algorithmes basés sur des modèles explicites de signaux et/ou de mélange. Enfin, nous évoquerons les représentations mi-niveau.

Nous nous focaliserons essentiellement sur les tâches de classification de sources instrumentales. Nous intéresserons aussi aux problèmes d'estimation de hauteur de note et de séparation de source, étant donné que ces problèmes ne sont pas complètement indépendants les uns des autres.

2.1.1 Les approches par sacs de trames

La classe de méthodes la plus employée en indexation automatique des signaux audio est l'approche par sacs de trames. Celle-ci s'appuie sur les étapes suivantes :

- L'extraction des caractéristiques sur des trames de signaux, et éventuellement, une sélection automatique des caractéristiques pertinentes pour la tâche de classification envisagée,
- La modélisation des classes à discriminer dans l'espace de ces caractéristiques sur des signaux d'entraînement,

– L'opération d'indexation. Dans le cas d'une classification, cette opération consiste à comparer des observations à des modèles précédemment appris.

Cette classe de méthodes a l'avantage de conduire à des solutions techniques relativement simples, en utilisant un des nombreux classifieurs existants (voir (Duda et al., 2000)) après avoir extrait des caractéristiques supposées pertinentes pour la discrimination des classes.

2.1.1.1 Caractéristiques

La première étape dans ce type d'approche est l'extraction des caractéristiques (*features*) sur des trames de signal, généralement de longueur fixe. Formellement, le signal x est représenté à un échantillonnage temporel plus grossier que l'échantillonnage temporel original (typiquement, $T = 30ms$), mais sur plusieurs dimensions, définies par les caractéristiques extraites (voir Figure 2.1).

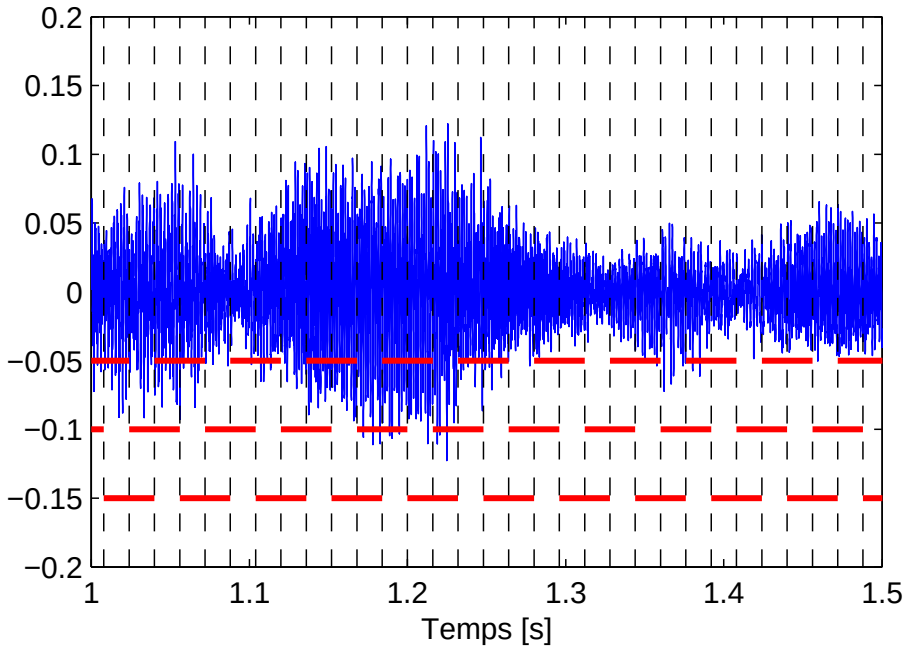


FIG. 2.1 – *Schéma représentant une extraction de caractéristiques classiques sur un signal audio. Les lignes rouges indiquent les trames sur lesquelles les caractéristiques sont calculées.*

Les caractéristiques extraites sont censées présenter des valeurs significatives des classes qu'on cherche à discriminer. Par exemple, dans le domaine du traitement automatique de la parole (reconnaissance de la parole ou du locuteur), les *Mel-Frequency Cepstral Coefficient* (MFCC) sont d'une remarquable efficacité de par leur capacité à décorréler les caractéristiques de la source (onde glottale) qui impose la fréquence fondamentale (*pitch*), et du filtre (cavités de résonnance de l'appareil phonatoire) qui donne des indications sur les phonèmes prononcées. Ces caractéristiques ont été reprises dans de nombreuses analyses concernant l'indexation automatique de signaux musicaux car elles capturent relativement bien l'information timbrale donnée par le spectre (Brown (1999); Eronen & Klapuri (2000); Logan (2000); Tzanetakis & Cook (2002)). Certaines études utilisent uniquement ces caractéristiques et leurs dérivées temporelles pour caractériser les

signaux musicaux, par exemple concernant la similarité musicale (Aucouturier, 2006). D'autres caractéristiques permettant de caractériser les formants, nommées *line spectrum frequencies* (LSF), ont été proposées et on montré de bonnes performances en classification d'instruments (Chétry & Sandler, 2006).

De nombreuses autres caractéristiques ont été développées. Souvent, les caractéristiques supplémentaires sont conçues en utilisant de la connaissance acquise dans d'autres domaines scientifiques. Concernant la reconnaissance automatique des instruments de musique, on notera par exemple que des études en acoustique musicale indiquent que le rapport entre l'énergie des harmoniques paires et impaires permet de distinguer les instruments à vent à résonnateur coniques ou cylindriques. Des études en perception ont montré que la durée d'attaque et la fréquence de coupure sont des caractéristiques pertinentes pour discriminer différents timbres (McAdams et al., 1995). Lorsqu'on aborde la reconnaissance automatique de genre (Tzanetakis & Cook, 2002), d'autres caractéristiques provenant de la musicologie sont notamment utilisées .

Des caractéristiques venant d'autres domaines sont parfois testées, avec peu d'a priori sur leur pouvoir discriminatif mais dont on espère qu'ils apporteront une information supplémentaire pour discriminer : par exemple, en traitement du signal, des caractéristiques sur l'impulsivité du signal peuvent être extraites à partir de la transformée en ondelettes (Leveau, 2004; Li & Ogihara, 2005). On notera également que les approches "bag-of-frames" peuvent aussi être appliquées pour des problèmes de transcription (Ellis & Poliner, 2006), en prenant comme caractéristiques les valeurs de la Transformée de Fourier à Court Terme.

Enfin on peut mentionner une dernière méthode d'élaboration de caractéristiques qui est l'EDS (*Extractor Discovery System*, Zils (2004)). Cette approche permet, étant donnés un critère de discrimination des classes ou un classifieur et des signaux appartenant à ces classes, de calculer les formules de descripteurs qui permettent de maximiser soit la corrélation avec le critère de discrimination, soit le score du classifieur. La méthode met en jeu des algorithmes génétiques qui permettent de parcourir efficacement l'espace des combinaisons possibles de formules mathématiques élémentaires.

Ainsi, nous voyons que l'élaboration des caractéristiques est un point critique pour cette classe de méthodes : elles sont le plus souvent construites afin qu'elles exhibent les similarités et les dissimilarités voulues entre les classes. Dans les étapes suivantes, il n'y a plus de retour au signal : l'ensemble des informations provenant du signal est extrait à ce stade.

Concernant la reconnaissance des instruments de musique, sujet que nous aborderons largement dans le chapitre applications (Chapitre 6), on peut trouver une revue des caractéristiques employées pour la reconnaissance des instruments de musique sur des notes isolées dans (Peeters, 2004), et sur des solos dans (Essid, 2005).

Les systèmes élaborés travaillent sur des ensembles de caractéristiques conséquents (543 pour (Essid, 2005), environ 500 pour (Peeters & Rodet, 2003)). Il est alors nécessaire de réduire ce nombre pour des raisons de complexité et parfois de mauvaise gestion des grandes dimensions par les algorithmes de classification. On utilise alors des méthodes de sélection de caractéristiques. Ces méthodes permettent de retenir les caractéristiques qui maximisent des critères de discrimination des classes.

2.1.1.2 Représentation des classes

Les algorithmes de classification supervisée permettent de représenter les classes à partir d'un entrainement sur une base d'apprentissage, puis de classer des données de classes inconnues en comparant leur représentation dans l'espace des caractéristiques aux modèles appris.

Les algorithmes d'apprentissage utilisent plusieurs paradigmes pour représenter les classes à distinguer. Certaines approches utilisent des modélisations statistiques des classes comme les Modèles de Mélanges de Gaussiennes (MMG ou *GMM*), les Modèles de Markov Cachés (MMC ou *HMM*) et les Réseaux Bayésiens Dynamiques (RBD ou *DBN*). D'autres sont basées sur la discrimination des classes mises en jeu, comme les arbres de décision qui délimitent des hyperplans séparant les classes, les Machines à Vecteurs Supports (MVS ou *SVM*) qui déterminent également des hyperplans séparant les classes mais cette fois-ci dans des espaces de dimension supérieure à l'espace de caractéristiques initial. Enfin, d'autres mettent en jeux des distances avec des représentants des classes comme les K-moyennes (*K-Means*) où la distance est calculée à partir des centroïdes des observations d'apprentissage, et les K-plus-proches voisins (KPPV ou *KNN*) où l'on choisit la classe la mieux représentée parmi les plus proches voisins de l'observation à classifier.

2.1.1.3 Décision

Si l'on excepte les classifieurs qui modélisent la dynamique de l'évolution des caractéristiques (DBN et HMM), les algorihmes de classification donnent un résultat pour chaque trame de signal. Certains donnent intrinsèquement des vraisemblances[1] (GMM), d'autres donnent des sorties qui peuvent être transformées en vraisemblances (SVM), enfin certains classifieurs ne peuvent donner qu'une décision brute par classe (KNN).

Le problème de l'aggrégation de ces décisions pour une fenêtre temporelle se pose quand cette fenêtre possède plusieurs trames d'analyse. Par exemple, le système d'Essid (2005) met en jeu des observations sur des trames temporelles de 30 ms environ, puis fusionne les décisions sur des fenêtres plus longues (par exemple 4 secondes) en considérant toutes les trames indépendantes. Dans ce cas, la vraisemblance de la classe sur la fenêtre de décision est le produit des vraisemblances de la classe sur les trames d'analyse. L'instrument qui possède la vraisemblance la plus haute est choisi comme résultat de la classification pour le segment considéré. D'autres stratégies peuvent être envisagées : si des décisions dures sont prises pour chaque trame temporelle, ce sont des stratégies de votes qui peuvent être mises en place.

2.1.1.4 Raffinements pour la musique polyphonique

L'emploi de ces méthodes à l'identification d'une ou plusieurs sources lorsque qu'elles sont activées simultanément dans un signal nécessite certaines adaptations. En effet, l'extraction de caractéristiques est un processus non-linéaire qui ne permet pas de tenir compte de la nature additive des mélanges de sources musicales. Les approches développées pour en tenir compte sont les suivantes :

– Eggink & Brown (2003) : les paramètres perturbés par les autres sources ne sont pas pris en compte par le classifieur (théorie de la caractéristique manquante).

[1] La vraisemblance d'un modèle M par rapport à une observation O est la probabilité conditionnelle $(O|M)$ que le modèle inconnu M ait généré les données connues O.

- Eggink & Brown (2004) : les peignes harmoniques prépondérants sont extraits par estimation de la fréquence fondamentale par somme spectrale pondérée, puis classifiés en utilisant des MMG. Cette approche a été appliquée à la reconnaissance automatique d'instruments solistes dans les sonates.
- Jinachitra (2004) : une séparation de sources par Analyse en Sous-espaces Indépendants est mise en oeuvre, puis une classification des signaux résultants par KPPV ou MMG selon un ensemble varié de caractéristiques est effectuée. La méthode a été appliquée à des mélanges de performances solos synthétiques, pour des instruments assez distincts dans leur timbre.
- Kitahara et al. (2005) : les modèles d'instruments dépendent de la hauteur de note, et l'apprentissage des modèles se fait sur des signaux déjà mélangés. La même idée a été reprise dans (Kitahara et al., 2006) pour une visualisation qui permet de représenter la présence des instruments dans un plan temps-pitch sans effectuer de détection de note.
- Essid et al. (2006a) : les ensembles d'instruments sont modélisés comme des classes individuelles en utilisant une classification hiérarchique. Cette méthode a été appliquée à des ensembles de jazz avec un succès satisfaisant.
- Ellis & Poliner (2006) : les modèles de notes sont appris sur des notes mélangées, afin de ne pas faire de surapprentissage sur des notes jouées seules. Ces modèles sont appliqués à la détection de mélodie.

Nous voyons ainsi que des méthodes basées sur l'approche sac de trames ont été développées pour traiter différent types de musique polyphonique.

Nous allons maintenant examiner une autre classe de méthodes qui ont pour base des modélisations explicites des signaux et des mélanges.

2.1.2 Méthodes par modèles de signaux et de mélanges

La seconde grande classe de méthodes d'analyse de signaux audio regroupe les techniques qui mettent en jeu un modèle de signal qui est censé représenter la ou les classes analysées. Ces méthodes utilisent une modélisation des sources et du mélange *qui permettent de resynthétiser les signaux séparés* au moins partiellement. Elles visent donc à représenter le signal sous la forme d'un ensemble d'objets sonores, éventuellement en y ajoutant un bruit qui représente la partie non-modélisée du signal. Ces approches ont pour application naturelle la séparation de source, la transcription automatique et l'identification d'instruments de musique. Les trois tâches mentionnées sont complémentaires : chacune sera d'autant mieux réalisée que les deux autres seront réussies.

Afin de mettre en oeuvre de telles méthodes, la structuration générale d'un signal musical doit être examinée plus précisément. Dans la musique occidentale, on peut considérer qu'une grande partie des oeuvres musicales est faite de notes produites par des instruments acoustiques ou électroniques. Les notes peuvent être considérées comme des évènements possédant donc un début (*onset*) et une fin (*offset*). Ils possèdent également une intensité (ou vélocité), et éventuellement une hauteur (*pitch*). Dans ce cas, leur agencement définit les harmonies et les mélodies. Les signaux musicaux sont donc le résultat de la captation de ces évènements de jeu sur des micros. Des traitements (incluant les effets de salle) sont ensuite appliqués sur les enregistrements. Ils peuvent modifier la spatialisation du son, sa dynamique, le profil spectro-temporel des notes etc.

La Figure 2.2 présente un schéma général d'enregistrement musical, qui consiste en

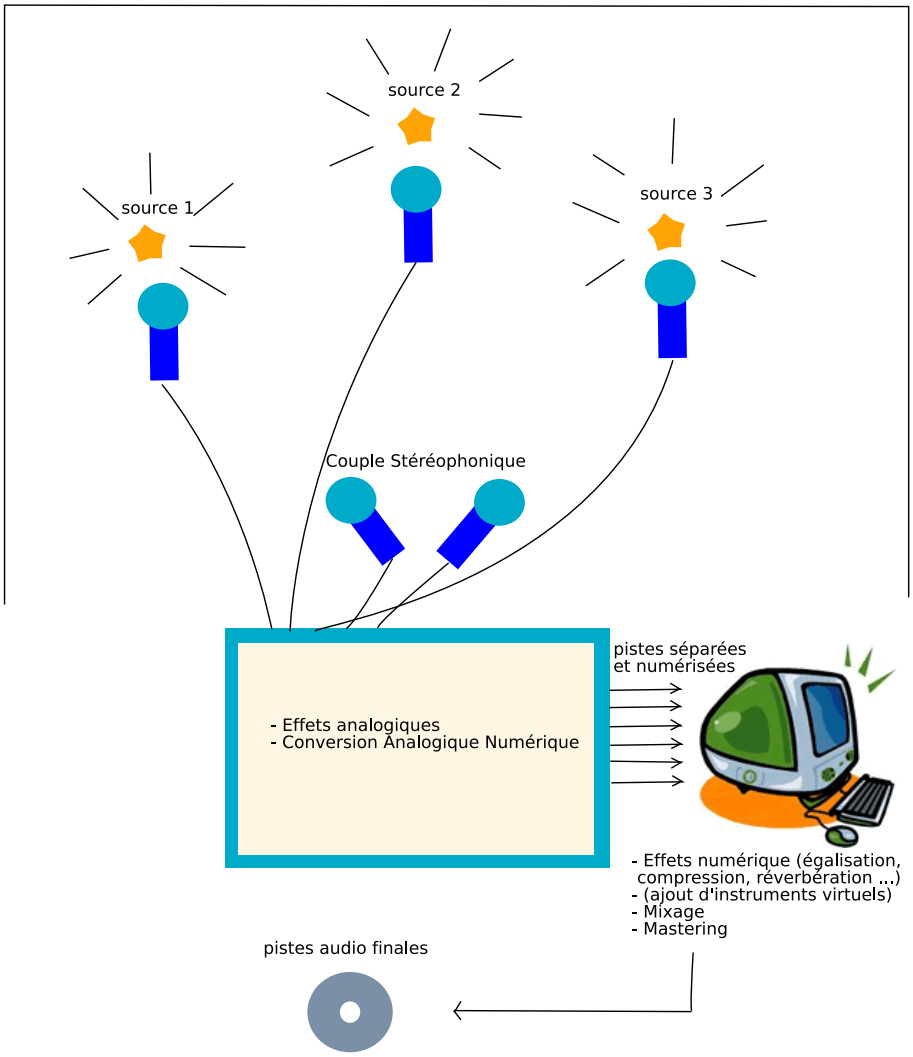


FIG. 2.2 – *Schéma d'un enregistrement moderne.*

un mélange et un traitement de sources musicales distinctes.

Formellement, un tel mélange peut s'écrire de cette façon (Vincent (2004)) :

$$x_i(u) = \sum_{j=1}^{n} \sum_{\tau=0}^{+\infty} a_{ij}(u-\tau,\tau) s_j(u-\tau) + n_i(u) \tag{2.1}$$

où x_i est le signal capté en i, a_{ij} les fonctions qui définissent le filtre entre la source j et le capteur i, u le temps, τ le délai entre l'émission en j et la réception en i, s_j le signal de la source et n_i le bruit sur le capteur i.

En gardant cette généralité, on qualifie ce mélange de *convolutif*. Si l'on pose $a_{ij}(u-\tau,\tau) = A_{ij}$ (constante), on parle alors de mélange instantané. Cette hypothèse de mélange n'est pas réaliste pour la majorité des enregistrements musicaux qui contiennent une réverbération naturelle ou synthétique.

Cependant, dans le cas monophonique, on peut noter que certains de ces traitements *a posteriori* comme la réverbération, les délais courts et la compression de la dynamique, peuvent être inclus dans les modèles de sources car difficilement discernables de l'enregistrement direct de la source. Il suffit donc d'inclure des signaux traités dans les ensembles d'apprentissage pour prendre en compte ces effets. Dans ce cas, le mélange est *vu* comme un mélange instantané, même s'il n'en est pas un. Le mélange peut alors se formaliser de cette façon :

$$x_i(u) = \sum_{j=1}^{n} s'_j(u) + n_i(u) \tag{2.2}$$

où s'_j sont les signaux des sources après les filtrages issus des traitements et des effets de salle. Cependant, en gardant cette hypothèse, le cas des délais longs ne peut pas être inclu dans ces traitements : si le délai est suffisamment long, la source est dupliquée dans le temps, et peut donc être vu comme une source à part entière pendant l'analyse. Nous allons donc passer en revue les méthodes qui prennent ces types de modélisation en compte.

2.1.2.1 Séparation de sources aveugles à un seul microphone

Les différents algorithmes de séparation de sources se distinguent par la quantité de connaissance injectée dans l'algorithme sur les sources et sur leur mélange. On parle de séparation de sources aveugle lorsque très peu d'hypothèses sont faites sur ces deux aspects.

Dans cette catégorie, on peut citer l'Analyse en Sous-espaces Indépendants (Casey & Westner (2000)). Dans ce cas de figure, le choix de la représentation du signal est fondamental. Si l'on traite des signaux audio, travailler sur le signal PCM brut ne conduit pas à des résultats probants. On travaille alors sur des représentations sur lesquelles les sources apparaissent plus séparées, comme des représentations temps-fréquence. Dans ce type d'approche, l'hypothèse faite sur le mélange est que les sources sont statistiquement indépendantes.

On peut également supposer que le mélange est une combinaison linéaire de vecteurs appartenant à un dictionnaire inconnu. Le travail de séparation de sources revient alors à un travail de *clustering*, où l'on essaye d'inférer le dictionnaire de sources à partir du mélange[2]. Pour traiter ce problème, divers algorithmes peuvent être utilisés. Une repré-

[2]Ces méthodes seront également évoquées dans 2.2.4.7

sentation temps-fréquence X peut être considérée comme une matrice, qui peut s'exprimer de la façon suivante : $X = DW + R$, où D est la matrice représentant les atomes du dictionnaire, W celle des poids des atomes et R un résidu. Il s'agit alors de minimiser une fonction de coût tenant compte de la norme de R et de la parcimonie de W. Des solutions à ce problème peuvent être trouvées grâce à des algorithmes de Factorisation de Matrices Non-Négatives (*NMF*,Lee & Seung (2001); Smaragdis & Brown (2003)). L'algorithme de K-SVD (Aharon et al. (2006)) peut également être utilisé. Il peut être vu comme une généralisation de l'algorithme des K-Moyennes, en ceci qu'il permet d'obtenir un *dictionnaire* avec lequel le signal est décomposé de façon parcimonieuse. On peut noter que la NMF peut être modifiée de telle sorte que de la connaissance soit injectée : Virtanen (2007) tient compte de contraintes temporelles et de parcimonie dans le calcul conduisant à une factorisation.

2.1.2.2 Les méthodes basées sur des modèles de sources

Nous allons maintenant nous intéresser aux méthodes d'analyse mettant en jeu des modèles des sources. L'intérêt d'utiliser de tels modèles est multiple. Tout d'abord, intégrer plus de connaissance *a priori* dans le processus de décomposition permet de mieux localiser les parties intéressantes du signal. Ensuite, si les classes de sources sont modélisées séparément, l'information de classe apparaît explicitement dans la décomposition. Cependant, elles présentent quelques désavantages : d'une part la modélisation des sources nécessite un apprentissage, d'autre part, bien entendu, la décomposition n'est intéressante que si le signal met en jeu les sources apprises. Voici des approches que nous pouvons évoquer :

- Kashino & Murase (1999) proposent une technique d'adaptation de modèle après extraction de la fréquence fondamentale. La méthode est basée sur des filtres adaptés : des filtres sont appris sur des notes isolées, puis la corrélation entre la sortie de ces filtres et leur entrée permet de déterminer quel instrument joue. Concernant la phase, une méthode de tracking de la phase a été développée, en se basant sur l'évolution linéaire de la phase en fonction du temps pour une fréquence donnée.
- Vincent (2004) a utilisé un modèle probabiliste pour les sources instrumentales et leur mélange. Concernant les sources instrumentales, la structure de signal utilisée est la transformée de Fourier à Court Terme du signal. Une fois un certain nombre de ces structures collectées sur des notes isolées d'instruments, un modèle graphique probabiliste est développé : il permet d'expliquer le signal par une somme de notes d'instruments, commandés en timbre par des paramètres qui déterminent l'écart du spectre observé par rapport à un spectre prototypique. Une idée similaire a été présentée par Vogel et al. (2005).
- Ozerov et al. (2005) propose une approche par filtrage de Wiener adaptatif pour séparer voix et musique. Chacune de ces sources est modélisée par des MMG sur la transformée de Fourier à Court Terme. L'avantage de la méthode proposée ici est qu'elle permet une adaptation des filtres au signal, en estimant un filtre global qui vient s'ajouter aux spectres existants.

Nous allons maintenant examiner une troisième manière de traiter des signaux musicaux. Il s'agit d'en extraire des caractéristiques pertinentes pour les traitements désirés, mais en obtenant une représentation plus structurée que celle à base de caractéristiques et sans forcément viser la transcription parfaite comme dans (Vincent, 2004).

2.1.3 Représentations "mi-niveau"

Une autre approche intéressante pour nos problématiques est celui des représentations dites de *mi-niveau*. Nous parlerons également de représentations *objet* dans la suite. Parler de représentations mi-niveau peut être considéré comme abusif, cela suppose en effet qu' il y aurait une gradation sur laquelle elles seraient comparées aux représentations "bas-niveau", issues de mesures physiques à l'aide de capteurs et aux représentations "haut-niveau" qui présentent les données sémantiques. Ici, on ne considérera les représentations mi-niveau proposées que comme des éléments d'une solution technique pour des tâches d'indexation ou de traitement audio, sans se placer sur une quelconque échelle de niveau.

Le principe de ces représentations objet est de représenter le signal d'une façon plus structurée en exhibant des éléments plus faciles à interpréter que l'information brute (signal PCM). Comme dans les représentations présentées en 2.1.1, la construction de telles représentations exige aussi l'introduction de connaissance a priori sur le signal analysé. Prenons le cas de la Transformée de Fourier à Court Terme : cette classe de représentation est utilisée très fréquemment dans le signal audio. Elle consiste à projeter le signal sur des sinusoïdes complexes fenêtrées. C'est sur les paramètres de ce processus (nombre d'échantillons de calcul de la transformée, taille de la fenêtre d'analyse, pas de calcul et type de fenêtre) que porte l'introduction d'une connaissance haut-niveau : si on analyse des phénomènes impulsifs, des fenêtre courtes seront choisies, tandis que des phénomènes nécessitant une bonne résolution fréquentielle seront analysés avec des fenêtres longues.

Ainsi, au sens strict, toute représentation (utile) d'un signal résultant d'un traitement des informations brutes (signal PCM) peut être considérée comme "objet", en ceci qu'elle nécessite l'introduction de connaissance pour sa construction. La différence entre ces représentation se fait au niveau des attributs que possèdent chacun des éléments de la représentation.

Ellis & Rosenthal (1998) énoncent un certain nombre de propriétés désirables pour la représentation mi-niveau idéale :

1. **Séparation des sources** : chaque élément de la représentation doit pouvoir représenter sans ambiguïté exactement une source.
2. **Inversibilité** : une représentation mi-niveau doit permettre de resynthétiser le signal original à l'identique.
3. **Réduction du nombre de composantes** : une représentation mi-niveau doit représenter le signal avec le moins d'éléments possibles.
4. **Pertinence des caractéristiques** : les paramètres des éléments doivent être en relation avec des paramètres pertinents des sources.
5. **Plausibilité physiologique** : la représentation doit être liée à des caractéristiques physiologiques.

Avant de nous lancer dans la quête de nouvelles représentations mi-niveau, on peut néanmoins émettre quelques critiques sur ces énoncés. En effet, l'énumération de ces propriétés a pour objectif de donner un guide pour la construction de représentations à partir desquelles n'importe quelle tâche audio pourrait être réalisée. Or, il n'est pas possible de connaître *a priori* l'étendue des traitements que l'on pourrait effectuer sur le signal. Il n'est donc pas concevable de déterminer tous les paramètres pertinents des signaux à extraire, ce qui contredit la quatrième propriété. De même, il est difficile de connaître

a priori la quantité d'information suffisante, ce qui s'oppose à la troisième. Cependant, si l'on se fixe un cadre applicatif, les paramètres peuvent être déterminés. Par exemple, pour une recherche de tonalité, une représentation de type MIDI peut suffire, sans que des caractéristiques timbrales soient extraites. Une autre propriété criticable est celle de l'inversibilité : dès lors qu'on cherchera à introduire une modélisation resynthétisable du signal, une erreur d'approximation sera présente, excepté si le signal a été généré à l'aide du modèle.

Des représentations dites mi-niveau ont donc été développées spécifiquement pour des tâches de *Music Information Retrieval* pour la recherche de mélodie (Song et al. (2002)), l'analyse en grilles harmoniques (Bello & Pickens (2005)), ou la similarité musicale (Marolt (2006)). Les propriétés de ces représentations sont finalement assez éloignées de celles décrites par Ellis : elles n'exhibent que les paramètres pertinents pour les applications considérées, et ne permettent donc pas de définir une représentation mi-niveau *unique* et universelle, dont d'ailleurs rien ne prouve l'existence. De plus, aucune d'entres elles ne permet une resynthèse du signal, même approximative. Ainsi, la collection de ces représentations mi-niveau permet d'obtenir différentes *grilles de lecture* du signal, et d'en déduire des caractéristiques symboliques ou relationelles assez simplement. Cependant, prises individuellement, elles ne permettent pas directement de rendre compte de la structuration en note de signaux musicaux.

Dans leur article, Ellis & Rosenthal (1998) caractérisent quelques représentations de bas-niveau selon cette grille d'analyse : les modèles sinusoïdaux, la transformée de Fourier à Court Terme, la Transformée à Q constant, le spectre de modulation, le corrélogramme, et la trame (*weft*). Cette trame est en fait une somme de sinusoïdes dont les amplitudes et fréquences instantanées varient de façon corrélée. Ces représentations sont de bonnes candidates pour la représentation des notes de musique produites par des instruments harmoniques, étant données qu'elles captent toute la partie harmonique des signaux et leurs fluctuations internes.

2.1.4 Synthèse sur les approches en indexation audio

Les approches par extraction de paramètres et celles par modèles de signaux présentent de nombreux points de convergence concernant le traitement de la musique polyphonique. Nous avons en effet remarqué que la mise en valeur de structures de signal est utilisée pour appliquer les méthodes sac-de-trames. Parallèlement, l'extraction de caractéristiques pertinentes sur les structures de signal modélisées commence à émerger pour tirer partie des algorithmes efficaces d'apprentissage statistique. Nous essaierons donc de garder à l'esprit cette notion pour la construction de notre modèle de signal et des algorithmes permettant de décomposer un signal quelconque selon ce modèle.

Nous allons maintenant présenter un état de l'art en représentations parcimonieuses des signaux audio, et voir en quoi elles peuvent correspondre à des représentations mi-niveau intéressantes.

2.2 Représentations parcimonieuses

Les conditions mentionnées pour l'obtention de bonnes représentations mi-niveau nous conduisent à chercher des structures additives de signaux. Comme nous allons le voir, gérer formellement ce type de problème est possible en rentrant dans le cadre des

représentations parcimonieuses, à la fois dans la recherche des bons modèles de signaux et dans les algorithmes pour décomposer les signaux selon ces structures.

2.2.1 De la parcimonie

L'utilisation du concept de parcimonie dans les sciences est très ancien. Son introduction est le plus souvent attribuée au moine franciscain et philosophe Guillaume d'Occam (XIVe siècle), et demeure une des fondations de la science actuelle : " Les entités ne doivent pas être multipliées par delà ce qui est nécessaire ". Celà se traduit dans l'approche scientifique par une invitation "à ne pas utiliser de nouvelles hypothèses tant que celles déjà énoncées suffisent, à utiliser à fond les hypothèses qu'on a déjà faites, avant d'introduire de nouvelles hypothèses, ou autrement dit à ne pas apporter aux problèmes une réponse spécifique, ad hoc, avant d'être (pratiquement) certain que c'est indispensable (sinon on risque d'escamoter le problème, et de passer à côté d'un théorème ou d'une loi physique)" (Wikipedia (2007)). Cependant, si la parcimonie est utilisée comme méthode pour la recherche de lois scientifiques, elle ne garantit pas que la solution trouvée est vraie.

Dans le domaine du traitement du signal, la parcimonie sera également vue comme une approche du problème de l'approximation d'un signal : le principe d'une représentation parcimonieuse d'un signal est de concentrer l'énergie du signal sur un faible nombre d'éléments. Ainsi, on cherchera à représenter le signal comme une combinaison linéaire de formes d'ondes, appelées *atomes* :

$$x \simeq \sum_{\lambda \in \Lambda} \alpha_\lambda g_\lambda \tag{2.3}$$

où x est le signal, (g_λ) les atomes et (α_λ) leurs poids. Ces atomes seront choisi dans un ensemble, appelé *dictionnaire*. L'ensemble des atomes sélectionnés pour représenter x sera appelé *livre* dans la suite. Ce dictionnaire peut contenir des exemples d'atomes, ainsi que des formes d'ondes définies par un nombre restreint de paramètres, comme nous le verrons dans le paragraphe 2.2.4. Représenter un signal de façon très parcimonieuse ne permet pas forcément d'en comprendre son contenu, tout dépendra des connaissances introduites dans l'algorithme de décomposition et le dictionnaire.

Nous voyons donc dans ce préambule que l'objectif de ces représentations est lié à la propriété 3 du paragraphe 2.1.3 (réduction du nombre de composantes). Un lien peut être également fait avec la propriété 2 (inversibilité), mais de façon approximative car dans le cas général, la nullité du résiduel à l'aide d'une représentation parcimonieuse du signal ne sera pas atteinte. On peut cependant atteindre une décomposition qui est identique du point de vue perceptif à l'original (Verma & Meng (1999)), sans que le résiduel soit nul.

2.2.2 Position du problème

Si l'on veut obtenir des représentations parcimonieuses de signaux, deux aspects sont à considérer : l'ensemble des atomes sur lesquels on décompose le signal (*dictionnaire*), et l'algorithme utilisé pour effectuer la décomposition. Ces deux aspects ne sont pas strictement disjoints : s'il existe des algorithmes permettant de décomposer un signal avec n'importe quels atomes, il est aussi possible de construire des algorithmes permettant

de tirer parti de certaines propriétés du dictionnaire. Si les atomes sont paramétrés, un échantillonnage de leurs paramètres est nécessaire afin de rendre le dictionnaire discret.

L'objectif d'un processus de représentation parcimonieuse peut être défini de plusieurs façons :

- Etant donné une valeur de norme ν sur l'espace des coefficients $\mathcal{N}(\alpha_\lambda)$, trouver les atomes et leurs poids associés pour minimiser une distance $\mathcal{D}$ entre le signal x et son approximation $\sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda$:

$$(\Lambda_0,(\alpha_{\lambda i})_{i=1..|\Lambda|}) = \arg\min_{\Lambda,(\alpha_\lambda)_{i=1..|\Lambda|}} \left\{ \mathcal{D}(x,\sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda) \quad | \quad \mathcal{N}(\alpha_\lambda) < \nu \right\} \tag{2.4}$$

- Etant donné une valeur de distance entre le signal et son approximation ϵ, trouver la combinaison d'atomes dont une norme donnée $\mathcal{N}$ dans l'espace des coefficients est minimale :

$$(\Lambda_0,(\alpha_{\lambda i})_{i=1..|\Lambda|}) = \arg\min_{\Lambda,(\alpha_\lambda)_{i=1..|\Lambda|}} \left\{ \mathcal{N}(\alpha_\lambda) \quad | \quad \mathcal{D}(x,\sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda) < \epsilon \right\} \tag{2.5}$$

- Fixer une fonction de coût qui dépende du nombre d'atomes sélectionnés (critère de parcimonie) et du niveau de résiduel atteint (critère de fidélité aux données) :

$$(\Lambda_0,(\alpha_{\lambda i})_{i=1..|\Lambda|}) = \arg\min_{\Lambda,(\alpha_\lambda)_{i=1..|\Lambda|}} \left\{ \mathcal{D}(x,\sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda) + \gamma\mathcal{N}(\alpha_\lambda) \right\} \tag{2.6}$$

  Suivant la distance $\mathcal{D}$ et la norme $\mathcal{N}$ choisies, différents algorithmes de résolution peuvent être employés (voir 2.2.5.1). Le facteur γ permet d'accentuer la contrainte soit sur le terme de fidélité aux données (premier terme), soit sur le terme de parcimonie (second terme).

On peut noter que la norme 0 (cardinal de l'ensemble) est un choix désirable pour $\mathcal{N}$ dans un bon nombre d'applications. Cette norme rend cependant les processus d'optimisation assez difficile. Quant à $\mathcal{D}$, il est commun d'utiliser la distance euclidienne (norme 2 de la différence entre le signal et l'approximant) afin d'utiliser les nombreux résultats issus de la théorie des espaces euclidiens. Minimiser $||x - \sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda||_2$ revient également à poser un problème de Maximisation de la Vraisemblance par rapport au dictionnaire et aux coefficients, en posant un modèle gaussien pour le résiduel $x - \sum_{\lambda\in\Lambda}\alpha_\lambda g_\lambda$ de cette approximation.

Concernant le choix du terme de parcimonie, Gribonval & Nielsen (2003) ont montré que la décomposition la plus parcimonieuse est indépendante du choix de la norme pour le terme de parcimonie sous certaines conditions très strictes sur les poids des atomes de la décomposition et du dictionnaire.

2.2.3 Atomes et molécules

Un atome est une forme d'onde d'énergie unité. Lorsqu'il est utilisé pour décomposer un signal, on l'associe à un scalaire α_λ que l'on nomme son poids.

Une molécule est un ensemble d'atomes. Lorsqu'elle est sélectionnée dans une décomposition, on l'associe à un vecteur de poids qui est constitué des poids respectifs des atomes. La forme d'onde ainsi obtenue s'écrit :

$$\mu(t) = \sum_{\lambda\in\mathcal{M}}\alpha_\lambda g_\lambda(t) \tag{2.7}$$

Les atomes et les molécules sont des entités différentes en chimie : une molécule est composée d'atomes, et ne peut pas être considérée comme un atome. La métaphore n'est pas directement transposable en traitement du signal : ces deux entités peuvent être représentées par un signal, et suivant le point de vue qu'on adopte, une même structure de signal peut être considérée comme un atome ou une molécule.

En effet, μ peut être normalisé à un : dans ce cas, l'atome ainsi constitué est paramétré par les paramètres et les poids respectifs des atomes qui forment ce nouvel atome. On peut alors utiliser plusieurs hiérachies en même temps pour décrire un signal. Par exemple, on peut très bien comparer les poids respectifs d'une sinusoïde et d'une somme de sinusoïdes en rapports harmoniques, et utiliser ces deux types de stuctures dans une même décomposition du signal. Des nouveaux atomes peuvent être à nouveau construits à partir de tels atomes, et ainsi de suite jusqu'à ce que le niveau de granularité désiré soit obtenu. Il est cependant facilement imaginable que la construction de dictionnaires composés de plusieurs couches d'atomes peut être empêchée par le grand nombre de combinaison d'atomes possible. Cela rend donc nécessaire la construction d'algorithmes *ad hoc* pour extraire les molécules, ce que nous verrons dans la suite.

2.2.4 Bestiaire des formes d'ondes

Dans cette section, nous décrivons quelques dictionnaires de formes d'ondes qui ont déjà été développés.

2.2.4.1 L'ondelette de Gabor

Pour le traitement des signaux audio, différentes formes d'ondes ont été utilisées. La première d'entre elle est l'ondelette de Gabor (1947) (voir Figure 2.3). Elle s'exprime sous la forme d'une exponentielle complexe pondérée par une fenêtre Gaussienne w :

$$g_{s,u,f}(t) = w\left(\frac{t-u}{s}\right) e^{2j\pi ft} \tag{2.8}$$

où u est la localisation en temps, s l'échelle de l'atome (sa longueur effective) et f sa fréquence.

La raison pour laquelle elle s'adapte bien aux signaux audio est la suivante : les signaux musicaux sont essentiellement composés de morceaux de sinusoïdes. On peut aussi remarquer qu'effectuer la projection d'un signal sur un dictionnaire d'atomes de Gabor avec une échelle s constante, un espacement entre les localisations d'atomes Δu constant et un échantillonnage en fréquence linéaire revient exactement à en calculer la Transformée de Fourier à Court Terme (TFCT) avec une fenêtre de pondération Gaussienne. La TFCT est utilisée abondamment dans toutes les tâches d'analyse et de synthèse du son, qu'elles soient automatiques ou non, par exemple par McAulay & Quatieri (1986); Serra (1989); Depalle et al. (1993); Lagrange et al. (2004). Dans ces tâches, le choix des paramètres d'analyse, c'est-à-dire de l'ondelette de Gabor, est effectué par l'opérateur en fonction de la tâche à effectuer. Dans le domaine des représentations parcimonieuses, nous verrons que ce choix ne sera pas nécessaire : le critère qui permet de choisir une taille d'atome plutôt qu'une autre est le maximum de la corrélation entre les atomes de différentes tailles et le signal.

Nous remarquons aussi que ces atomes sont complexes. Ainsi, la décomposition d'un signal réel sur de tels atomes reviendra à les utiliser par paire d'atomes conjugués (Mallat & Zhang

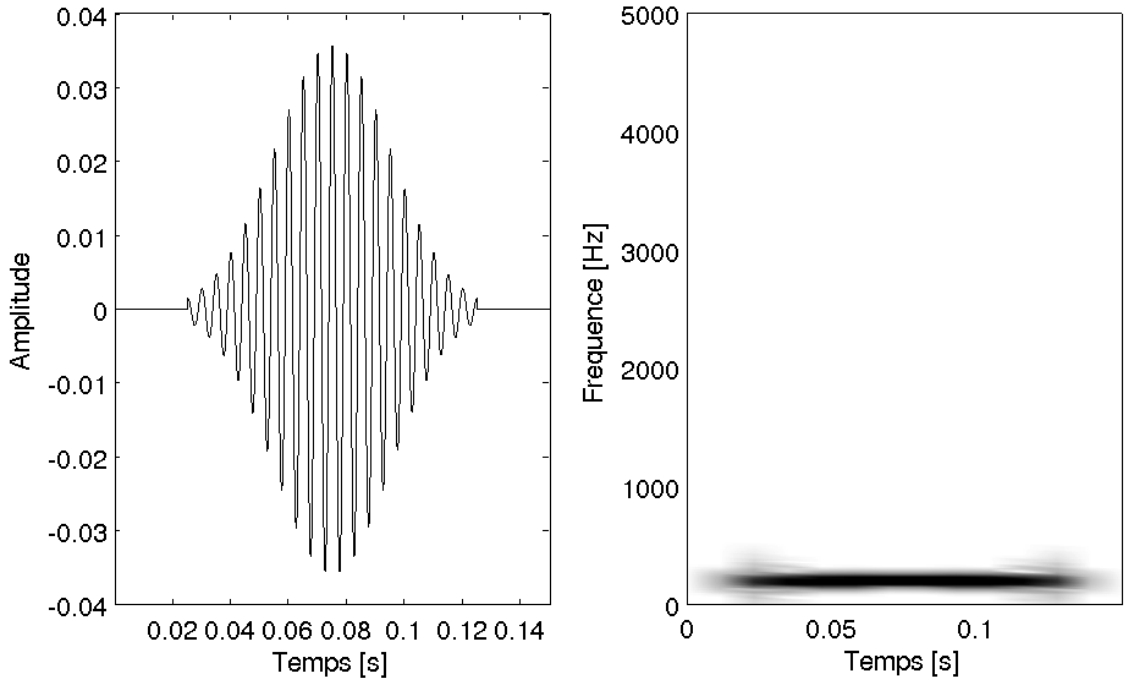


FIG. 2.3 – représentations temporelle (gauche) et spectrogramme (droite) d'un atome de Gabor

(1993); Goodwin & Vetterli (1999); Goodwin (2001)). Dans ce cas, les poids et phases des atomes réels doivent être calculés par projection orthogonale du signal réel sur les atomes complexes conjugués. On parle alors de dictionnaires de molécules di-atomiques.

Ce type de dictionnaire a été utilisé par Mallat & Zhang (1993) avec différentes échelles s. Il a ensuite été employé par Goodwin (2001) et Heusdens et al. (2002) pour le codage audio. On peut également mentionner une application à la visualisation, en reconstruisant une carte temps fréquence à partir de la somme des transformées de Wigner-Ville des atomes lorsqu'on utilise plusieurs résolutions (Mallat & Zhang (1993)).

On peut également noter que les autres fenêtre utilisées en traitement du signal audio peuvent être utilisées (Hann, Hamming etc.) ainsi que des fenêtres asymétriques comme la Formant-wave-Function (FOF, Rodet (1984)).

2.2.4.2 Sinus et cosinus locaux

Les sinus et les cosinus locaux sont assez proches des atomes de Gabor. S'ils capturent des caractéristiques similaires aux atomes de Gabor (structures temps-fréquence), l'échantillonnage temporel particulier de leurs phases peuvent rendre ces dictionnaires orthogonaux, on parle alors de MDCT ou MDST (*Modified Discrete Cosinus (Sinus) Transform*). Dans ce cas, ces atomes s'écrivent (en notation signal discret) :

$$g_{u,f}[t] = g_u[t]\sqrt{\frac{2}{s}}\cos\left[\frac{\pi}{s}\left(f+\frac{1}{2}\right)\left(t+\frac{1}{2}\right)\right] \tag{2.9}$$

avec $g_u[t] = \sin[(\pi/2s)(t + s/2 + 1/2)]$ par exemple. D'autres fenêtres g_u peuvent être utilisées.

Cependant, le fait de fixer la phase de ces atomes a pour conséquence de rendre l'amplitude de la projection du signal dépendante de la phase relative de l'atome par rapport à la celle d'une sinusoïde sous-jacente dans le signal. Ainsi une sinusoïde sera représentée par des amplitudes de projection oscillantes. Cet aspect n'est pas génant pour la reconstruction du signal s'ils sont collectés sur des points fréquentiels successifs, mais peut l'être néanmoins à des fins d'analyse, à moins que les projections ne soient régularisées (Daudet & Sandler (2004)).

De tels dictionnaires ont été utilisés dans de nombreuses études mettant en jeu des décompositions du son, notamment en modification de la parole (George & Smith (1992)) ou en codage audio (Vos et al. (1999)). C'est notamment ce type de représentation qui est utilisé pour le codage MP3 et AAC (ISO/IEC 14496-3 :2001 (2001)).

2.2.4.3 Atomes de Gabor modulés en fréquence (*chirps*)

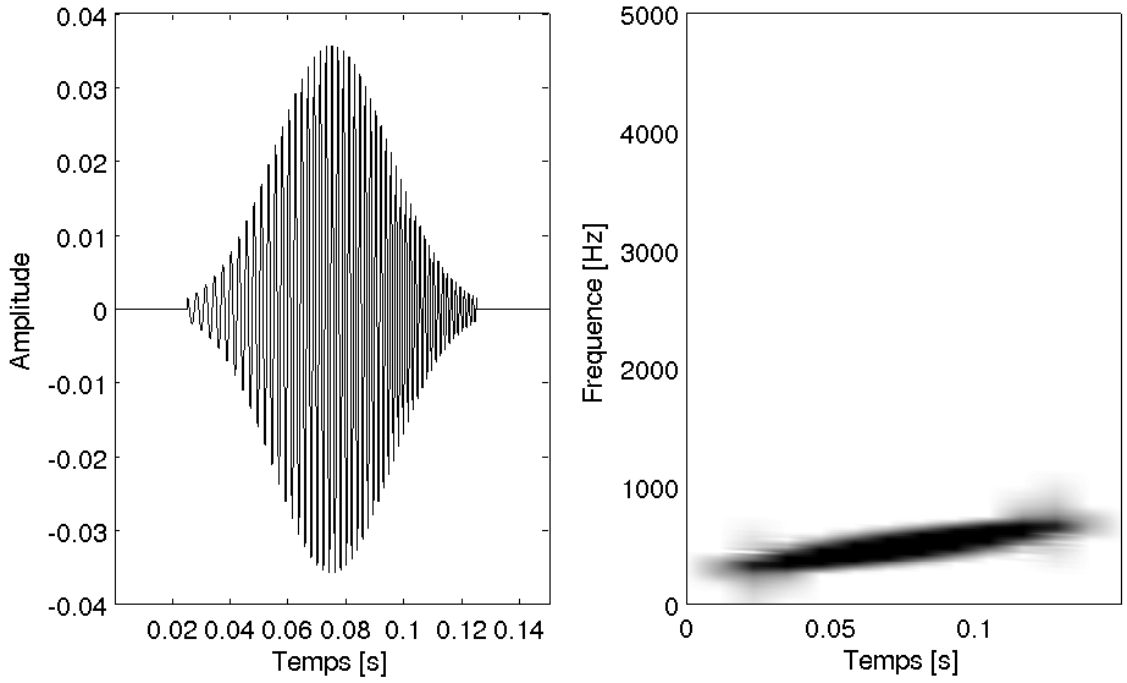


FIG. 2.4 – *représentations temporelle (gauche) et spectrogramme (droite) d'un atome de Gabor modulé en fréquence.*

Un raffinement des atomes de Gabor consiste à introduire un terme de modulation de fréquence :

$$g_{s,u,f,c}(t) = \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) e^{2j\pi\left(f(t-u)+\frac{c}{2}(t-u)^2\right)} \tag{2.10}$$

où c est le taux de modulation de fréquence, ou *chirp rate* (voir Figure 2.4). De tels atomes ont été utilisés par Bultan (1999); Gribonval (2001). Ils ont l'avantage de représen-

ter de façon plus adéquate les modulations de fréquence qui sont contenues dans certains sons musicaux (vibrato, glissando).

2.2.4.4 Sinusoïdes amorties

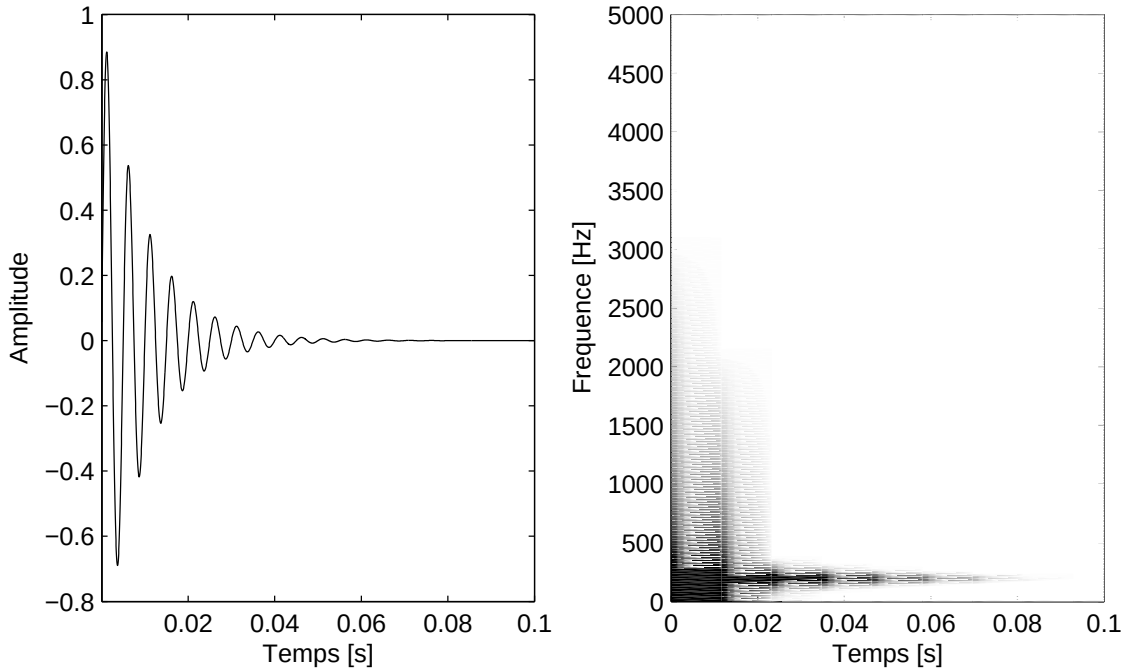


FIG. 2.5 – *représentation temporelle (gauche) et spectrogramme (droite) d'un atome sinusoïdal amorti.*

Goodwin & Vetterli (1999) introduisent les atomes sinusoïdaux amortis, et calculent les projections sur ces atomes par une approche en bancs de filtres :

$$g_{s,u,f,\alpha,\phi}(t) = \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) e^{-\alpha t} \sin(2\pi\,(f(t-u)+\phi) \tag{2.11}$$

où α est un facteur d'amortissement (voir Figure 2.5), w une fenêtre rectangulaire. Ces atomes permettent de bien représenter les signaux percussifs. On peut également obtenir des atomes asymétriques en pondérant des exponentielles complexes par d'autres fenêtres que les fenêtres symétriques classiques (Gribonval & Bacry (2003)).

2.2.4.5 Atomes harmoniques

Les atomes harmoniques ont été proposés par Gribonval & Bacry (2003). Ils mettent en jeu des combinaisons linéaires d'atomes de Gabor, dont les fréquences fondamentales sont en rapport harmonique :

$$h_{s,u,f_0,A,\Phi} = \sum_{m=1}^{M} a_m\, e^{j\phi_m} g_{s,u,m.f_0} \tag{2.12}$$

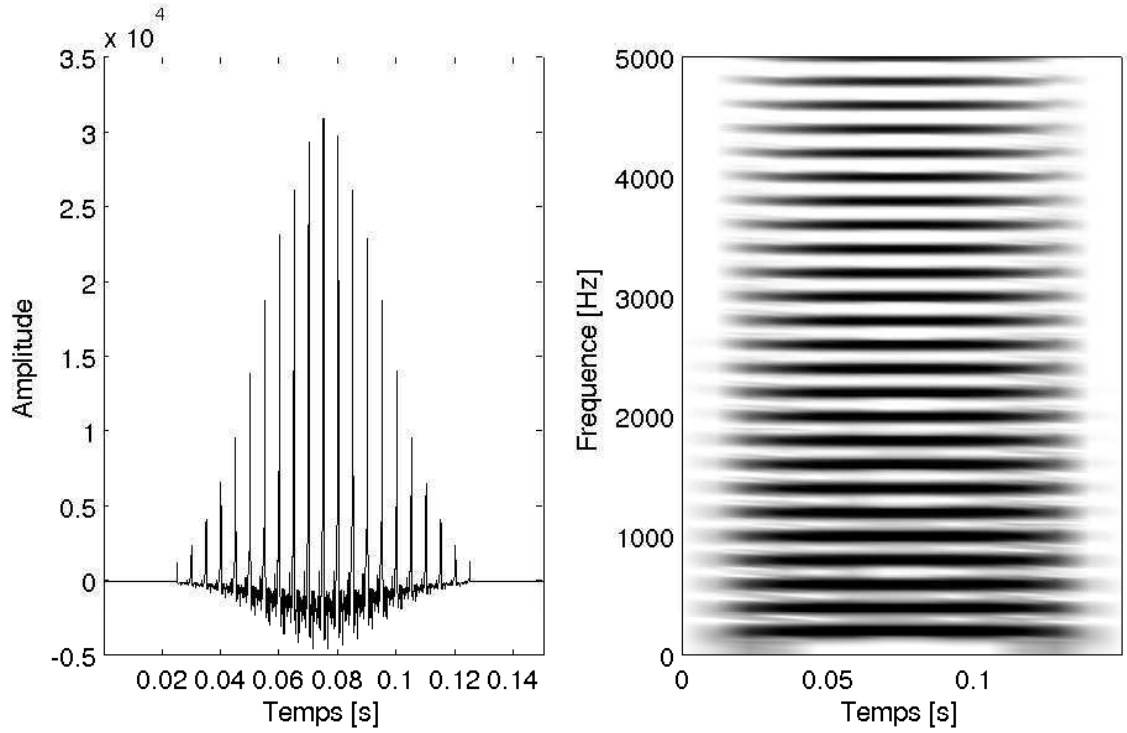


FIG. 2.6 – *représentation temporelle (gauche) et spectrogramme (droite) d'un atome harmonique.*

où $g_{s,u,m.f_0}$ sont des atomes de Gabor, a_m les amplitudes des partiels et ϕ_m leurs phases (voir Figure 2.6).

2.2.4.6 Ondelettes dyadiques

Les ondelettes dyadiques peuvent être utilisées pour construire un pavage temps-fréquence plus adapté à la représentation des signaux transitoires (Molla (2003), Daudet (2006)). Le pavage temps-fréquence qui découle de l'utilisation de ce type de base permet d'obtenir des bonnes résolutions temporelles en hautes fréquences, et des bonnes résolutions fréquentielles en basses fréquences. De plus, lorsqu'une singularité se produit dans le signal (transitoire rapide), les valeurs des projections selon les différentes échelles sont très corrélées (Mallat (2000)). On peut alors extraire des molécules qui permettent de capturer le transitoire en entier (Daudet (2006)).

Un exemple d'ondelette dyadique est présenté sur la Figure 2.7.

2.2.4.7 Atomes quelconques appris

Tous les atomes présentés précédemment possèdent une structure bien définie. Cependant, il est tout à fait possible de travailler avec des formes d'ondes arbitraires. Elles ne peuvent avoir un intérêt que si elles présentent une certaine similarité avec le signal analysé, et que le dictionnaire ainsi constitué permette de satisfaire correctement les contraintes de parcimonie.

Afin de construire des dictionnaires qui satisfont ces contraintes, de nombreux algorithmes ont été développés. Nous n'entrerons pas dans le détail de ces méthodes, le

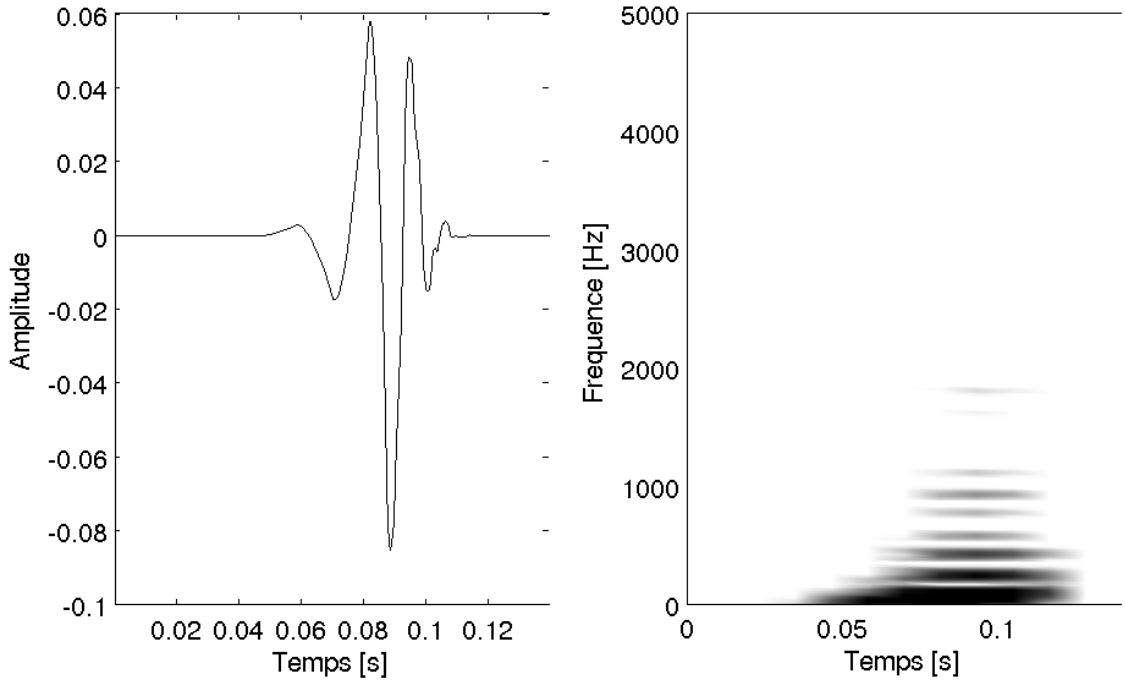


FIG. 2.7 – *représentation temporelle (gauche) et spectrogramme (droite) d'une ondelette de Daubechies 8.*

lecteur intéressé pourra se référer à (Lesage, 2007) pour en avoir une vue d'ensemble.

2.2.4.8 Atomes stéréo

Les atomes décrits ci-dessus peuvent être définis en stéréo (Gribonval (2002)). On écrit alors dans le cas général :

$$g_{st\,\lambda,\theta,\tau}(t) = [\cos(\theta) g_\lambda(t), \sin(\theta) g_\lambda(t-\tau)] \tag{2.13}$$

où λ est l'éventuel vecteur de paramètres de g, θ un paramètre qui définit la panoramique de l'atome dans l'espace et τ le temps de décalage entre les deux canaux.

L'information de panoramique des atomes peut être utilisée pour identifier la position des sources dans l'espace.

2.2.5 Algorithmes de décomposition

2.2.5.1 Présentation générale

Les algorithmes développés pour obtenir des décompositions parcimonieuses résolvent l'un des problèmes d'approximation présentés en 2.2.2. En général, les approches frontales qui consistent à effectuer une recherche exhaustive de toutes les combinaisons possibles d'atomes sont beaucoup trop complexes. Nous allons donc détailler les stratégies développées pour contourner cet obstacle.

L'algorithme de meilleure base orthogonale (*Best orthogonal basis, BOB*) présenté par Coifman & Wickerhauser (1992) utilise un critère d'entropie de représentation afin de choisir quelle est la base sur laquelle la représentation est la plus parcimonieuse pour des segments de signal audio. Les bases utilisées dans l'application sont des sinus locaux et des ondelettes dyadiques.

L'algorithme de Matching Pursuit a été présenté par Mallat & Zhang (1993). Nous le détaillerons dans la section suivante. Il est condidéré comme un algorithme *glouton* : il approxime le signal de façon itérative, en sélectionnant à chaque fois l'atome le plus corrélé avec le signal.

Une autre classe d'algorithmes traite le problème suivant, cas particulier du problème (2.6) :

$$(\Lambda_0, (\alpha_{\lambda i})_{i=1..|\Lambda|}) = \arg \min_{\Lambda,(\alpha_\lambda)_{i=1..|\Lambda|}} \left\{ ||x - \sum_{\lambda \in \Lambda} \alpha_\lambda g_\lambda||_2 + \gamma ||\alpha_\lambda||_1 \right\} \tag{2.14}$$

La norme 1 est retenue pour le critère de parcimonie. Bien que minimiser un critère faisant intervenir la norme 0 est souvent plus intéressant d'un point de vue applicatif, l'utilisation de la norme 1 permet de mettre en oeuvre des algorithmes de résolution moins complexes que la recherche exhaustive.

Parmi eux, on peut mentionner l'algorithme FOCUSS (Gorodnitsky & Rao (1997)), algorithme itératif composé de deux étapes. La première consiste à obtenir une estimation grossière du signal qui initialise la représentation, la seconde permet de rendre la représentation parcimonieuse, en se basant sur le fait que l'initialisation rend la solution du problème unique.

Le principe du Basis Pursuit (Chen et al. (2001)) propose de mettre en oeuvre des algorithmes de programmation linéaire pour résoudre le même problème.

Une autre classe de techniques utilisent le principe du codage parcimonieux (*Sparse Coding*). Il s'agit dans ce cas de formuler le problème d'approximation du signal comme une Maximisation a Posteriori : des *a priori* sont fixés sur la distribution des coefficients du dictionnaire, de telle sorte qu'ils expriment la concentration de l'énergie du signal sur un faible nombre d'atomes. Cette formulation du problème peut également servir à l'apprentissage du dictionnaire si celui-ci n'est pas fixé. Le premier type d'*a priori* utilisé est une distribution de Laplace qui présente un lobe plus fin que la distribution gaussienne (Olshausen & Field, 1997), et qui permet de bien représenter la concentration de l'énergie. On peut également utiliser un mélange de deux Gaussiennes, l'une à faible variance et l'autre à forte variance (Molla, 2003), ou des lois de Student (Févotte et al., 2004). Si les coefficients présentent une structure, par exemple un chaînage temporel pour des atomes représentant des sinusoïdes longues, on peut également introduire des modèles de chaînages des coefficients, par exemple markoviens (Févotte et al., 2006).

La factorisation en matrices non-négatives (NMF) que nous évoquions précédemment peut également être cité comme un algorithme de décomposition parcimonieuse. Cet algorithme a la particularité de permettre à la fois un apprentissage du dictionnaire et la décomposition du signal à l'aide de ce dictionnaire. Cependant, dans le cadre du signal audio, il nécessite une représentation temps-fréquence préalable (le plus souvent une TFCT) : il n'y a pas de retour au signal temporel, ce qui empêche l'utilisation de dictionnaires multi-résolution.

2.2.5.2 L'algorithme de Matching Pursuit

L'algorithme de Matching Pursuit (Mallat & Zhang (1993)) est un algorithme qui consiste à sélectionner itérativement l'atome qui, soustrait au signal avec son poids optimal, minimise le résidu. Il permet de trouver une solution en général sous-optimale aux problèmes (2.4) et (2.5).

Considérons la première itération. Soit x le signal à décomposer et g_λ un atome du dictionnaire. Par projection orthogonale de x sur g_λ, on a alors :

$$x = \langle x, g_\lambda \rangle g_\lambda + Rx \tag{2.15}$$

où Rx est le résidu de l'approximation de x selon g_λ. Le poids de l'atome extrait est $\alpha_\lambda = \langle x, g_\lambda \rangle$.

Le processus décrit ci-dessus est ensuite réitéré sur les résiduels successifs jusqu'à ce qu'une condition d'arrêt est atteinte, sur le niveau de résiduel (ou Rapport Signal à Résiduel, RSR) ou sur le nombre d'atomes sélectionnés. Le signal peut alors être approché par :

$$\hat{x} = \sum \alpha_\lambda g_\lambda \tag{2.16}$$

On peut noter que, comme g_λ et Rx sont orthogonaux, on a :

$$||x||^2 = |\langle x, g_\lambda \rangle|^2 + ||Rx||^2 \tag{2.17}$$

On en déduit que minimiser l'énergie du résidu Rx revient à maximiser le module du produit scalaire $\langle x, g_\lambda \rangle$.

Si le dictionnaire est complet, c'est-à-dire qu'il contient au moins une base, la convergence de l'énergie du résidu vers 0 est assurée. L'algorithme est qualifié de "glouton" : il consiste à sélectionner des atomes optimaux localement, mais l'ensemble des atomes sélectionnés associés à leur poids n'est pas garanti comme optimal pour représenter le signal.

On peut noter que la première application de cet algorithme par Mallat & Zhang (1993) concernait des atomes de Gabor de résolutions différentes pour l'analyse de signaux de parole, afin d'en obtenir une représentation temps-fréquence mieux définie que les spectrogrammes et sans les termes d'interférence de la transformée de Wigner-Ville.

2.2.5.3 Les variantes de l'algorithme de Matching Pursuit

L'algorithme de Matching Pursuit possède de nombreuses variantes. Le *Matching Pursuit Orthogonal* (Pati et al. (1993)) en est une permettant d'optimiser les poids des atomes sélectionnés. Il consiste à effectuer une projection orthogonale du signal original sur les atomes déjà sélectionnés à chaque itération de l'algorithme. L'intérêt apparaît donc lorsque les atomes du dictionnaire ne sont pas orthogonaux. Cependant, l'orthogonalisation demande des calculs supplémentaires qui peuvent le rendre prohibitif pour certaines applications.

On peut également mentionner l'algorithme de Matching Pursuit Haute-Résolution (Gribonval et al., 1996), qui consiste à remplacer le produit scalaire par une autre mesure de corrélation permettant d'éviter les phénomènes de pré-écho inhérents au Matching Pursuit avec des fenêtres longues.

D'autres variantes portent sur la sélection de l'atome à chaque itération, en s'appuyant sur le résultat de Temlyakov (2000) concernant le Matching Pursuit *faible*. En

sélectionnant un atome qui n'est pas optimal, la convergence vers zéro de l'énergie du résiduel peut en effet être assurée sous la condition suivante :

$$\exists \rho, 0 < \rho < 1 \quad | \quad \forall n, \quad |\langle x, g_{n sel} \rangle| \geqslant \rho |\langle x, g_{n opt} \rangle| \tag{2.18}$$

où ρ est un paramètre qui caractérise la sous-optimalité, $g_{n opt}$ est l'atome optimal et $g_{n sel}$ l'atome sélectionné.

En s'appuyant sur ce résultat, on peut alors élaborer des stratégies sous-optimales mais plus rapides, qui permettront d'obtenir des décompositions néanmoins utilisables.

Gribonval (2001) introduit un algorithme permettant d'extraire rapidement des atomes de Gabor modulés en fréquence (2.2.4.3). Au lieu de construire des dictionnaires d'atomes de Gabor chirpés en échantillonnant le chirp rate c, une estimation du meilleur c est effectuée pour chaque atome d'un dictionnaire de Gabor standard.

Enfin on peut mentionner qu'il est possible d'introduire de l'information a priori dans les décompositions en effectuant une pondération des produits scalaires entre les atomes et le signal : par exemple, si l'atome a peu de chance a priori d'être sélectionné, il peut être normé à une valeur inférieure à 1. Le comportement de tels algorithmes de Matching Pursuit *Pondéré* a été étudié par Escoda et al. (2006).

2.2.5.4 Cohérence d'un dictionnaire

Il est nécessaire de souligner que l' approximation du signal à l'aide d'un algorithme de décomposition peut être plus ou moins difficile selon le dictionnaire utilisé. En particulier, les corrélations entre les atomes du dictionnaire joue un rôle crucial dans la facilité qu'on aura à approcher un signal à l'aide de ce dictionnaire. Dans le cas extrême ou tous les atomes du dictionnaire sont orthogonaux, un simple seuillage des projections du signal sur les atomes permet d'obtenir la représentation la plus parcimonieuse possible sur un critère de nombre d'atomes ou de rapport signal à bruit. En pratique, les atomes sont rarement orthogonaux deux à deux. On définit alors la cohérence d'un dictionnaire comme le maximum entre les produits scalaires de deux atomes d'un dictionnaire :

$$\mu = \max_{i \neq j} \{ |\langle h_i, h_j \rangle| \, | (h_i, h_j) \in \mathcal{D}^2 \} \tag{2.19}$$

Une valeur faible de cohérence indique donc que les atomes sont fort orthogonaux entre eux. Des mesures plus informatives peuvent être utilisées, comme la cohérence cumulative ou le spark (Tropp (2004)).

Connaître la cohérence d'un dictionnaire permet de déterminer certains comportements des algorithmes d'approximation. Dans le cas du Matching Pursuit, la convergence est exponentielle lorsque le dictionnaire est quasi-incohérent. Lorsque le signal est la somme exacte d'atomes du dictionnaire, les algorithmes de Matching Pursuit Orthogonal et de Basis Pursuit permettent de retrouver la combinaison d'atomes utilisée par le signal sous certaines conditions sur la cohérence du dictionnaire.

Si le dictionnaire est très cohérent, l'algorithme de Matching Pursuit présente une certaine faiblesse : la soustraction d'un atome modifie les projections du signal sur les atomes qui lui sont corrélés. Nous verrons dans ce document que cet aspect est assez gênant, notamment lorsque qu'il s'agit d'identifier plusieurs sources instrumentales à un instant donné.

2.2.5.5 Matching Pursuit Moléculaire

Etant donnée une molécule, on définit son poids $\alpha_{\mathcal{M}}$ par rapport au signal en calculant la norme de la projection orthogonale du signal sur le sous-espace vectoriel engendré par les atomes :

$$\alpha_{\mathcal{M}} = ||\mathcal{P}_{Vec(g_\lambda, \lambda \in \mathcal{M})} x||_2 \tag{2.20}$$

Si les atomes sont orthogonaux, on a :

$$||\mathcal{P}_{Vec(g_\lambda)} x||_2^2 = \sum_{\lambda \in \mathcal{M}} ||\mathcal{P}_{(g_\lambda)} x||_2^2 = \sum_{\lambda \in \mathcal{M}} |\langle x, g_\lambda \rangle|^2 \tag{2.21}$$

Cette propriété est intéressante en pratique : elle permet de calculer directement le poids d'une molécule en fonction des poids des atomes qui la composent.

D'autres cas permettent aussi un calcul direct des poids a posteriori des molécules. Dans le cas de molécules diatomiques composées d'atomes complexes conjugués, on peut sélectionner la meilleure molécule diatomique à partir des poids des atomes complexes qui la composent. Cette propriété est utilisée très souvent en pratique, notamment lorsqu'on utilise la Transformée de Fourier Rapide qui permet de calculer efficacement des produits scalaires complexes.

L'algorithme de Matching Pursuit a aussi été modifié afin de traiter des dictionnaires de molécules plus complexes que la molécule diatomique d'atomes de Gabor conjugués. Gribonval & Bacry (2003) présentent un algorithme permettant d'extraire des molécules harmoniques (2.2.4.5). En définissant des sous-espaces harmoniques à partir de fréquences d'atomes de Gabor, et en supposant que les vecteurs qui le définissent (les atomes de Gabor représentant les partiels) sont orthogonaux, on peut sélectionner le meilleur sous-espace harmonique en sélectionnant la meilleure somme des poids des atomes qui la composent au carré :

$$||\mathcal{P}_{Vec(g_\lambda)} x||_2^2 \simeq \sum_{\lambda \in \mathcal{M}} |\langle x, g_\lambda \rangle|^2 \tag{2.22}$$

Dans (Daudet, 2006), des molécules tonales et transitoires sont définies. Les molécules tonales sont construites à partir de cosinus locaux, et les molécules transitoires à partir d'ondelettes dyadiques (Haar). Afin de former les molécules tonales, les poids des molécules sont calculés en effectuant des sommations de coefficients régularisés selon le temps, et sont tronquées a posteriori par seuillage sur ces coefficients. Concernant les molécules transitoires, le poids des molécules est cette fois-ci une somme des coefficients de la transformée selon les branches de la réprésentation (un coefficient par échelle), et sont tronqués là aussi par seuillage sur les coefficients.

Nous voyons donc que la recherche des molécules se base souvent sur des approximations qui permettent de sélectionner une bonne molécule à partir des projections du signal sur les atomes qui pourraient la composer. Dans ce cas, le choix de la molécule optimale n'est possible que si les atomes sont orthogonaux.

2.3 Bilan

Les approches utilisées en indexation audio s'appuient essentillement sur trois modèles :

– les approches *sac de trames*, basées sur l'extraction de caractéristiques pertinentes pour modéliser ou discriminer les classes à identifier ;

- les approches par *modèles de signaux*, qui modélisent explicitement les sources et/ou le mélange des sources mises en jeu ;
- les approches par *représentations mi-niveau*, qui permettent d'obtenir des représentations plus structurées que celles par caractéristiques mais moins précises que celles désirées pour l'approche par modèle de signaux.

Concernant les représentations parcimonieuses, nous avons vu que de nombreuses formes d'ondes ont été développées, que ce soit pour les parties tenues des notes ou pour les transitoires rapides. On remarquera que les parties bruitées ne sont pas modélisées explicitement dans les approches de représentations parcimonieuses. Le cadre de travail qu'elles introduisent donne une flexibilité qui, à condition de définir les dictionnaires et les algorithmes adéquats, permettent d'envisager des représentations mi-niveau intéressantes.

Dans notre étude, nous allons nous attarder sur le cas des sons contenant des notes possédant une structure harmonique, ou quasi-harmonique. En effet, une majeure partie des sons en musique est constituée de notes possédant une forte structure harmonique. Comme nous le verrons dans la suite du document, nous nous baserons donc sur des atomes harmoniques ou quasi-harmoniques, éventuellement modulés en fréquence, et dont les amplitudes des partiels nous permettront d'introduire de la connaissance sur les sources considérées. Cette information permettra de rendre moins ambigus certains mélanges, mais aussi d'avoir une information sur le timbre des instruments intervenant dans les signaux dès la fin de la décomposition. Chaque élément de la décomposition portera donc une part d'information haut-niveau, ce qui reviendra à obtenir une description mi-niveau du son. Á partir de la représentation, d'autres algorithmes spécialisés dans le traitement des données symboliques pourront être appliqués directement pour déterminer les lignes mélodiques, les instruments mis en jeu, la tonalité, la similarité etc.

Chapitre 3

Modèles de signaux

Dans le chapitre précédent, nous avons vu que le domaine des représentations parcimonieuses pouvait aider à rendre plus explicite le signal musical, pourvu que le dictionnaire d'atomes, le critère d'optimalité et l'algorithme qui y conduisent soient bien choisis. Dans cette partie, nous allons définir les atomes et les molécules conçus pour analyser les signaux musicaux. Ils sont formellement proches d'atomes introduits dans des études précédentes, leur originalité consistant essentiellement en l'inclusion de connaissances sur les sources potentiellement présentes dans le signal. Ces connaissances portent sur l'enveloppe spectrale des sources. Nous verrons que plusieurs variantes de ces atomes peuvent être définies, ainsi que des molécules composées de ces atomes. Nous montrerons que la projection du signal sur ces atomes peut être mise en relation avec des opérations effectuées dans des tâches d'estimation de paramètres du signal musical comme l'identification d'instruments et l'estimation de pitch.

3.1 Les atomes Notes : les atomes idéaux

Comme nous l'avons vu dans le chapitre 2.1.2, nous ferons l'hypothèse que le signal est le résultat d'un mélange instantané de notes, en considérant que tous les traitements ultérieurs, tels que les effets de salle, sont inclus dans le modèle de note.

Dans notre cas, les atomes idéaux seront l'espace de toutes les notes pouvant être produites par les instruments possibles, avec tous les instruments possibles et suivi de tous les traitements possibles. Etant donné un certain nombre de paramètres qui permettent de synthétiser ces atomes, on peut alors penser les échantillonner afin d'obtenir un dictionnaire représentatif de toutes les notes possibles. Cependant, nous ne pouvons pas réaliser un tel travail : échantillonner toute la variabilité des notes de musique et des traitements ultérieurs n'est pas envisageable. En effet, les paramètres qui peuvent varier sont la durée de la note, la hauteur de note, les modulations de fréquence et d'amplitudes, l'intensité, des paramètres d'enveloppe spectrale etc.

Cela nous conduit donc à adopter une autre stratégie : les atomes retenus seront à une granularité temporelle plus fine que la note. Les paramètres de fréquence fondamentale et de localisation temporelle seront échantillonnés. Les amplitudes des partiels caractériseront les spectres pouvant être produits, et seront échantillonnées en les contraignant à appartenir à un dictionnaire de vecteurs d'amplitudes (ou d'enveloppes). Un paramètre d'échelle temporelle pourra être soit échantillonné sur un faible nombre d'échelles différentes, soit unique, mais dans ce cas la durée des notes sera calculée de

manière *ad hoc*, en utilisant des algorithmes moléculaires. Ces algorithmes auront pour but de grouper les atomes en notes, et ceci dès l'extraction.

Revenons sur l'hypothèse du mélange instantané considérée dans notre étude. Cette hypothèse aura une conséquence sur l'étude des mélanges stéréo. Si le signal stéréo est issu d'un mélange instantané de sources fixes sur l'axe du panoramique, les atomes intéressants à extraire seront simplement situés à l'endroit où les sources ont été positionées. Cependant, si le signal est issu d'un mélange convolutif, cette hypothèse implique que les localisations des atomes ne seront plus strictement à un endroit précis, mais plutôt distribuées autour de la position spatiale de la source originale. Retrouver des atomes situés à un endroit précis demanderait d'estimer la matrice convolutive entre les sources et les capteurs, c'est-à-dire de faire l'hypothèse d'un mélange convolutif. Nous examinerons les conséquences de l'hypothèse de mélange instantané sur de tels signaux dans la section 6.4.

3.2 Conventions

Tout d'abord, il est nécessaire de mentionner que, comme dans (Mallat & Zhang, 1993; Gribonval, 1999), nous n'emploierons que des atomes complexes. Cette notation fait porter les ajustements de phase sur le poids des atomes, et permet ainsi de ne pas échantillonner ce paramètre de phase.

Pour analyser un signal réel avec des atomes complexes, il suffit de les combiner par paires d'atomes conjugués (ou molécules di-atomiques), comme l'ont proposé Mallat & Zhang (1993); Goodwin & Vetterli (1999). Dans ce cas, la molécule optimale peut être calculée en prenant le maximum du module des produits scalaires entre le signal et un des atomes complexes.

3.3 Atomes Harmoniques Spécifiques à des Instruments

Dans le cadre de travail que nous nous sommes fixé, nous abordons les signaux contenant des sons auquels on peut attribuer une hauteur. Il est alors naturel d'utiliser des atomes harmoniques ou quasi-harmoniques. Ils permettront ainsi de modéliser des sons entretenus, à condition qu'ils puissent s'adapter aux modulations de fréquences et d'amplitudes qui peuvent se produire, ou alors produits à partir de systèmes en oscillation libre comme les cordes frappées ou pincées, s'ils prennent en compte le décalage éventuel des partiels qui en est la conséquence. Comme mentionné dans la partie précédente, Gribonval & Bacry (2003) ont introduit ce type d'atomes, ainsi qu'un algorithme afin de les extraire. Le calcul des coefficients des partiels par projection du signal sur le sous-espace formé par ces partiels permet d'obtenir les coefficients *quasi*-optimaux en terme de Rapport Signal-à-Résiduel (RSR). En effet, comme les atomes de partiels ne peuvent être considérés comme strictement orthogonaux, les projections du signal sur les atomes ne permettent pas de calculer directement la projection sur le sous-espace qu'ils forment, ce qui empêche de connaître quel sous-espace est optimal.

Cependant, si cette approche permet d'enlever les structures harmoniques quasi-optimales, elle ne tient pas compte de l'enveloppe spectrale de ces structures. Or des *a priori* sur l'enveloppe spectrale peuvent être utiles afin de rendre le mélange moins ambigu, ainsi que pour l'estimation de fréquences fondamentales (simples ou multiples) comme l'ont montré certaines méthodes basées sur la somme spectrale ((Emiya et al.,

2007) pour le piano, (Klapuri, 2006) dans le cas général). En effet, comme nous le soulignerons en 4.1.1, nous chercherons prioritairement à extraire des atomes dont la hauteur correspond à celle des notes jouées.

Nous introduisons donc un nouveau vecteur de paramètres pour les atomes harmoniques. Ainsi, contrairement au Matching Pursuit Harmonique, ce n'est pas le peigne optimal du signal qui sera choisi, mais le peigne dont les amplitudes de partiels seront les plus corrélées avec des enveloppes spectrales appartenant à un dictionnaire.

Supposons un vecteur d'amplitude de partiels $A = (a_m)_{m=1..M}$, un atome Harmonique Spécifique à un Instrument (HSI) s'écrit alors de la façon suivante :

$$h_{s,u,f_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m.f_0}(t) \tag{3.1}$$

Si l'on suppose que les partiels sont orthogonaux, on obtient un atome $h_{s,u,f_0,A,\Phi}$ normé à 1 sous la condition $\sum_{m=1}^{M} a_m^2 = 1$. Cette hypothèse est d'autant plus valide que le produit $f_0 \times s$ est élevé. Par exemple, pour une taille de fenêtre de $s = 46$ ms et la fréquence fondamentale f_0 la plus basse du violoncelle (65 Hz), le produit scalaire entre deux atomes de partiels successifs est de $|\langle g_{s,u,m\times f_0}, g_{s,u,(m+1)\times f_0}\rangle| = 0.026$. En pratique, les dictionnaires que nous construirons contiendront uniquement des atomes dont le produit $f_0 \times s$ sera au-dessus d'une valeur minimale : certaines fréquences fondamentales basses ne seront représentées que par des atomes d'échelles plus longues.

Le vecteur A permet une paramétrisation du dictionnaire selon une caractéristique importante pour la définition du timbre : l'enveloppe spectrale. En apprenant des vecteurs caractéristiques d'un instrument ou d'un groupe d'instruments, on introduit de la connaissance sur les sources sonores au niveau du dictionnaire, qui permettra donc d'obtenir une représentation mi-niveau du signal pertinente dans un contexte où l'on connait les instruments pouvant être mis en jeu dans le signal musical.

Afin d'avoir un dictionnaire d'amplitudes représentatives des timbres des instruments, on apprendra des vecteurs A spécifiques à la fois aux notes jouées et à l'instrument. Dans la suite, on distinguera la hauteur de note p de la fréquence fondamentale f_0 : p est un index entier correspondant au code MIDI de la note jouée, telle qu'elle serait écrite sur une partition, et f_0 est un paramètre continu exprimé en Hz. Ainsi, on obtient la hauteur de note p correspondant à une fréquence fondamentale f_0 grâce à la formule $p_{f_0} = \text{arrondi}(69 + 12.\log_2(f_0/440))$. On décomposera donc le dictionnaire de vecteurs d'amplitudes $\mathcal{A}$ en sous-dictionnaires d'amplitudes $\mathcal{C}_{ip}$, qui contiendront chacun des vecteurs d'amplitudes $(A_{ipk})_{k=1..K}$ qui peuvent être joués par l'instrument i à la note p. Dans la suite, on dira par abus qu'un atome h est dans $\mathcal{C}_{ip}$ si le vecteur d'amplitude à partir duquel il est généré appartient à $\mathcal{C}_{ip}$.

La Figure 3.1 présente des exemples de tels atomes pour différents instruments. Nous verrons dans le chapitre 5.2 comment générer les ensembles $\mathcal{C}_{ip}$.

3.4 Atomes de Gabor chirpés harmoniques avec enveloppe déterminée

Pour certains sons harmoniques, notamment les instruments à sons entretenus, des modulations de fréquence peuvent être observées au niveau des partiels. Une propriété de ces modulations est qu'elles se produisent de façon quasi-synchrones entre les partiels.

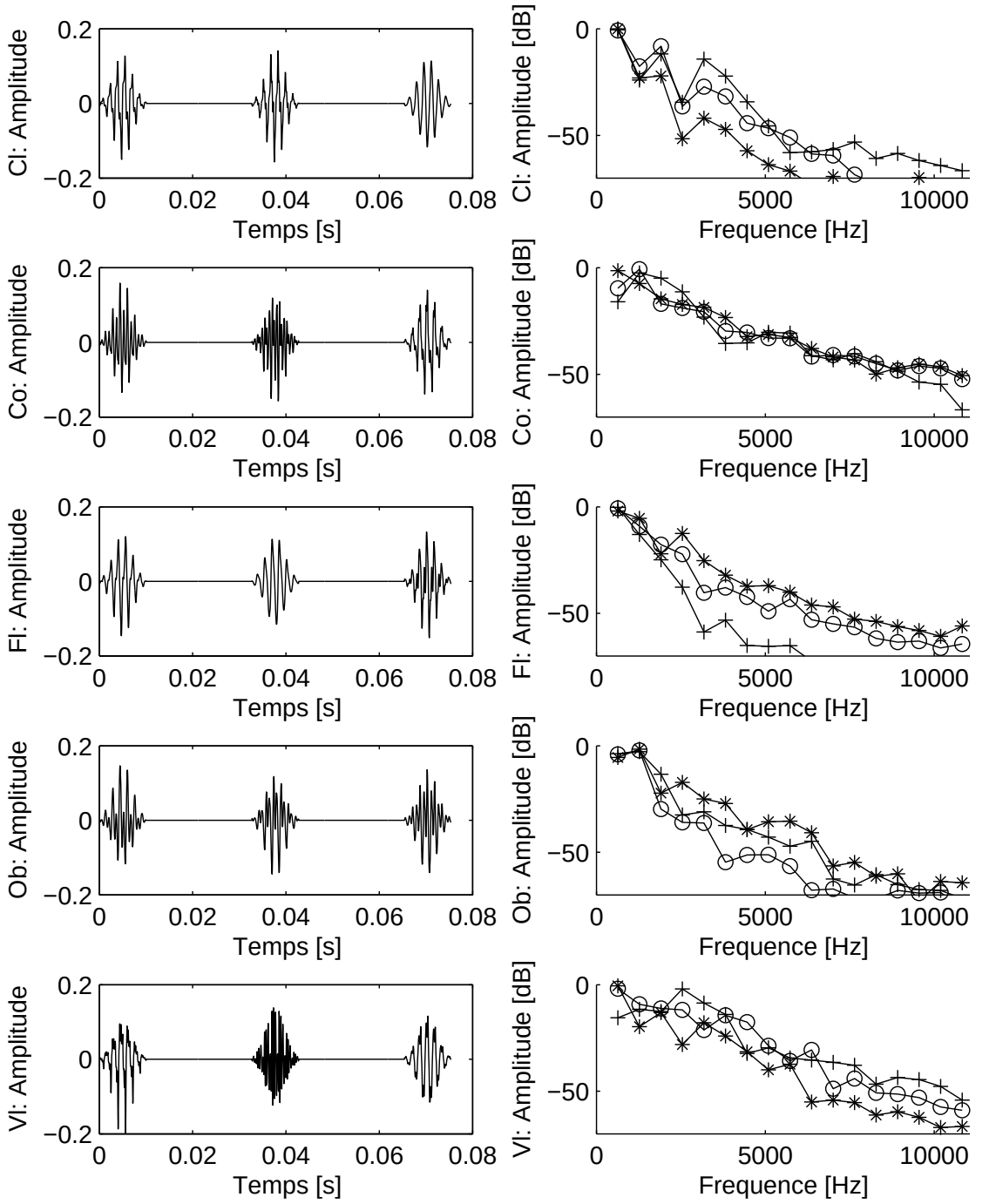


FIG. 3.1 – *Atomes Harmoniques Spécifiques à un Instrument pour cinq instruments : Clarinette (Cl), Violoncelle (Co), Flute (Fl), Hautbois (Ob), Violon (Vl), pour une note à 650 Hz. Les amplitudes des partiels ont été extraits sur des sources différentes. Gauche : formes temporelles, Droite : amplitudes des partiels. Sur les figures de droite, les courbes marquées par des o, des * et des + correspondent respectivement aux formes d'ondes sur les figures de gauche prises de gauche à droite.*

Il est alors naturel d'introduire des atomes Harmoniques Spécifiques à des Instruments *Chirpés* (atomes HSIC) :

$$h_{s,u,f_0,c_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m.f_0,m.c_0}(t) \tag{3.2}$$

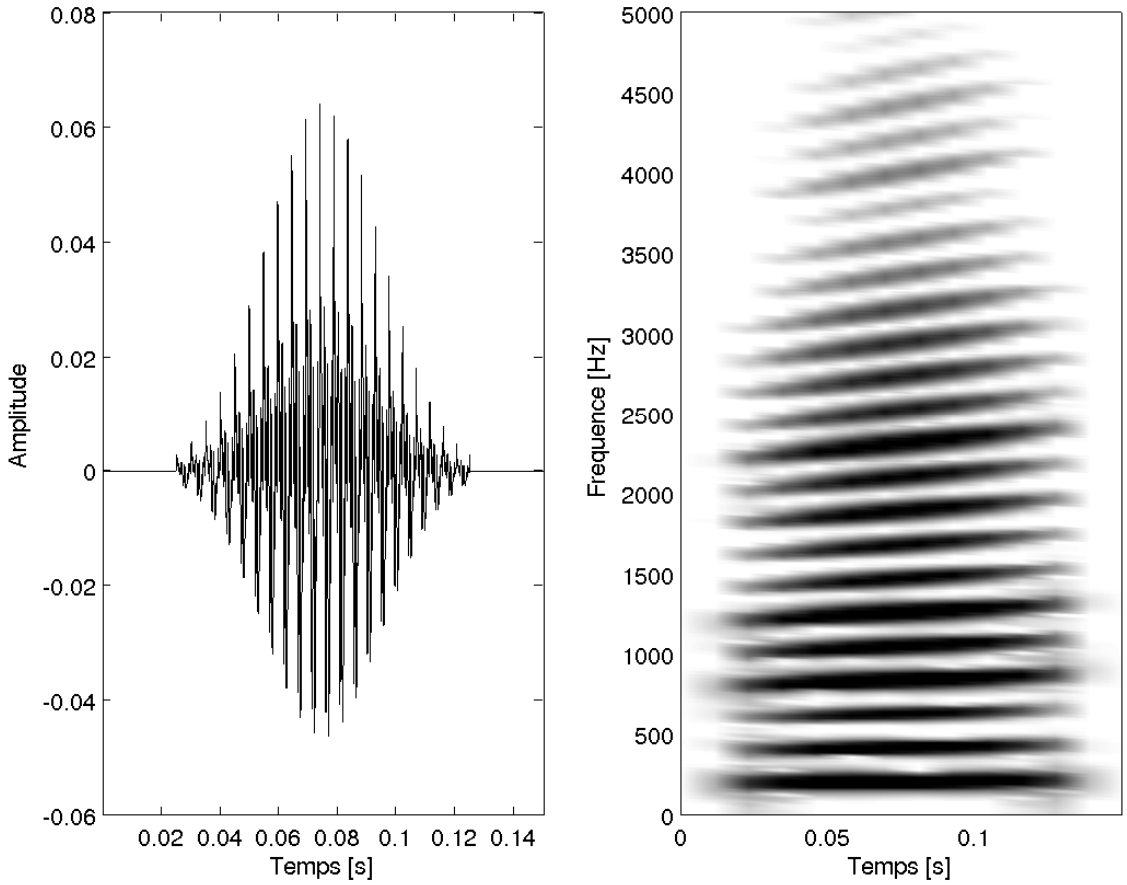


FIG. 3.2 – *Atome Harmonique Spécifique à un Instrument Chirpé. Gauche : représentation temporelle, Droite : spectrogramme*

Les taux de modulation des partiels ("taux de *chirp*") sont proportionnels au taux de modulation fondamental c_0, afin de garantir les rapports harmoniques entre les fréquences instantanées des partiels.

3.5 Atomes de Gabor légèrement inharmoniques avec enveloppe déterminée

Un autre type d'atome peut s'adapter convenablement aux sources présentant une légère inharmonicité, comme cela peut se produire pour les instruments mettant en jeu

des cordes tendues (piano, guitare). En effet, dans ce cas, la déviation de la fréquence des partiels par rapport à un spectre purement harmonique entraine des confusions dans l'estimation de leur énergie (voir Figure 3.3). Dans ce cas de figure, on utilise un paramètre β, qui peut être calculé en fonction des paramètres physiques de la corde, dont on déduit la position des partiels. On a alors :

$$h_{s,u,f_0,\beta,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m.\sqrt{1+\beta(m^2-1)}.f_0}(t) \tag{3.3}$$

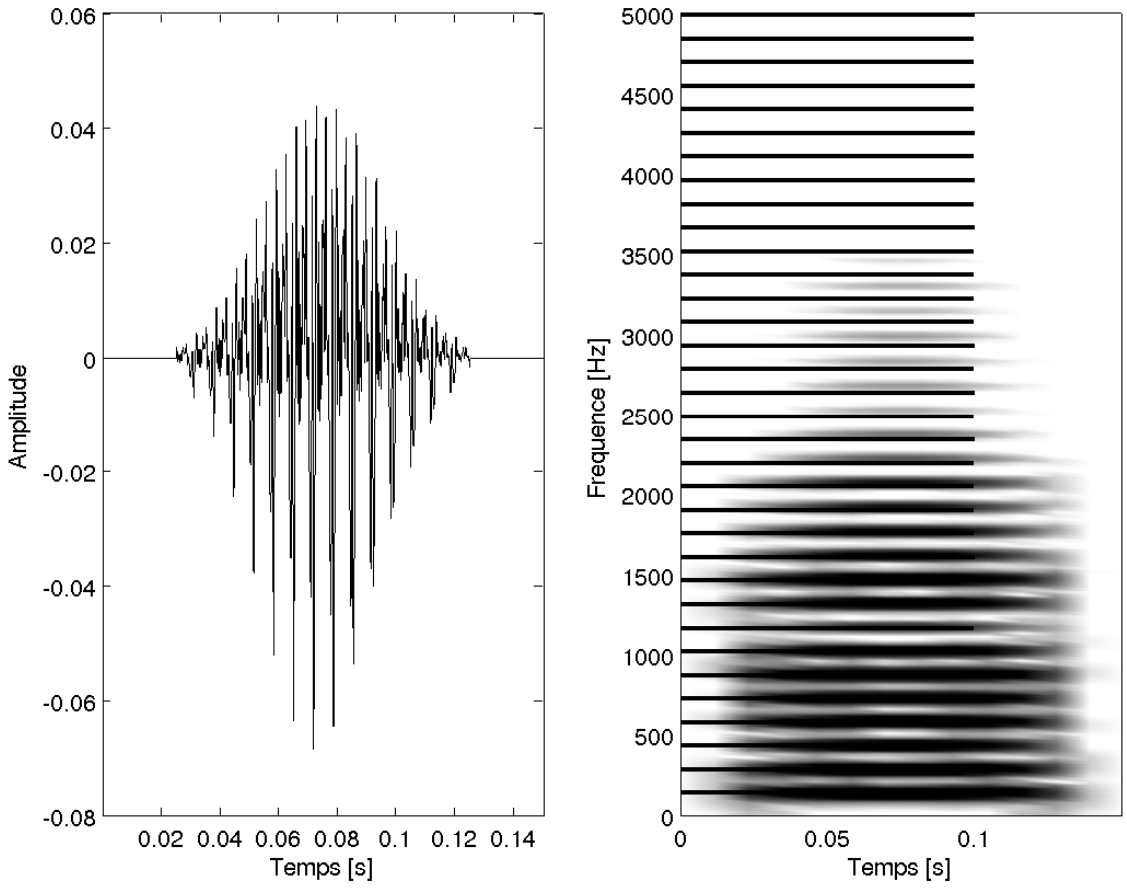


FIG. 3.3 – *Atome Inharmonique Spécifique à un Instrument. Gauche : représentation temporelle, Droite : spectrogramme (les lignes verticales représentent les positions des partiels du peigne harmonique accordé sur sur f_0, c'est-à-dire pour $\beta = 0$).*

L'utilisation de cette modélisation de l'inharmonicité a été utilisée avec succès par Emiya et al. (2007) pour l'estimation de hauteur de note sur le piano.

Dans la suite, on fera référence aux atomes Inharmoniques Spécifiques à un Instrument (IHSI) pour ce modèle.

3.6 Version stéréo

Tous les atomes décrits ci-dessus peuvent être dérivés en atomes stéréophoniques. Par exemple, pour l'atome harmonique décrit en 3.3, on peut écrire l'atome stéréo :

$$h_{st_{s,u,f_0,A,\Phi,\theta}}(t) = [\cos(\theta)h_{s,u,f_0,A,\Phi}(t) \quad \sin(\theta)h_{s,u,f_0,A,\Phi}(t-\tau)] \tag{3.4}$$

Le paramètre τ, bien qu'il soit pertinent pour la validité physique du modèle, peut souvent être négligé quand on le compare à l'échantillonnage temporel du dictionnaire de période Δu : τ est le plus souvent de l'ordre de la milliseconde. Il est également difficile d'obtenir des estimations robustes de ce paramètre. De plus, les atomes que nous avons définis ont pour but de modéliser les parties quasi-stationnaires des signaux. Cela implique que, la plupart du temps, les sources sont activées en même temps sur les deux canaux, en étant simplement déphasées. Ainsi, pour des raisons d'adaptation au signal, nous avons préféré différencier le vecteur des phases selon chacun des canaux :

$$h_{st_{s,u,f_0,A,\Phi,\theta}}(t) = [\cos(\theta)h_{s,u,f_0,A,\Phi_1}(t) \quad \sin(\theta)h_{s,u,f_0,A,\Phi_2}(t)] \tag{3.5}$$

où Φ_1 et Φ_2 sont les vecteurs des phases respectifs sur chacun des deux canaux.

On peut penser à introduire d'autres degrés de liberté dans ce modèle, par exemple attribuer un paramètre de panoramique à chacun des partiels. Cela permettrait de tenir compte de la variabilité de la directivité des sources selon les fréquences. Nous n'étudierons pas cette modélisation dans la suite, nous garderons donc une localisation unique des partiels sur l'axe stéréo.

3.7 Molécules

Comme évoqué en 2.2.3, une molécule peut être considérée comme un atome. Ici on appellera molécule la structure de plus haut niveau de notre modèle, c'est-à-dire celle qui regroupe des atomes de l'un des trois types mentionnés ci-dessus. Dans notre cas, étant donné que les atomes capturent une grande partie de la structure fréquentielle d'une note, la molécule consistera à introduire une dimension temporelle "long-terme" (à l'échelle de la note) dans nos structures de signaux. Ici, on ne construira pas de dictionnaires d'enveloppes temporelles, bien que ce problème mériterait qu'on y porte de l'attention étant donné les caractéristiques propres de certains instruments à cette échelle (vibrato, tremolo, décroissances d'amplitude...). Dans ce cas, on permettrait une granularité en atomes plus proche encore de la note, afin de tenir compte des variations de paramètres dans la zone 4-10 Hz. Dans notre cas, afin de tenir compte de ces phénomènes pour la classification, on pourra néanmoins extraire des caractéristiques sur ces molécules une fois la décomposition effectuée.

Les molécules seront donc formées de façon *ad hoc* (comme dans le Matching Pursuit Harmonique dans le domaine spectral), c'est-à-dire qu'on les construira afin qu'elles correspondent au mieux au signal en maximisant la projection orthogonale du signal sur les atomes qui composent la molécule.

On posera les contraintes suivantes pour la construction des molécules d'atomes HSI, afin d'obtenir des molécules ressemblant à celles présentées sur la Figure 3.4 :

- les atomes couvrent des plages de localisations successives u, avec exactement un atome par localisation,
- tous les atomes viennent d'un seul instrument i,

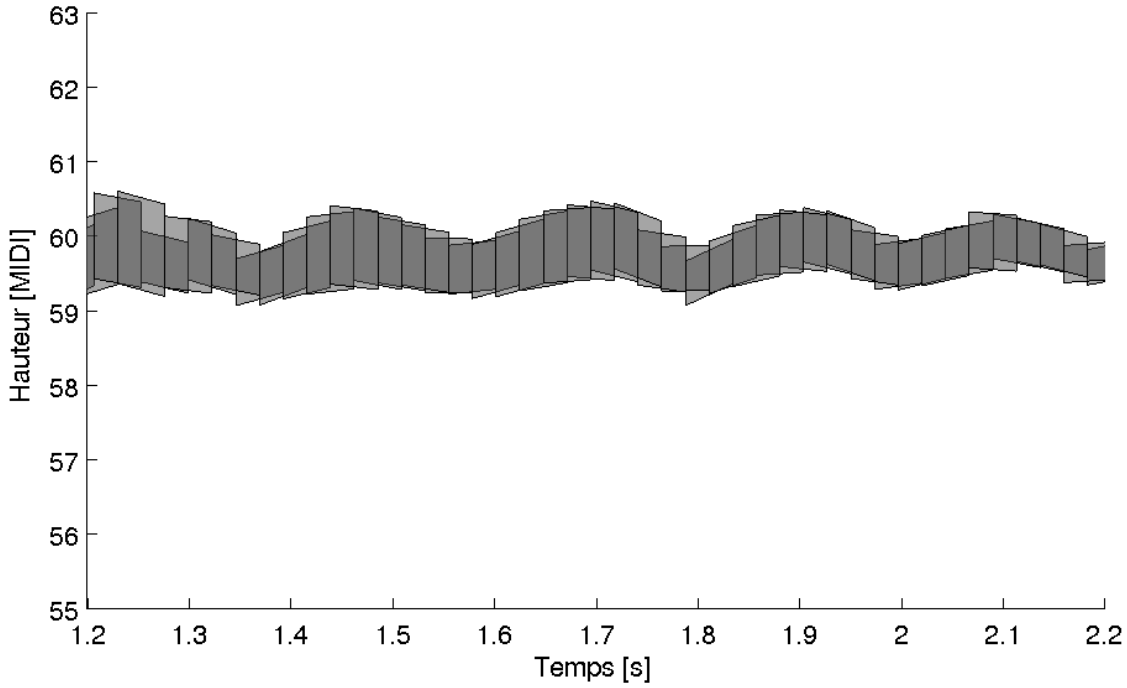


FIG. 3.4 – *Molécule composée d'atomes HSI de flûte. Chaque parallélogramme représente un atome, la dimension verticale correspond à leur amplitude et l'inclinaison est proportionnelle au taux de modulation.*

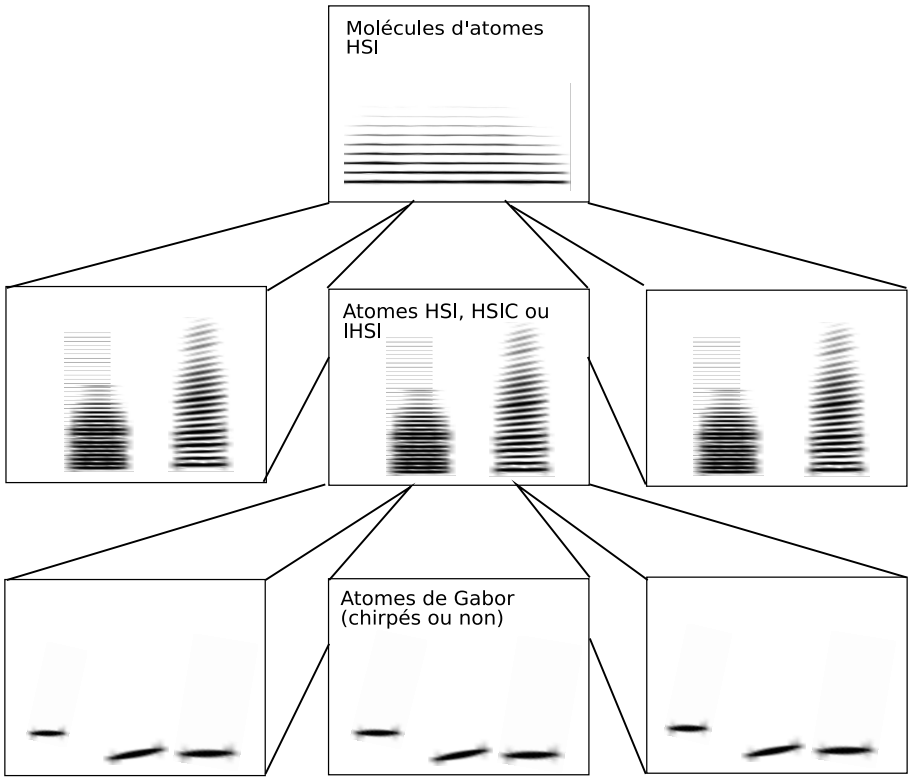


FIG. 3.5 – *Hiérarchie du modèle de signal*

– la log-variation de la fréquence fondamentale entre deux atomes consécutifs est bornée par un seuil D :

$$|\Delta \log f_0| \leq D. \tag{3.6}$$

Une fois ce modèle posé, on peut représenter la hiérarchie entre les différentes structures de modèles évoquées jusqu'à présent sur la Figure 3.5.

3.8 Interprétation des produits scalaires

Il s'agira dans la suite de sélectionner les atomes h_λ qui sont très corrélés avec le signal, c'est-à-dire avec des modules de produits scalaires $|\langle x, h_\lambda \rangle|$ élevés. Dans ce paragraphe, nous allons voir en quoi le produit scalaire entre un signal et un atome HSI peut être mis en relation avec des techniques utilisées en indexation audio, notamment en estimation de hauteur de note et en classification.

Examinons le produit scalaire entre un signal x et un atome HSI h :

$$\langle x, h_\lambda \rangle = \left\langle x, \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m\times f_0,0} \right\rangle \tag{3.7}$$

$$= \sum_{m=1}^{M} a_m \, e^{-j\phi_m} \langle x, g_{s,u,m\times f_0,0} \rangle \tag{3.8}$$

Comme nous le verrons par la suite, les phases des partiels seront calculées de telle sorte

qu'elles soient adaptées au signal :

$$e^{j\phi_m} = \frac{\langle x, g_{s,u,m\times f_0,m\times c_0}\rangle}{|\langle x, g_{s,u,m\times f_0,m\times c_0}\rangle|} \tag{3.9}$$

On a alors

$$\langle x, h_\lambda\rangle = \sum_{m=1}^{M} a_m \, |\langle x, g_{s,u,m\times f_0,0}\rangle| \tag{3.10}$$

Cette quantité, qui est donc un réel positif, peut donc être écrite grâce à un facteur de normalisation C :

$$\langle x, h_\lambda\rangle = C \sum_{m=1}^{M} a_m \, b_m \tag{3.11}$$

où

$$C = \left(\sum_{m=1}^{M} |\langle x, g_{s,u,m\times f_0,0}\rangle|^2 \right)^{1/2}, \tag{3.12}$$

$$b_m = \frac{1}{C} |\langle x, g_{s,u,m\times f_0,0}\rangle|. \tag{3.13}$$

En notant $B = \{b_m\}_{m=1\ldots M}$ le vecteur des amplitudes du signal normalisé ($\sum_{m=1}^{M} b_m^2 = 1$), on obtient finalement :

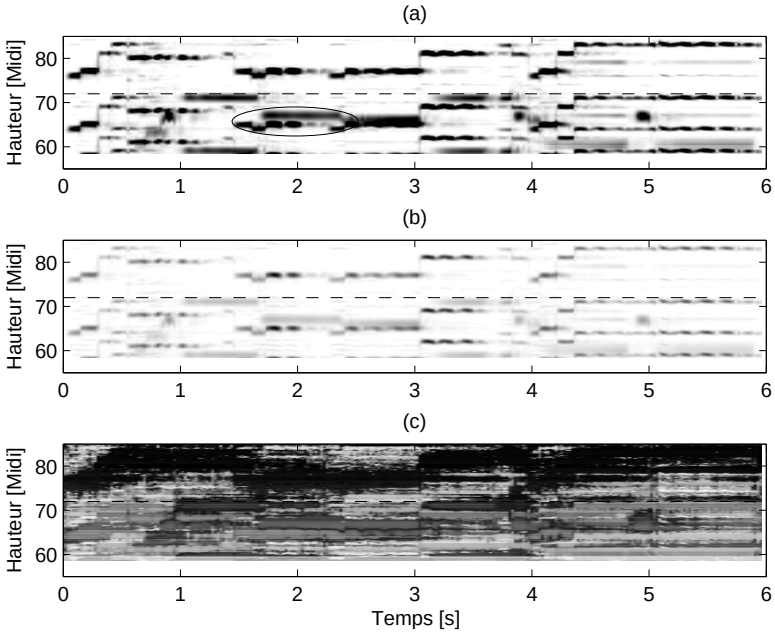
$$|\langle x, h\rangle| = C\langle A, B\rangle. \tag{3.14}$$

C est un facteur calculé par une somme spectrale sur des points du périodogramme correspondant aux harmoniques de f_0, et reste indépendant des vecteurs d'amplitudes A. Cette quantité peut être mise en relation avec une technique couramment utilisée en estimation de hauteur : un ensemble de valeurs C est calculé pour un certain nombre de fréquences fondamentales, puis la fréquence fondamentale pour laquelle C est le plus élevé détermine la hauteur de note. La valeur C est également celle qui permet de sélectionner le meilleur atome harmonique dans l'algorithme présenté par Gribonval & Bacry (2003).

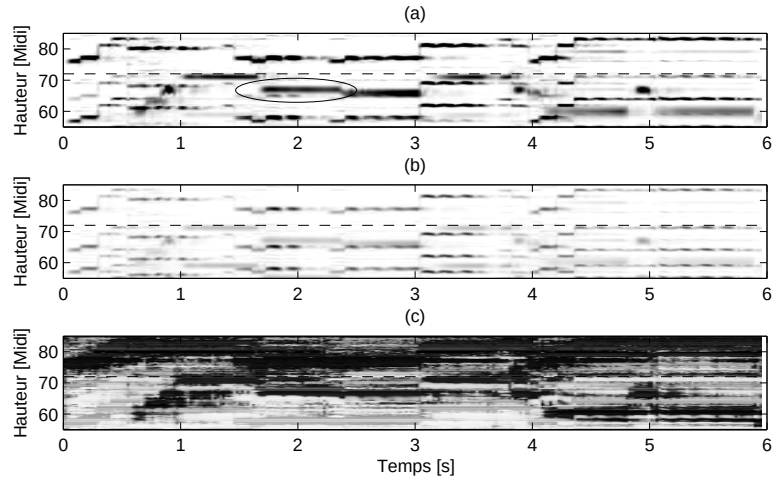
L'autre terme du produit $\langle A, B\rangle$ est un produit scalaire entre deux vecteurs d'amplitude normalisés, qui indiquent donc la similarité entre l'enveloppe du signal B et celles contenues dans le dictionnaires. Ce terme peut alors être vu comme une pondération de la somme spectrale représentée par le vecteur C, et devrait donc améliorer une estimation de hauteur basée uniquement sur l'énergie portée par les peignes harmoniques. De plus, ce produit porte de l'information permettant d'identifier l'instrument mis en jeu dans le signal : connaissant la hauteur d'une note p présente dans le signal, nous postulerons que la corrélation entre B et une enveloppe A d'une classe $\mathcal{C}_{i,p}$ sera liée à la vraisemblance que l'instrument i joue ladite note.

Au total, on remarquera donc que la maximisation produit scalaire entre le signal x et un atome HSI h est lié à la fois à l'estimation de la hauteur de note et à celle de la classe de l'instrument qui l'a produite.

La Figure 3.6 permet de visualiser les contributions respectives de la salience de fréquence fondamentale C et des similarités entre enveloppes spectrales $\langle A, B\rangle$ dans le produit scalaire entre le signal et les atomes. La représentation Piano Roll de la partition jouée (annotée manuellement) apparaît sur la Figure 3.7. On peut remarquer que la ligne mélodique de clarinette possède des projections plus faibles sur les atomes de flûte que la ligne mélodique de flûte.



(a) Flûte



(b) Clarinette

FIG. 3.6 – *Décomposition du produit scalaire entre le signal et les atomes HSI. La ligne mélodique du haut est jouée par la flûte, et celle du bas par la clarinette (la représentation MIDI est présentée sur la Figure suivante). Elles sont délimitées par une ligne pointillée. (a) : Produit scalaire total ($C.\langle A, B\rangle$), (b) : Somme spectrale (C, indépendante de l'instrument), (c) : Corrélation des enveloppes normalisées ($\langle A, B\rangle$, dépendant de l'instrument). Haut : Flûte, Bas :Clarinette. L'ellipse entoure une note de clarinette projetée sur les atomes HSI de flûte en haut, et de clarinette en bas. La note apparaît avec plus de contraste par projection sur les atomes de clarinette que par projection sur les atomes de flûte.*

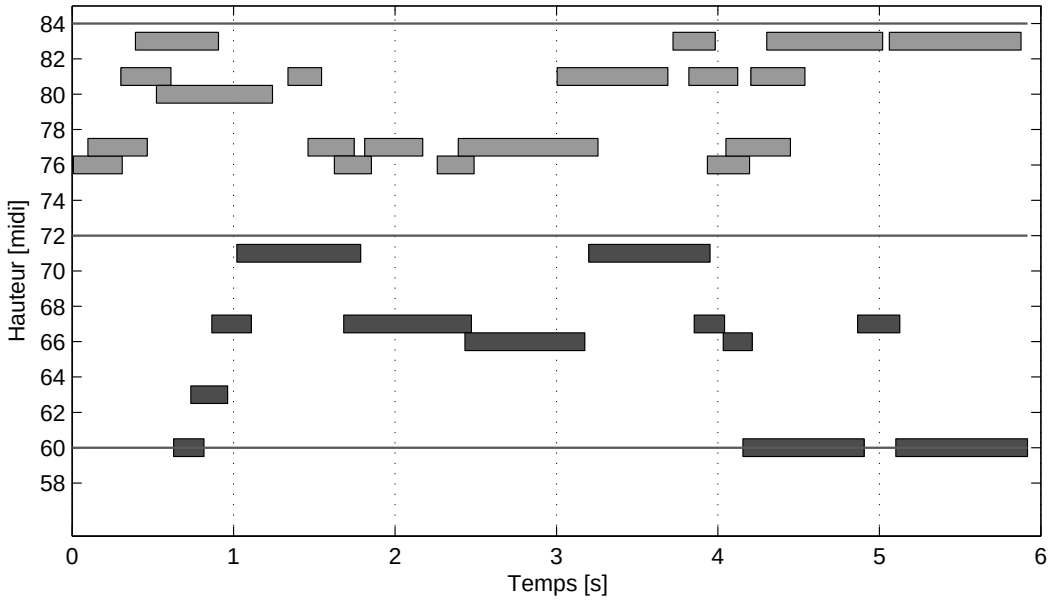


FIG. 3.7 – *Représentation "Piano-Roll" du duo Flûte-Clarinette analysé. La ligne mélodique du haut (au dessus du pitch 72) est jouée par la flûte, et celle du bas (en dessous de 72) par la clarinette.*

3.9 Bilan

Dans ce chapitre, nous avons défini des modèles de signaux permettant d'analyser des signaux musicaux composés de sources spécifiques. Les atomes Harmoniques Spécifiques à des Instruments (HSI) permettront dans une première approche d'identifier des peignes harmoniques spécifiques à des sources. Les atomes Harmoniques Spécifiques à un Instrument Chirpés (HSIC) offrent une adaptation possibles aux modulations de fréquences de certains instruments à son entretenu, et les atomes Inharmoniques Spécifiques à des Instruments (IHSI) permettent de tenir compte de l'inharmonicité faible de certains instruments comme le piano. Nous avons également défini des molécules, qui permettent de capturer les variations temporelles à long terme intervenant dans les notes de musique. Enfin nous avons indiqué qu'un produit scalaire de module élevé entre un atome HSI et un signal donnait une indication sur l'instrument et la note jouée dans le signal.

Les modèles développés sont ainsi restreints à des sources possédant des structures harmoniques fortes. Même en restreignant notre étude à ce type de source, les modèles développés ne permettent pas de représenter les signaux qu'elles peuvent produire de façon exhaustive. Nous ne traitons en effet que la partie sinusoïdale de ces sources, qui permet d'extraire une bonne partie de l'information utile à l'indexation. Elle porte en effet la totalité de l'information de hauteur de note, et une partie de l'information permettant d'identifier l'instrument produisant le signal. Les autres structures remarquables dans le signal, telles que la partie "bruit" et les transitoires rapides, ne seront pas modélisées. L'inconvénient principal de l'absence de ces modélisations est que ces structures seront modélisées par des atomes HSI, ce qui introduira des atomes d'erreur de modélisation dont il faudra se débarasser en post-traitement.

On peut néanmoins tracer quelques perpectives pour la prise en compte de ces structures. La partie stochastique stationnaire, bien qu'utile pour la discrimination entre les instruments, ne sera pas modélisée dans notre travail. Il serait cependant intéressant

de pouvoir en tenir compte dans le processus d'approximation du signal. Seulement, il apparaît compliqué de définir des "atomes de bruit" : si une modélisation dans le domaine spectral est possible et a déjà été utilisée dans différents travaux (Serra (1989)), on ne peut pas générer explicitement les formes d'ondes temporelles avec un nombre restreint de paramètres, ce qui empêche d'inclure de telles modélisation dans un algorithme faisant intervenir des analyse-synthèse successive comme l'algorithme de Matching Pursuit. Une approche permettant de contourner ce problème pourrait consister à combiner une analyse-synthèse effectuée sur le spectre tels que le filtrage de Wiener pour le bruit, et une autre dans le domaine temporel pour les autres atomes. Concernant les transitoires rapides, on pourrait donc ajouter au dictionnaire des sinusoïdes amorties, éventuellement en rapport harmonique ou quasi-harmonique, ou alors des molécules d'ondelettes dyadiques. Une représentation adaptée aux transitoires pourrait être utile, étant donné que ceux portent une partie importante de l'information permettant de discriminer les instruments, comme nous l'avons montré dans Essid et al. (2005). On peut également apprendre des formes d'ondes qui leur sont adaptées grâce aux méthodes d'apprentissage de dictionnaire.

Dans le chapitre suivant, nous allons décrire notre démarche pour extraire les atomes définis dans cette partie, puis certaines configurations de molécules.

Chapitre 4

Algorithmes

Dans le chapitre précédent, nous avons défini les modèles de signaux que nous allons utiliser. Dans ce chapitre, nous préciserons tout d'abord quels sont les objectifs d'une décomposition parcimonieuse avec les atomes et molécules que nous avons définis, en vue d'opérations d'indexation. Nous présenterons ensuite les algorithmes permettant de décomposer les signaux selon ces modèles. Les algorithmes sont dérivés de l'algorithme de Matching Pursuit : des modifications de cet algorithme permettront d'extraire les structures que nous avons définies en évitant la construction exhaustive des dictionnaires. Nous verrons enfin quelles perspectives ils offrent en vue d'objectifs applicatifs.

4.1 Discussion sur les objectifs des représentations

4.1.1 Atomes physiques, Atomes d'erreur de modélisation

En décomposant un signal avec les atomes présentés, les atomes (ou molécules) obtenus auront un rôle intéressant ou gênant suivant le rapport qu'ils entretiennent avec la réalité physique du jeu musical. On peut en effet postuler qu'un atome extrait est utile s'il correspond à une note ou une partie de note de même hauteur que la note jouée par un instrument, et que son support temporel est inclu entre le début et la fin de la note dans l'enregistrement. Si l'on considère une molécule telle qu'on l'a définie dans le chapitre précédent, la condition supplémentaire est que l'amplitude instantanée de la molécule soit proportionnelle à celle de la note jouée. Dans la suite, on les qualifiera d'atomes *physiques*.

Même s'ils améliorent le RSR (Rapport Signal à Résidu), les atomes extraits supplémentaires sont indésirables pour l'indexation audio, mais aussi, comme nous le verrons, pour le codage audio très-bas-débit. En effet, leurs paramètres ne permettent pas de décrire une source instrumentale. Les atomes supplémentaires peuvent être sélectionnés pour plusieurs raisons : ils peuvent compenser les erreurs de modélisation des vecteurs d'amplitude, la quantification de la fréquence fondamentale, les modulations de la fréquence instantanée et de l'amplitude, ou alors être extraits sur le bruit. Ces atomes seront appelés *atomes d'erreur de modélisation*. Nous essaierons d'éviter leur extraction, ou alors de les éliminer par un post-traitement de la décomposition.

4.1.2 Parcimonie temporelle, parcimonie en pitch

Etant donnés ces modèles de signaux, les techniques de décompositions parcimonieuses viseront à décomposer le signal de façon optimale avec ces atomes. L'optimalité est définie par un critère de parcimonie (2.2.2). Or, comme nous l'avons mentionné dans le paragraphe 2.2.1, rien ne garantit qu'il existe une fonction de coût dont le minimum indique que tout les atomes extraits seront des atomes physiques, excepté si une représentation très parcimonieuse peut être trouvée avec une fonction de coût donnée (Gribonval & Nielsen (2003)). Dans notre cas, pour que l'optimalité définie par la fonction de coût corresponde au mieux à l'explication physique du signal, nous règlerons la balance entre l'erreur de modélisation et la parcimonie grâce à une optimisation sur des ensembles de développement, au regard du critère qui quantifie la performance d'une (ou de plusieurs) application(s).

Sans pour autant chercher la transcription exacte du signal à l'aide d'atomes HSI, deux types de parcimonie seront donc recherchées, afin de d'obtenir une représentation facilement exploitable des signaux musicaux étudiés :

- La **parcimonie temporelle** : une décomposition présentant une parcimonie temporelle est une décomposition qui représentera des sources de durée variable avec peu d'éléments. Par conséquent, la parcimonie temporelle peut être considérée comme optimale si une note est représentée par exactement un atome. Par exemple, l'utilisation de dictionnaires multi-résolution permet de représenter des sources de durée variable, ainsi que des transitoires rapides si de très petites échelles sont utilisées. Echantillonner un grand nombre d'échelles possibles serait cependant très coûteux. On pourra néanmoins obtenir une bonne parcimonie temporelle si l'on utilise des molécules constituées de tels atomes.
- La **parcimonie en pitch** (ou hauteur de note) : étant donné un mélange de sources à un instant donné, une décomposition ayant une bonne parcimonie en hauteur représente efficacement un mélange de sources avec peu d'éléments. Ainsi, à un instant donné, la parcimonie en hauteur est considérée comme optimale si chacune des sources est représentée par exactement un atome.

Nous verrons que la parcimonie temporelle peut être approchée en utilisant un algorithme moléculaire avec les atomes proposés. La parcimonie en hauteur sera approchée en post-traitement, étant donnée la limitation de l'algorithme développé pour optimiser ce critère.

4.2 Recherche des meilleurs atomes

4.2.1 Matching Pursuit

Parmi les algorithmes présentés dans la section 2.2.5.1, nous avons choisi de nous baser sur l'algorithme de Matching Pursuit. Parmi les avantages qu'il présente pour notre problème, on peut citer sa rapidité, et sa flexibilité qui permet de chercher des atomes ou des sous-espaces sous-optimaux sans remettre en cause sa convergence. De plus, il extrait de façon prioritaire les atomes qui sont les plus corrélés avec le signal, donc les plus pertinents pour en obtenir une explication.

Le Matching Pursuit peut-être appliqué directement pour décomposer des signaux avec des atomes harmoniques spécifiques. Afin de construire le dictionnaire, les paramètres des atomes doivent être échantillonnés, excepté le vecteur des phases Φ. Le cal-

cul de chacune des phases se fait comme dans (Mallat & Zhang, 1993; Gribonval & Bacry, 2003) : une fois l'atome sélectionné, les phases de chaque partiel ϕ_m sont obtenues à l'aide de l'équation (4.1) :

$$e^{j\phi_m} = \frac{\langle x, g_{s,u,m\times f_0,m\times c_0}\rangle}{|\langle x, g_{s,u,m\times f_0,m\times c_0}\rangle|} \tag{4.1}$$

Dans le cas des atomes HSI, les paramètres seront donc échantillonnés de la façon suivante :

- l'échelle est typiquement échantillonnée en puissances de 2 successives (en nombre d'échantillons). Le nombre d'échelles ne doit pas être trop grand afin d'éviter des calculs trop coûteux.
- les localisations temporelles u sont espacées linéairement avec un pas Δu par échelle s,
- la fréquence fondamentale f_0 est échantillonnée logarithmiquement, ce qui constitue une différence notable avec l'algorithme introduit par Gribonval & Bacry (2003),
- le vecteur des amplitudes de partiels A est un des vecteurs $(A_{i,p,k})_{k=1\ldots K}$ de l'instrument i et dont la hauteur p est la plus proche de f_0.

L'échantillonnage logarithmique de la fréquence fondamentale permet de rendre l'échantillonnage de notre dictionnaire uniforme en hauteur de notes (qui est proportionnelle au logarithme de la fréquence fondamentale), ce qui est bien adapté au tempérament égal qui caratérise une grande partie de la musique occidentale.

Pour les atomes HSIC et IHSI, l'échantillonnage des paramètres supplémentaires mis en jeu ne sera pas effectué. Nous utiliserons une procédure spécifique détaillée dans le paragraphe suivant pour les estimer *a posteriori*.

Il faut noter que cet algorithme, ainsi que les suivants, traitent le signal en entier comme dans (Mallat & Zhang, 1993; Krstulovic & Gribonval, 2006), alors que la plupart des autres approches le traitent par trame de la même taille que les atomes employés (Goodwin (2001); Heusdens et al. (2002)). Traiter le signal en entier permet d'utiliser des atomes possédant plusieurs résolutions. La mise en oeuvre est légèrement plus complexe : soustraire un atome implique que toutes les projections du signal sur les atomes dont le support temps-fréquence possède un recouvrement avec l'atome soustrait doivent être mis à jour.

Le choix du dictionnaire a une influence sur le comportement de l'algorithme de Matching Pursuit. On peut noter tout d'abord que rien ne garantit la convergence de l'algorithme vers 0, étant donné que la complétude du dictionnaire n'est pas assurée. Cependant, le nombre d'itérations sera suffisamment réduit pour que cette propriété ne soit pas indispensable. Ensuite, les atomes du dictionnaire peuvent être très corrélés entre eux, en particulier lorsqu'ils ont en commun des partiels de même fréquence, ou lorsque leur supports temporels se recouvrent. Le dictionnaire est donc très cohérent, ce qui présente quelques inconvénients comme nous le soulignerons en 4.2.3.

4.2.2 Matching Pursuit avec réestimation des paramètres

L'échantillonnage "brutal" du dictionnaire selon tous les paramètres peut être assez coûteux. Dans notre cas, pour garder des temps de calcul raisonnables, nous avons choisi de ne pas échantillonner les paramètres de raffinement de la structure harmonique, c'est-à-dire les paramètres de taux de modulation de fréquence fondamentale c_0 et d'inharmonicité β. Dans (Gribonval, 2001), l'échantillonnage de c est évité en estimant le meilleur c sur chacun des sous-dictionnaires $\mathcal{D}_{u,f,s}$ composés de tous les atomes de paramètres

(u, f, s). L'estimation pouvait se faire en se basant sur les valeurs des atomes "plats" ($c = 0$), de fréquences immédiatement inférieures et supérieures sur la grille de f.

Dans le cas d'atomes HSIC et IHSI, nous n'avons pas trouvé une formule explicite d'estimation respectivement du meilleur taux de chirp et du meilleur coefficient d'harmonicité à partir des projections sur les atomes HSI. Nous avons donc choisi une stratégie différente, qui consiste à sélectionner le meilleur atome HSI, puis de maximiser son poids en fonction du paramètre supplémentaire par une méthode de gradient. Cette produire conduit donc à un maximum local du dictionnaire d'atomes HSIC ou IHSI. Les méthodes d'optimisation seront plus détaillées dans le paragraphe 4.6. La Figure 4.1 montre le schéma-bloc de cet algorithme.

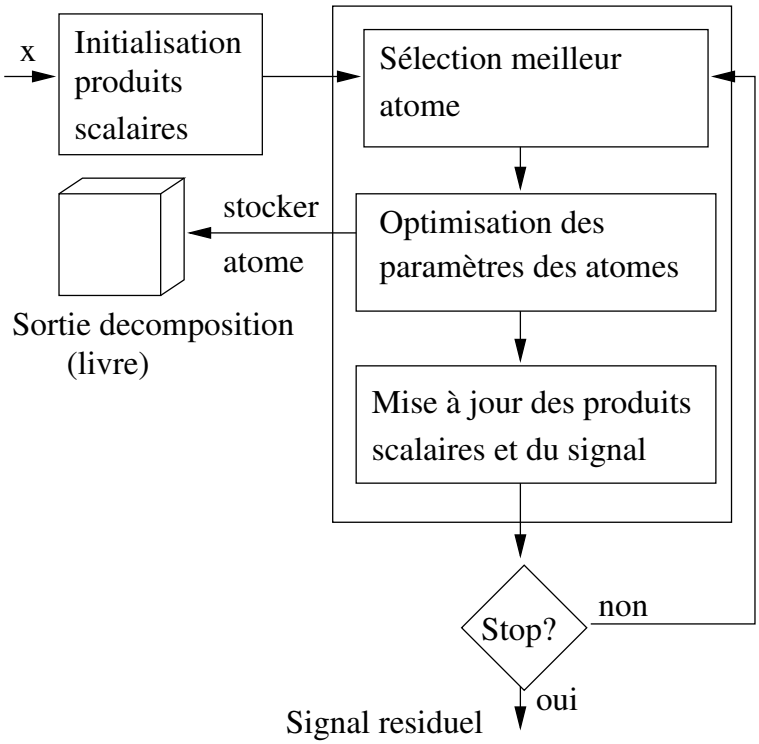


FIG. 4.1 – *Schéma-Bloc de l'algorithme atomique avec réestimation des paramètres.*

Dans le cas de la réestimation des paramètres, la convergence est accélérée pour les premières itérations, étant donné que l'atome sélectionné est plus énergétique que le meilleur atome du dictionnaire plat. La convergence de l'algorithme est aussi assurée car l'énergie du signal décroit. Cependant, la convergence vers 0 n'est pas établie car on sort du cadre du Matching Pursuit (même faible) : on ne sait pas en effet si l'on peut borner le ratio maximum local/maximum global par un réel compris entre 0 strictement et 1. Dans les cas pratiques que nous aborderons, le problème du comportement asymptotique n'aura cependant pas une grande importance : les décompositions seront arrêtées quand le résidu est environ 20 dB en dessous du signal original.

4.2.3 Inconvénient des algorithmes atomiques

Parmi les inconvénients des algorithmes "gloutons", on peut citer leur relative faiblesse lorsque les dictionnaires ont une forte cohérence. En effet, lorsque deux atomes

sont corrélés avec le signal, et qu'ils sont corrélés entre eux, l'extraction du premier atome prend une grande partie de l'énergie et en laisse peu pour le suivant. Dans notre cas d'étude, ce problème peut-être particulièrement gênant pour extraire des structures longues (comme une note longue) à l'aide d'atomes courts, mais aussi lorsque deux sources sont en rapport harmonique à un instant donné. Dans notre étude, nous chercherons essentiellement à contourner le premier obstacle. Comme nous l'avons souligné dans (Leveau & Daudet, 2006) dans le cas d'un dictionnaire de Gabor multi-résolution, l'algorithme de Matching Pursuit ne modélise pas les structures longues de façon optimale : lors des premières itérations, l'algorithme extrait des atomes dont les supports temporels ne coïncident pas. Il s'ensuit qu'une fois que les atomes enlevés recouvrent tout le support temporel de la sinusoïde, la sinusoïde resynthétisée présente des modulations d'amplitude, comme le montre la Figure 4.2, ce qui constitue un inconvénient pour les problématiques de codage.

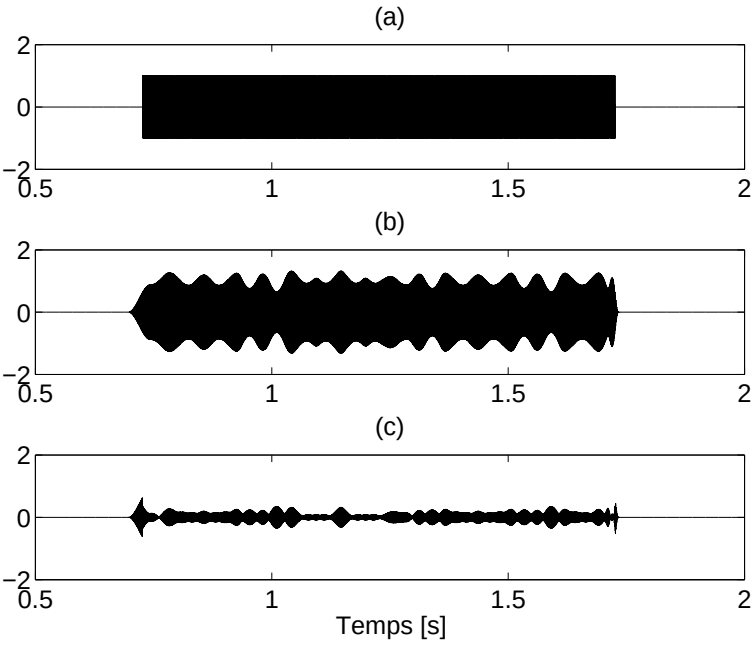


FIG. 4.2 – *Analyse d'une sinusoide (1 seconde, 1000 Hz) avec un dictionnaire de Gabor multi-résolution (échelles 256, 512, 1024, 2048, 4096 à $f_s = 44100Hz$). (a) Signal original, (b) Signal reconstruit à partir des 24 premiers atomes sélectionnés, (c) Signal résiduel.*

Les itérations suivantes permettent de combler les trous d'énergie, mais au prix d'une faible parcimonie, et d'une mauvaise répartition des poids sur les atomes. En effet, il serait désirable que l'énergie des atomes successifs modélisant une structure longue suive les modulations d'amplitude de la note.

Nous allons voir dans la section 4.4 que les approches moléculaires que nous avons développées permettent de contourner cette difficulté en optimisant les poids d'atomes conjointement lorsque ceux-ci ne sont pas orthogonaux, dans le cas particulier où ils occupent des localisations temporelles successives. Les algorithmes moléculaires développés présentent des similarités avec l'algorithme de Matching Pursuit Orthogonal (MPO), la différence résidant principalement sur les ensembles d'atomes sur lesquels portent les

optimisations : l'algorithme de MPO fait l'optimisation des poids sur tous les atomes extraits, tandis que dans notre cas elle est faite localement, sur un sous-ensemble d'atomes ayant de bonnes prédispositions à représenter une note (ou un partiel dans le cas présenté dans (Leveau & Daudet, 2006)). N'effectuer cette optimisation que sur un sous-ensemble d'atomes permet aussi de contourner le problème de complexité algorithmique du MPO.

4.3 Recherche des meilleurs atomes en stéréo

Pour obtenir une décomposition en atomes stéréo, nous utilisons l'algorithme de Stereo Matching Pursuit introduit par Gribonval (2002), modifié à l'étape de calcul des phases (étape 4). L'algorithme complet procède donc de la façon suivante :

1. Les corrélations entre chaque canal $x_l(t)$ et $x_r(t)$ du signal stéréo $x_{st}(t)$ et tous les atomes monophoniques h du dictionnaire sont calculées :

$$\langle x_l, h\rangle = \sum_{t=1}^{T} x_l(t)\,\overline{h}(t) \tag{4.2}$$

$$\langle x_r, h\rangle = \sum_{t=1}^{T} x_r(t)\,\overline{h}(t) \tag{4.3}$$

2. la projection du signal stéréo x_{st} sur les atomes stéréo h_{st} en est déduite :

$$P_{h_{st}} x_{st} = |\langle x_l, h\rangle|^2 + |\langle x_r, h\rangle|^2 \tag{4.4}$$

3. L'atome h qui possède la plus grande valeur $P_{h_{st}} x_{st}$ est sélectionné ;
4. les vecteurs de phase respectifs de chaque canal $h_{l_n}(t)$ et $h_{r_n}(t)$ de l'atome stéréo h_{st} sont calculées comme dans l'équation 4.1.
5. l'atome h_{st} est soustrait du signal une fois son poids α_n et son paramètre de panoramique θ_n calculés :

$$x_{st_{n+1}} = x_{st_n} - 2\alpha_n \mathbf{Re}\left(\left[\cos(\theta_n)h_{l_n}(t) \quad \sin(\theta_n)h_{r_n}(t)\right]^T\right)$$

   les poids α_n et θ_n sont calculés comme suit :

$$\alpha_n = \|P_{h_{st}} x_{st}\| \tag{4.5}$$

$$\theta_n = \tan^{-1}\left(\frac{\langle x_r, h_n\rangle}{\langle x_l, h_n\rangle}\right) \tag{4.6}$$

6. Les projections sont mises à jour sur le signal résiduel, et l'algorithme est itéré à partir de l'étape 2 jusqu'à ce qu'une condition d'arrêt soit satisfaite.

Nous verrons dans la suite que l'on peut traiter les atomes HSIC et IHSI avec cet algorithme. Il suffit pour cela d'introduire une estimation des paramètres entre la sélection et la soustraction de l'atome (étape 5).

Par contre, nous ne présenterons pas d'algorithme moléculaire permettant de traiter le cas stéréo. L'extension de notre algorithme au cas moléculaire est également possible, mais ne sera pas présentée dans le document.

4.4 Recherche des meilleures molécules

Afin d'extraire les molécules définies dans le chapitre précédent (3.7), le Matching Pursuit doit être adapté de façon à extraire plusieurs atomes à la fois. Dans le chapitre 2, nous avons souligné qu'il était très coûteux de sélectionner la molécule qui est optimale *a posteriori*, c'est-à-dire une fois que les poids de ses atomes sont calculés par projection orthogonale du signal sur le sous-espace vectoriel formé par ces atomes. En effet, cela nécessite d'inverser une matrice pour chaque ensemble d'atome candidat. Nous allons donc faire l'hypothèse abusive que les atomes composant une molécule sont orthogonaux afin de calculer un poids de molécule *a priori*, en utilisant les projections individuelles du signal sur les atomes. La molécule ne sera donc pas forcément la molécule optimale *a posteriori*, mais sera au moins une bonne molécule, qui satisfera la condition du Matching Pursuit faible.

Une fois la meilleure molécule *a priori* sélectionnée, on soustraira néanmoins la molécule avec les poids des atomes optimaux, c'est à dire calculés par projection orthogonale du signal sur le sous-espace formé par les atomes de la molécule.

Nous définissons le poids d'une molécule *a priori* comme :

$$\delta(\mathcal{M}) = \left(\sum_{\lambda \in \mathcal{M}} |\langle x, h_\lambda \rangle|^2 \right)^{1/2} \tag{4.7}$$

La nature additive de ce poids par rapport à ceux des atomes permettra de mettre en oeuvre un algorithme de Viterbi pour l'optimiser, comme nous le verrons.

Dans la suite, nous n'aborderons que le cas d'une échelle s unique. La formation de molécules à partir d'atomes d'échelles différentes a été cependant abordée dans (Leveau & Daudet, 2006) dans le cas des atomes de Gabor, pour l'extraction de partiels de fréquence constante. Le principe n'est cependant pas directement transposable dans notre cas, où la fréquence fondamentale des molécules peut varier.

Etant données les contraintes définies pour les molécules, sélectionner une molécule $\mathcal{M}$ revient à sélectionner un chemin $\mathcal{P}$ d'atomes dans des grilles temps-hauteur spécifiques à des instruments. Ces grilles sont construites de la façon suivante : chaque noeud de la grille pour l'instrument i est indexé par sa localisation temporelle u et sa fréquence fondamentale f_0 (échantillonnée sur la grille logarithmique définie en 4.2.1). Chacun de ces noeuds porte la valeur $G_i(u, f_0)$, qui est le maximum du carré des produits scalaires $|\langle x, h_\lambda \rangle|^2$ entre le signal et les atomes h_λ qui ont les paramètres u, f_0 et A. Les vecteurs A considérés appartiennent à $\mathcal{C}_{i,p}$, avec p la hauteur de note correspondant à f_0 ($p = \text{arrondi}(12. \log_2 f_0/440) + 69$) :

$$G_i(u, f_0) = \max_{A \in \mathcal{C}_{i,p}} \{|\langle x, h_{s,u,f_0,A} \rangle|\} \tag{4.8}$$

On définit alors le poids d'une molécule *a priori* à partir des valeurs $G_i(u, f_0)$ par :

$$\delta(\mathcal{P}) = \left(\sum_{(u,f_0) \in \mathcal{P}} G_i(u, f_0) \right)^{1/2} \tag{4.9}$$

La section suivante montre comment cette recherche des meilleurs chemins peut s'intégrer dans un algorithme de poursuite. Deux méthodes différentes pour chercher de tels chemins seront également présentés.

4.5 Algorithmes moléculaires

Les algorithmes moléculaires consistent à remplacer l'étape de la sélection du meilleur atome par celle de la sélection de la meilleure molécule. Comme le montre la Figure 4.3, l'algorithme procède donc de la façon suivante :

1. Les atomes qui composent la molécule sont sélectionnés par recherche du meilleur chemin dans des grilles G spécifiques un instrument. Nous préciserons les approches permettant d'extraire les chemins d'atomes qui respectent les conditions fixées en 3.7 dans les paragraphes suivants.
2. Pour chaque atome sélectionné dans le chemin, les paramètres Φ, éventuellement f_0, c_0 ou β sont estimés (4.6),
3. Les poids respectifs de chacun des atomes sont calculés par projection orthogonale du signal sur le sous-espace formé par les atomes avec leurs paramètres estimés.

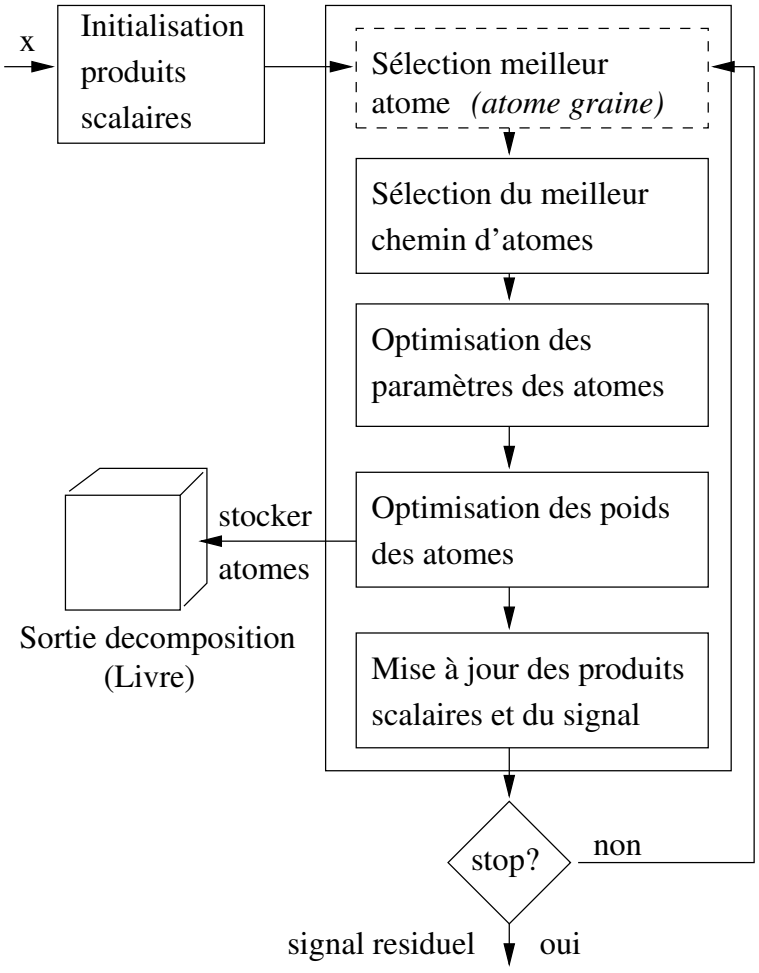


FIG. 4.3 – *Schéma-Bloc d'un algorithme moléculaire avec réestimation des paramètres. L'étape encadrée en pointillé n'est pas nécessaire pour l'algorithme par pénalisation de la longueur de chemin.*

La recherche des meilleurs chemins dans une grille en utilisant un algorithme de recherche exhaustif aurait une complexité très élevée. Un algorithme de Viterbi peut cepen-

dant être mis en oeuvre pour traiter ce type de problème (Forney Jr (1973)), permettant de réduire considérablement la complexité.

4.5.1 Algorithme de Viterbi

L'algorithme présenté dans ce paragraphe sera utilisé dans les deux approches précisées dans les paragraphes suivants.

Supposons que l'on cherche un chemin dans un intervalle de temps $[u_1, u_2]$ dans une grille G_i spécifique à un instrument i. En considérant un noeud au point (u, f_0), l'algorithme de Viterbi est basé sur le principe suivant :

$$\delta^2_{opt}(u, f_0) = \max_{f'_0 \in \mathcal{A}(f_0)} \delta^2_{opt}(u-1, f'_0) + G(u, f_0), \tag{4.10}$$

où $\mathcal{A}(f_0)$ est l'ensemble de fréquences fondamentales échantillonnées qui peuvent atteindre la fréquence f_0, en suivant la condition (3.6), et $\delta_{opt}(u, f_0)$ le poids du chemin optimal entre u_1 et u. En pratique, le meilleur chemin est construit itérativement du temps initial au temps final, en gardant en mémoire les meilleurs chemins intermédiaires. Une fois l'instant final atteint, une opération de rétro-propagation donne le meilleur chemin.

L'utilisation de cet algorithme permet de réduire la complexité de $O(N^U_{f_0})$ à $O(U^2)$, où N_{f_0} est le nombre d'échantillons de fréquence fondamentale et U le nombre d'échantillons temporels[1].

Jusqu'à présent, aucune zone de recherche de la meilleure molécule n'a été définie. Si l'on cherche la meilleure molécule au sens du poids δ maximal, la meilleure molécule *a priori* sera toujours la plus longue possible : ajouter un atome à la molécule augmente obligatoirement son poids. Un tel comportement n'est pas désirable, car il ne permettrait pas de conduire à une parcimonie suffisante pour qu'une molécule corresponde à une note. Il faut donc employer une stratégie qui permette de limiter la taille des molécules.

Deux méthodes ont été envisagées pour résoudre ce problème : une pénalisation de la longueur du chemin présentée en 4.5.2, et une délimitation de la taille des molécules en s'appuyant sur le poids du dernier atome de la molécule (4.5.3).

4.5.2 Approche par pénalisation de la longueur du chemin

Il s'agit de rendre le poids des molécules dépendant de la longueur du chemin en pénalisant la somme des poids des atomes par le nombre d'atomes mis en jeu :

$$\delta_\gamma(\mathcal{M}) = \frac{\left(\sum_{\lambda \in \mathcal{M}} |\langle x, h_\lambda \rangle|^2\right)^{1/2}}{|\mathcal{M}|^\gamma} \tag{4.11}$$

où $|\mathcal{M}|$ est le cardinal de la molécule.

Le coefficient γ permet de pondérer l'influence du nombre d'atomes utilisés et celle des poids de ces atomes, et donc constitue un paramètre permettant de régler la parcimonie *locale* de la molécule. Si $\gamma = 0$, nous nous retrouvons dans le cas précédent : les molécules les plus longues possibles seront sélectionnées. Si $\gamma = 0.5$, les molécules optimales seront les molécules composées d'un seul atome : ajouter un atome moins bon que l'optimal à la molécule fait baisser le poids total de la molécule. On retrouve alors le

[1] La complexité est égale à $O(U^2)$ car la grille de recherche est triangulaire (elle est de $O(N_{f_0}U)$ dans les implémentations classiques pour des grilles rectangulaires)

Matching Pursuit atomique. Le réglage de γ permet donc de fixer un compromis entre ces deux situations extrêmes, comme le montre la Figure 4.4.

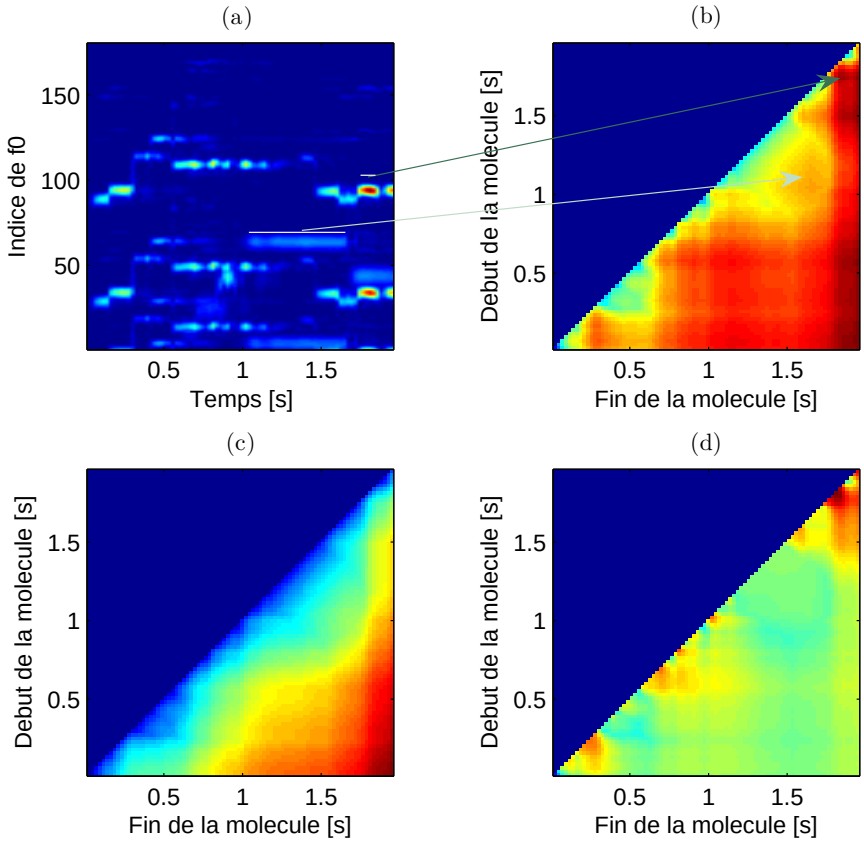


FIG. 4.4 – *Illustration de l'approche par pénalisation de longueur de chemin sur un duo flûte-clarinette. (a) Grille temps-hauteur obtenue à partir des projections sur les atomes de flûte, (b) Poids des chemins avec $\gamma = 0.3$: les chemins sont représentés par une paire de coordonnées (début, fin), la valeur étant le poids du meilleur chemin entre le temps de début et le temps de fin. Les flèches montrent des bons chemins candidats, matérialisés par des maxima locaux dans cette représentation, (c) Poids des chemins avec $\gamma = 0$: les chemins les plus longs sont les meilleurs, (d) Poids des chemins avec $\gamma = 0.5$: les chemins les plus courts sont les meilleurs (les valeurs maximales sont atteintes sur la diagonale : les atomes seuls sont les meilleurs candidats).*

On remarque que pour la valeur $\gamma = 0.3$, des maxima locaux apparaissent dans la matrice des poids de chemin. Ils possèdent des coordonnées $[u1, u2]$ constituant de bon candidats pour être des couples de début et de fin de notes de musique.

Dans ce cas de figure, il est nécessaire d'effectuer les recherches de meilleurs chemins sur tous les intervalles $[u1, u2]$ possibles. La complexité totale de la recherche de chemin devient donc de l'ordre de $O(U^3)$. Il convient donc de fixer une taille maximale de chemin, afin de rendre la recherche possible sur des signaux de durée finie. Si de plus une mise à jour locale des poids de chemin est mise en oeuvre, la coût total des recherches de chemin devient alors presque linéaire par rapport à U : dans ce cas, seuls les poids des

chemins dont le support temporel possède un recouvrement avec la molécule extraite à l'itération précédente sont calculés.

On peut noter que cette approche introduit un critère de parcimonie temporelle *a priori* : il permet de sélectionner un chemin qui capturera beaucoup d'énergie avec peu d'atomes.

4.5.3 Approche par délimitation de la zone de recherche

Une autre stratégie consiste à délimiter la zone de recherche de la meilleure molécule. Cette stratégie est présentée dans (Leveau et al., 2008). Dans un premier temps, un intervalle de temps est défini, puis la meilleure molécule est choisie dans celui-ci. Cette méthode permet d'éviter la comparaison de chemins de longueurs différentes, et limite la complexité de la recherche.

Nous avons choisi de limiter l'intervalle de recherche en calculant un chemin avant et un chemin arrière en partant du noeud *graine*. Ce noeud graine est défini par le noeud qui contient l'atome le plus corrélé avec le signal, c'est-à-dire celui qui est sélectionné dans un algorithme atomique. Les algorithmes permettant de construire ces deux chemins sont strictement symétriques. Nous ne détaillerons donc ci-dessous que l'algorithme pour le chemin avant.

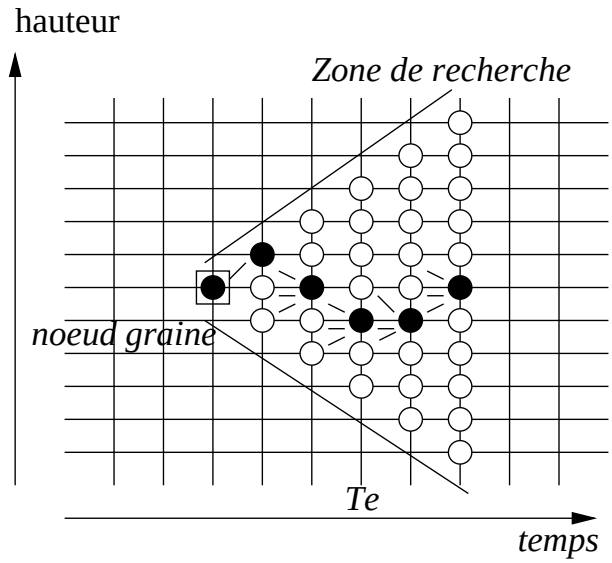


FIG. 4.5 – *Illustration de la recherche du meilleur chemin avant à partir du noeud graine. Les disques blancs indiquent les atomes non sélectionnés dans le chemin, tandis que les disques noirs indiquent les atomes sélectionnés.*

Le chemin qui donne le plus grand poids sous les contraintes fixées est estimé avec l'algorithme de Viterbi décrit en 4.5.1, comme le montre la Figure 4.5. La fin de la propagation avant du chemin est décidée lorsque la valeur du dernier noeud du chemin est inférieure à un seuil $A_{\mathcal{M}}$ défini comme suit :

$$A_{\mathcal{M}} = \max\{\mu_0 \alpha_0^2, \mu_e \alpha_e^2\} \tag{4.12}$$

où $\alpha_0 = |\langle x, h_0 \rangle|$ est le poids du premier atome graine h_0 sélectionné dans toute la décomposition, $\alpha_e = |\langle x, h_e \rangle|$ est le poids de l'atome graine h_e de la molécule courante.

μ_0 and μ_e sont des ratios fixes. Le terme $\mu_0\alpha_0^2$ est un seuil qui empêche la sélection d'atomes qui sont en-dessous de l'erreur d'approximation. Le terme $\mu_e\alpha_e^2$ introduit un niveau de bruit local à la molécule, qui limite la sélection d'atomes appartenant à des notes précédentes ou suivantes, ou à la réverbération. Il dépend de la dynamique voulue à l'intérieur d'une molécule. μ_e doit être plus grand que μ_0, sinon le coefficient n'a pas d'effet sur $A_{\mathcal{M}}$ car α_0 est presque toujours plus élevé que α_e dans une décomposition. Des valeurs typiques pour μ_0 et μ_e sont respectivement de l'ordre de 0,01 et 0,1.

Une fois la zone de recherche sélectionnée, l'algorithme de Viterbi est appliqué pour chaque instrument i sur la grille rectangulaire délimitée par $[u_{\min}, u_{\max}]$ et le registre de l'instrument, sans contraindre les noeuds initiaux et finaux en fréquence fondamentale. Un chemin $\mathcal{P}_i$ par instrument i est donc obtenu et le chemin avec le plus grand poids est finalement sélectionné. On peut remarquer que l'atome graine initial n'est plus utilisé et peut ne pas être inclus dans la molécule finale.

Cette méthode est plus rapide que la précédente, mais nécessite le réglage de deux paramètres (μ_0 et μ_e) permettant de définir un seuil de sélection des atomes, alors que l'algorithme précédent n'a qu'un paramètre γ.

4.6 Optimisation des paramètres après la sélection

Dans cette section, nous présentons des optimisations de paramètres des atomes permettant une convergence plus rapide au niveau des premières itérations et une meilleure adéquation des atomes au signal.

4.6.1 Taux de modulation et fréquence fondamentale

L'optimisation *a posteriori* consiste à maximiser le poids de l'atome par rapport à son paramètre de taux de modulation. Nous avons également choisi de régler conjointement la fréquence fondamentale f_0 avec ce paramètre, pour s'adapter au mieux à la structure harmonique sous-jacente. La fonction à maximiser est donc la suivante pour les atomes HSIC :

$$\mathcal{J}(f_0, c_0) = |\langle x, h\rangle|^2 \tag{4.13}$$

On cherchera à atteindre un maximum local en partant d'une valeur initiale. Pour les atomes HSIC, le vecteur initial (f_0, c_0) est initialisé à $(\hat{f}_0, 0)$, où $\hat{f}_0$ est la fréquence du meilleur atome HSI : on cherche à obtenir le minimum local en partant du meilleur atome plat.

La stratégie choisie est un algorithme du gradient. On peut en effet calculer explicitement la valeur du gradient :

$$\nabla\mathcal{J} = \left[\frac{\partial\mathcal{J}}{\partial f}\frac{\partial\mathcal{J}}{\partial c}\right]^T \tag{4.14}$$

En notant $\otimes$ le produit de deux signaux terme à terme, et avec $\frac{\partial g_m}{\partial f_0} = 2j\pi m(t-u) \otimes g_m$

et $\frac{\partial g_m}{\partial c_0} = 2j\pi m\frac{(t-u)^2}{2} \otimes g_m$, on obtient :

$$\frac{\partial \mathcal{J}}{\partial f_0} = 2\sum_{m=1}^{M} a_m^2.\Re\left(\langle x, 2j\pi m(t-u)\otimes g_m\rangle\overline{\langle x, g_m\rangle}\right), \tag{4.15}$$

$$\frac{\partial \mathcal{J}}{\partial c_0} = 2\sum_{m=1}^{M} a_m^2.\Re\left(\left\langle x, 2j\pi m\frac{(t-u)^2}{2}\otimes g_m\right\rangle\overline{\langle x, g_m\rangle}\right). \tag{4.16}$$

Ceci peut être écrit :

$$\nabla\mathcal{J} = -4\pi\sum_{m=1}^{M} a_m^2.m\times\Im\left(\left\langle\begin{bmatrix} t-u \\ \frac{(t-u)^2}{2}\end{bmatrix}\otimes x, g_m\right\rangle\overline{\langle x, g_m\rangle}\right) \tag{4.17}$$

Le calcul du gradient peut être accéléré en stockant les signaux $t\times x(t)$ et $t^2\times x(t)$.

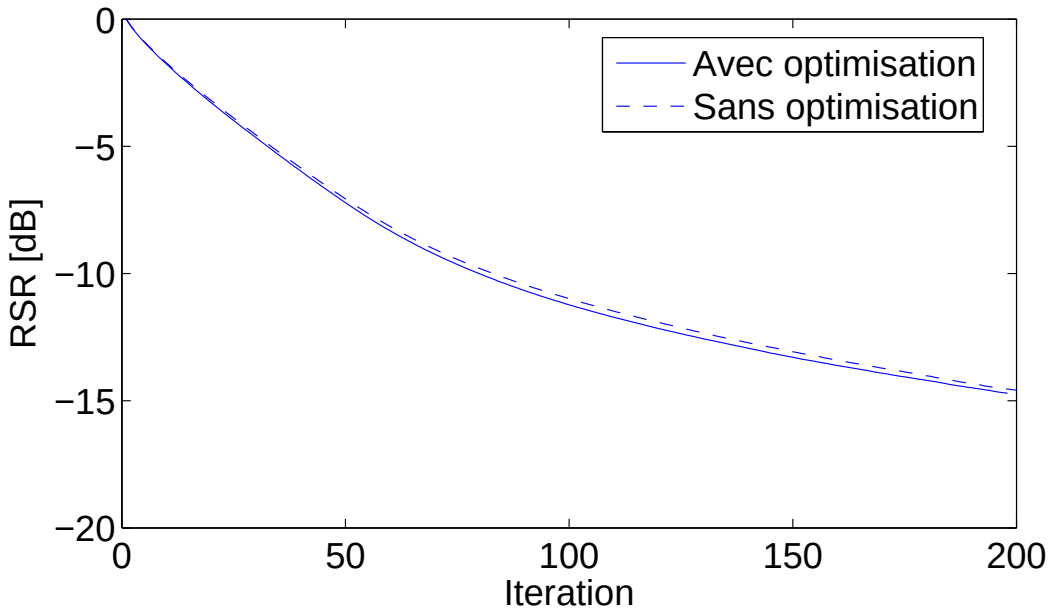


FIG. 4.6 – *Influence de l'estimation du taux de chirp fondamental sur la décroissance de l'énergie de l'algorithme atomique. La courbe est la moyenne de courbes de décroissance de l'énergie pour environ 1500 échantillons de 2s issus de performances solo d'instruments appartenant au dictionnaire (Violoncelle, Violon, Hautbois, Clarinette, Flûte). Les paramètres du dictionnaire sont :* $Fs = 22050Hz, s = 0,046s, \Delta u = 0,023s, \Delta\log_2 f_0 = 1/60$ *(1/10 de ton).*

L'apport d'un point de vue du RSR est représenté sur la Figure 4.6. Le gain est d'environ 0,4 dB après 200 itérations sur des exemples de 2 secondes, ce qui est relativement faible. Par contre, certains gains d'un point de vue applicatif seront évoqués dans le chapitre 6.

Ce calcul peut être étendu au cas stéréo, en remarquant que la fonction de coût en stéréo s'écrit comme la somme de celles des deux canaux :

$$\mathcal{J}_{st} = \mathcal{J}_l + \mathcal{J}_r \tag{4.18}$$

On obtient donc finalement :

$$\nabla \mathcal{J}_{st} = -4\pi \sum_{m=1}^{M} a_m^2 . m \times \Im \left(\left\langle \left[\frac{t-u}{\frac{(t-u)^2}{2}} \right] \otimes x_l, g_m \right\rangle \overline{\langle x_l, g_m \rangle} + \left\langle \left[\frac{t-u}{\frac{(t-u)^2}{2}} \right] \otimes x_r, g_m \right\rangle \overline{\langle x_r, g_m \rangle} \right) \tag{4.19}$$

4.6.2 Inharmonicité et fréquence fondamentale

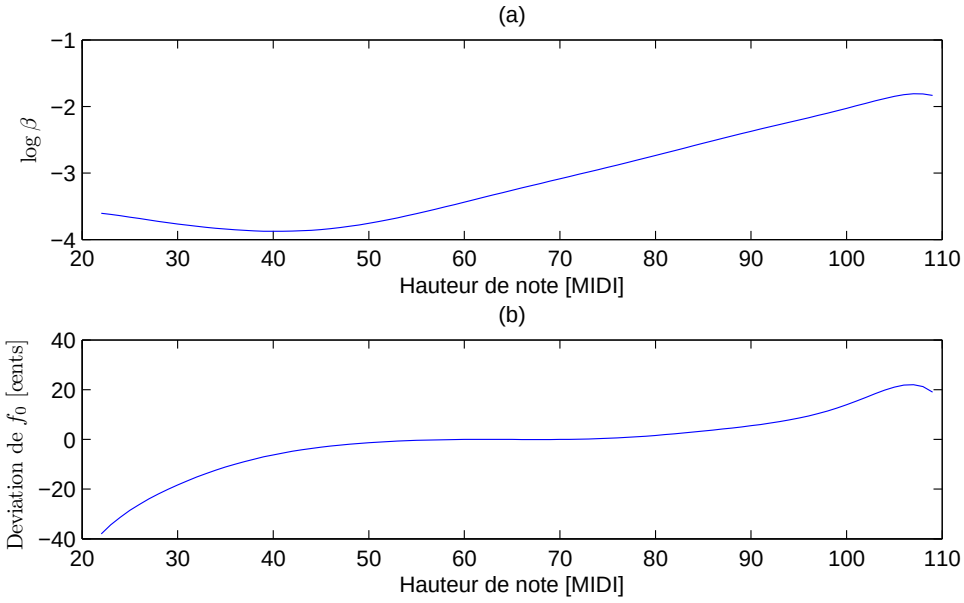


FIG. 4.7 – *Paramètres d'inharmonicité d'un piano moyen utilisés pour la construction du dictionnaire. (a) coefficient d'inharmonicité β, (b) déviation de la fréquence fondamentale du peigne harmonique par rapport à la fréquence fondamentale d'accord.*

Dans le cas des atomes inharmoniques,

$$\mathcal{J}_(f_0, \beta) = |\langle x, h \rangle|^2 \tag{4.20}$$

Bien que le calcul soit possible, nous ne prenons pas en compte ici le taux de modulation du fondamental dans cette fonction de coût. En effet, l'inharmonicité et la modulation de fréquence apparaissent de façon disjointe dans la grande majorité des sources musicales naturelles : l'inharmonicité se produit lorsque le son n'est pas entretenu (Piano), contrairement aux modulations de fréquence (instruments à cordes frottées, instruments à vent...).

Concernant les atomes IHSI, le vecteur initial est initialisé à $(f_{0_i}, \beta_i(f_0))$, où $\beta_i(f_0)$ est un coefficient d'inharmonicité moyen pour la fréquence fondamentale f_0, pris sur plusieurs pianos (Emiya et al. (2007)). Dans le cas du piano, on peut également observer une déviation entre la fréquence fondamentale physique f_{0_i} par rapport à la fréquence fondamentale perçue (visée par l'accordeur du piano). La valeur initiale f_{0_i} sera donc initialisée en utilisant une déviation moyenne. Ainsi, avec les paramètres ainsi définis, la position des partiels $f_{m,p}$ d'une note de pitch p sera donnée par :

$$f_{m,p} = m.\sqrt{1 + \beta(m^2 - 1)}.(f_{0_p}.2^{d_p/1200}) \tag{4.21}$$

où d_p est la déviation en cents de la fréquence fondamentale physique par rapport à la fréquence fondamentale d'accord.

Les références prises pour ces deux paramètres sont présentées sur la Figure 4.7, issue de (Rossing & Fletcher, 1998).

Dans ce cas le gradient s'écrit :

$$\nabla \mathcal{J} = -4\pi \sum_{m=1}^{M} a_m^2 . m \times \Im \left(\left\langle \left[f_0(1/2 \times \frac{t-u}{\frac{m^2-1}{\sqrt{1+\beta(m^2-1)}}}) \right] \otimes x, g_m \right\rangle \overline{\langle x, g_m \rangle} \right) \tag{4.22}$$

L'amélioration introduite par la réestimation des paramètres est illustrée sur la Figure 4.8. On remarque que l'optimisation des paramètres permet d'améliorer le SRR pour les premières itérations. Ensuite, la convergence de l'algorithme sans estimation des paramètres est plus rapide, ce qui indique que les atomes physiques sont extraits assez tôt. Il est possible que ce phénomène se manifeste également pour l'estimation du taux de modulation de chirp, peut-être plus tardivement car le dictionnaire est plus réduit (une seule échelle courte).

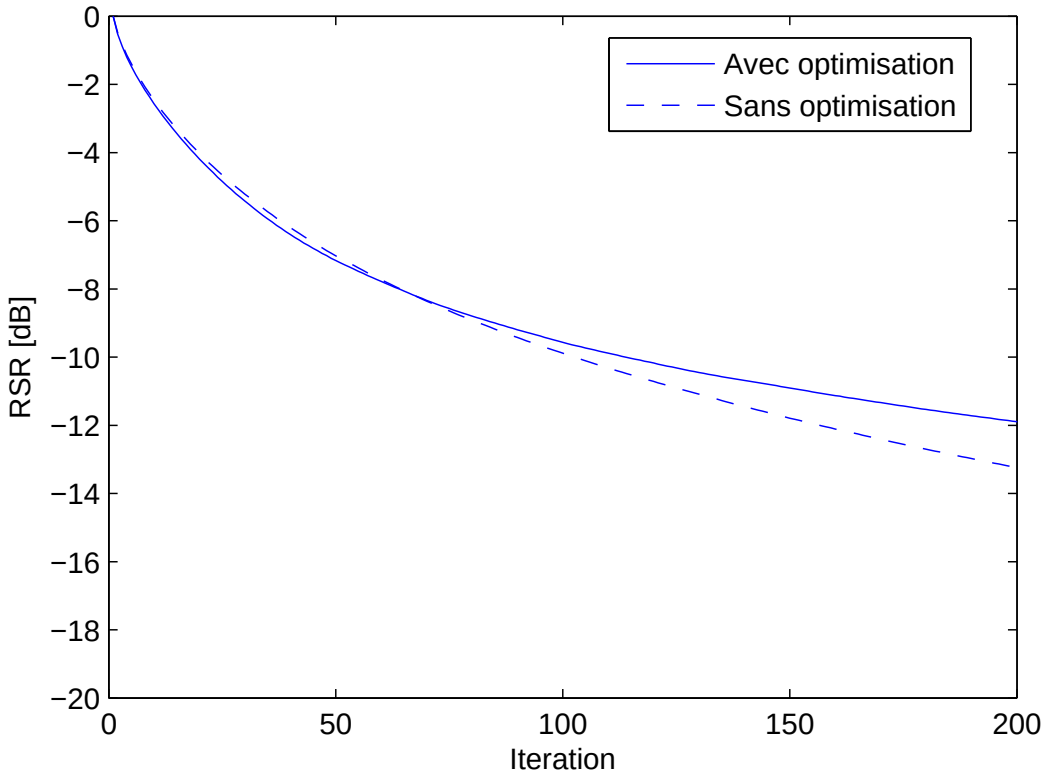


FIG. 4.8 – *Influence de l'estimation du paramètre d'inharmonicité et de la fréquence fondamentale sur la décroissance de l'énergie de l'algorithme atomique. La courbe est la moyenne de courbes de décroissance de l'énergie pour 355 échantillons de 2s issus de performances solo de Piano. Le dictionnaire utilisé ne contient que des atomes de piano. Les paramètres du dictionnaire sont :* $Fs = 22050Hz, (\Delta u, s) \in \{(512, 256), (1028, 2048), (1024, 4096), (1024, 8192)\}, \Delta \log_2 f_0 = 1/60$ *(1/10 de ton).*

Comme précédemment, la formule peut être étendue à un signal stéréo en écrivant la fonction à maximiser en stéréo comme la somme de celles en mono.

4.6.3 Vecteur des phases

Les phases des partiels Φ_m sont déterminées en prenant les angles des poids complexes, comme dans (Mallat & Zhang, 1993) :

$$e^{j\phi_m} = \frac{\langle x, g_{\lambda_m} \rangle}{|\langle x, g_{\lambda_m} \rangle|} \tag{4.23}$$

où λ_m est le vecteur des paramètres de l'atome de Gabor représentant le partiel m. Comme mentionné par Gribonval (1999), cette phase n'est pas la phase optimale de l'atome réel correspondant, qui sera ensuite retiré du signal. Dans le cas des atomes de Gabor, la phase de l'atome est d'autant moins optimale que le produit $s.f$ est faible. Dans notre cas, étant donné que les dictionnaires seront construits de telle sorte que les produits $s.f_0$ soient élevés pour des considération d'orthogonalité des partiels, cette limitation ne sera pas gênante.

4.6.4 Vecteur d'amplitudes

Le vecteur d'amplitudes peut être aussi modifié a posteriori pour être plus proche du signal, tout en restant similaire à des atomes de la classe $\mathcal{C}_{ip}$ sélectionnée. Pour cela, un vecteur A est construit comme une combinaison linéaire non-négative d'autres vecteurs de la classe :

$$A = \sum_{n \in \mathcal{C}_{ip}} \lambda_n A_{i,p,n} \tag{4.24}$$

avec $\lambda_n \geqslant 0$. Cette optimisation a lieu une fois l'atome h sélectionné : Il s'agit de trouver :

$$\hat{A} = \arg\max_{(\lambda_n)} \left\{ |\langle x, \sum_{n \in \mathcal{C}_{ip}} \lambda_n \sum_m a_{i,p,n,m} g_m \rangle| \right\} \tag{4.25}$$

on calcule alors la combinaison linéaire non-négative de vecteurs A la plus proche du vecteur $(|\langle x, g_m \rangle|)_{m=1..M}$ où g_m sont les vecteurs de partiels (notation simplifiée). On peut en effet écrire :

$$\hat{A} = \arg\max_{(\lambda_n)} \left\{ |\langle B, \sum_{n \in \mathcal{C}_{ip}} \lambda_n A_{i,p,n} \rangle| \right\} \tag{4.26}$$

où $B = (|\langle x, g_m \rangle|)_{m=1..M}$. Le problème est alors de trouver une factorisation non-négative de B avec la donnée des vecteurs $A_{i,p,n}$. Ce problème peut être résolu efficacement avec l'algorithme traitant le problème avec une distance euclidienne proposé par Lee & Seung (2001), en ne mettant pas à jour la matrice contenant les vecteurs d'amplitudes.

L'amélioration apportée par cette optimisation est illustrée sur la Figure 4.9. L'utilisation de cet algorithme permet d'obtenir une convergence légèrement plus rapide en gardant une complexité raisonnable. Son utilisation est intéressante pour des perspectives de codage : la partie harmonique du timbre peut être codée de façon plus précise, avec un dictionnaire de taille identique et à un coût de codage assez faible, en restant dans l'espace vectoriel des amplitudes de l'instrument codé. Cependant le gain en terme de RSR n'est peut-être pas suffisamment important pour que cette optimisation soit rentable d'un point de vue codage.

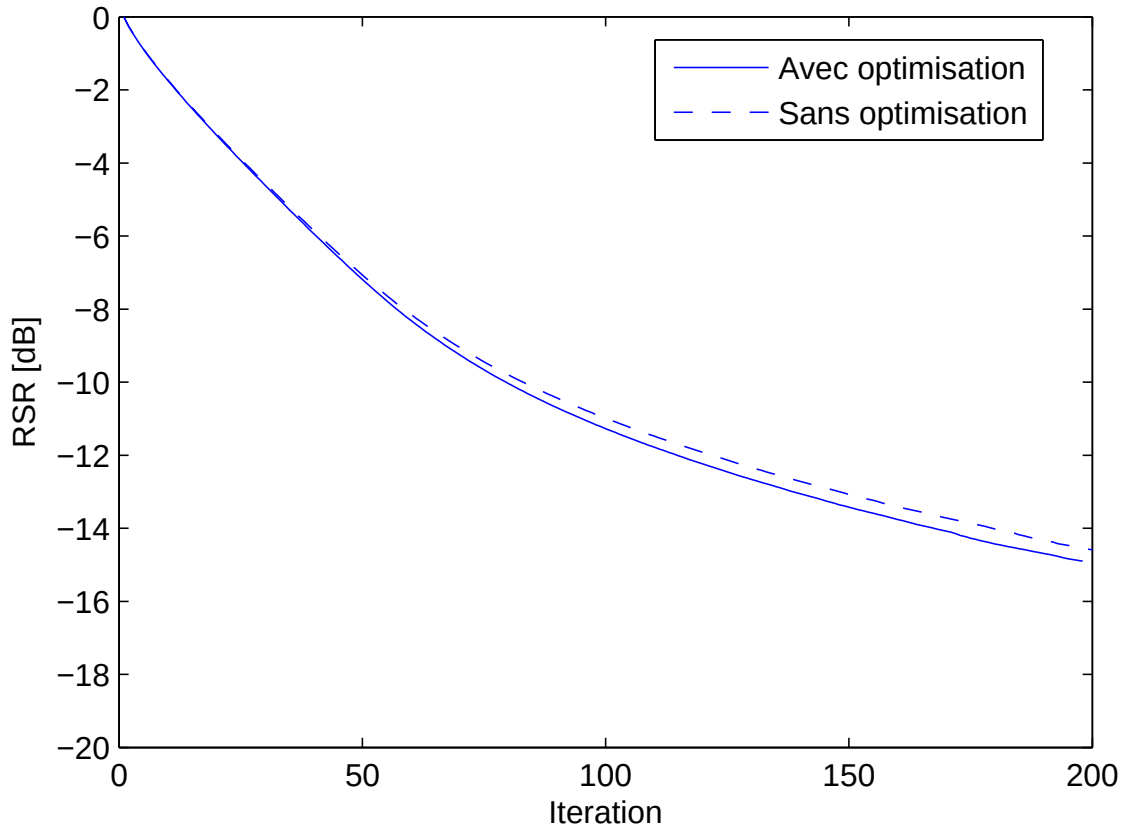


FIG. 4.9 – *Influence de l'optimisation du vecteur d'amplitude sur la décroissance de l'énergie de l'algorithme atomique. La courbe est la moyenne de courbes de décroissance de l'énergie pour environ 1500 échantillons de 2s issus de performances solo d'instruments appartenant au dictionnaire (Violoncelle, Violon, Hautbois, Clarinette, Flûte). Les paramètres du dictionnaire sont :* $Fs = 22050Hz, s = 0,046s, \Delta u = 0,023s, \Delta \log_2 f_0 = 1/60$ *(1/10 de ton).*

4.6.5 Calcul des poids (molécules)

Lorsqu'une molécule est extraite, et que tous les paramètres des atomes sont estimés, les poids de chacun des atomes de la molécule sont calculés par projection orthogonale des atomes sur l'espace formé par les atomes de la molécule :

$$\alpha_{0_i} = \frac{G(h_1, ..., h_{\lambda-1}, x, h_{\lambda+1}, ..., h_n)}{G(h_1, ..., h_{\lambda-1}, h_\lambda, h_{\lambda+1}, ..., h_n)} \tag{4.27}$$

où G est le déterminant de Gram[2] des vecteurs en argument.

Nous avons illustré l'apport de l'optimisation des poids en prenant l'exemple d'un dictionnaire d'atomes multi-résolution (Leveau & Daudet (2006)).

4.6.6 Remarque sur l'optimisation des paramètres des atomes d'une molécule

Nous avons présenté des optimisations de paramètres qui peuvent être effectuées sur des atomes sélectionnés isolément ou conjointement. Dans le second cas, étant donné que les atomes ne sont pas orthogonaux, tous les paramètres devraient être optimisés dans le même temps afin d'obtenir l'ensemble de paramètres qui définit le meilleur sous-espace, en partant de l'ensemble des paramètres initiaux. Nous n'effectuerons pas cette optimisation, en considérant que la corrélation entre les atomes adjacents n'est pas suffisamment forte pour induire un mauvais ensemble de paramètres pour toute la molécule.

4.7 Complexité

Dans la suite de ce paragraphe, nous évaluons la complexité d'une seule itération de chaque algorithme itératif. Le rappel des paramètres de l'algorithme figure dans le Tableau 4.1.

Paramètre	Signification
D	déviation maximale en hauteur de note entre deux atomes consécutifs d'une molécule
M	nombre de partiels d'un atome ISH
K	nombre d'atomes par classe $\mathcal{C}_{ip}$
U	nombre d'atomes contenu dans un chemin ($U = 1$ pour un algorithme atomique)
I	nombre d'instruments du dictionnaire
N_{f0}	nombre de fréquences fondamentales dans le dictionnaire
N_s	nombre d'échantillons pour l'échelle s
N_{U_s}	nombre de pas temporels à mettre à jour pour l'échelle s
$\Delta \log f_0$	échantillonage du logarithme de la fréquence fondamentale

TAB. 4.1 – *Paramètres des algorithmes*

La charge de calcul pour chacune des observations est détaillée dans le Tableau 4.2.

Dans les cas pratiques, la charge de calcul est dominée par la mise à jour des produits scalaires et par l'optimisation des paramètres pour chacun des algorithmes. La mise à jour

[2]Le déterminant de Gram $G(x_1, x_2, \ldots, x_n)$ est le déterminant de la matrice de Gram definie par ses éléments $G_{i,j} = \langle x_i, x_j \rangle$

MP.	At.	M1.	M2.	Opération	Coût
X	X	X	X	MAJ des corrélations (Gabor)	$\sum_s M \times N_{f0} \times N_s \times N_{U_s}$
X	X	X	X	MAJ des corrélations (A)	$\sum_s M \times K \times I \times N_{f0} \times N_{U_s}$
		X		Rech. de chemin (approche 1)	$I \times (D/\Delta \log f_0)^2 * U^3$
			X	Rech. de chemin (approche 2)	$I \times (D/\Delta \log f_0)^2 * U^2$
	X	X	X	Opt. des param. des atomes	$M \times N_s \times N_{it} \times U$
		X	X	Opt. des poids des atomes	$U^2 \times N_s$

TAB. 4.2 – *Complexité des étapes de calcul pour une itération d'un algorithme. Les étapes mises en jeu dans chacun des quatre algorithmes présentés dans notre étude sont indiquées (MP = Matching Pursuit, At = Algorithme atomique avec réestimation des paramètres, M1 = Algorithme moléculaire par pénalisation de la longueur de chemin (4.5.2), M2 = Algorithme moléculaire par délimitation de la zone de recherche (4.5.3).*

rapide des produits scalaires proposée dans (Mallat & Zhang, 1993) n'a pas été mise en oeuvre, étant donné qu'elle demandent une connaissance au préalable des produits scalaires entre les atomes du dictionnaire pris deux à deux. En effet, les algorithmes utilisant des optimisations de paramètres ne permettent pas de connaître à l'avance les paramètres des atomes qui seront sélectionnés. Néanmoins, il est possible de réduire la complexité du calcul des produits scalaires avec les atomes de Gabor en utilisant une transformée de Fourier à Court Terme. L'échantillonnage logarithmique du dictionnaire empêche cependant de l'appliquer directement avec une taille de fenêtre égale à celle des atomes : on peut alors utiliser un bourrage de zéro (*zero padding*) afin d'affiner l'échantillonnage fréquentiel de la transformée, puis procéder à une interpolation en splines cubiques afin d'obtenir une estimation de l'amplitude des projections aux points fréquentiels désirés.

L'ordre de grandeur de la complexité avec des paramètres tels que nous avons choisis dans le chapitre conduisent à des algorithmes ayant un temps d'exécution de 10 à 100 fois la durée du signal, avec une mise en oeuvre en Matlab exécutée sur un ordinateur monoprocesseur cadencé à 3GHz. Une mise en oeuvre plus efficace serait cependant envisageable dans un langage plus bas niveau, afin d'obtenir des temps d'exécution avoisinant le temps d'écoute du signal.

Le Tableau 4.2 montre que le coût de calcul a une variation affine par rapport au nombre d'instruments mis en jeu dans le dictionnaire. Ainsi, même avec un nombre important d'instruments, l'algorithme de décomposition reste applicable avec une complexité raisonnable. On peut éventuellement réduire la complexité en utilisant une procédure pour le calcul des produits scalaires comme présenté dans (Jost et al., 2006).

4.8 Relation avec l'indexation audio

Nous allons maintenant discuter de la capacité des algorithmes présentés à extraire des caractéristiques pertinentes du signal musical, en montrant leurs avantages et inconvénients respectifs, mais aussi en pointant les améliorations possibles et les opérations pouvant être réalisées en post-traitement.

4.8.1 Classification

Dans le chapitre précédent, nous avons vu que la sélection de l'atome optimal du point de vue du RSR revient à choisir un atome dont le timbre et la hauteur de note est similaire à une portion du signal à analyser. A chaque itération la classe $\mathcal{C}_{ip}$ dont l'atome la représentant est le plus proche du signal est sélectionné, puis soustrait du signal. Cette classification est assez rudimentaire, des développements permettant d'améliorer la sélection de la bonne classe seraient donc envisageables, par exemple en introduisant des algorithmes issus des méthodes avancées de classification statistique.

4.8.2 Estimation de la localisation temporelle des notes

Les algorithmes présentés permettent de s'adapter convenablement aux structures temporelles sous certaines hypothèses :

- l'algorithme atomique (4.2.2) permet de s'adapter automatiquement à la durée des notes sur un signal monophonique pourvu que les enveloppes temporelles des atomes soient proches de l'enveloppe temporelle de la structure harmonique de la source, et que les durées soient échantillonnées finement. Cependant, utiliser un échantillonnage très fin de l'échelle peut être prohibitif en temps de calcul.
- l'algorithme moléculaire (4.5) permet de prendre en compte des notes longues : elles sont représentées par plusieurs atomes d'échelle courte. L'approche par pénalisation de la longueur de chemin permet de régler la parcimonie temporelle de la représentation, tandis que l'approche par seuillage local ne permet pas de la contrôler, mais demeure plus rapide.

On peut noter que si les structures temporelles définies s'adaptent aux notes en localisation u et en durée s, la détection des débuts et des fins des notes est réalisée implicitement car ce sont les objets sonores qui sont identifiés.

4.8.3 Estimation de l'enveloppe temporelle des notes

Comme nous l'avons souligné dans 4.2.3, les algorithmes atomiques que nous avons présentés ne permettent pas de représenter les enveloppes temporelles des notes de façon adéquate. Par contre, l'algorithme moléculaire avec réestimation des paramètres est mieux adapté à cette tâche : il permet de calculer le poids des atomes avant extraction, ce qui permet d'éviter l'inconvénient de l'estimation de poids itérative. La réestimation des paramètres permet d'ajuster la fréquence et le taux de modulation fondamentaux de sorte que les modulations de fréquences sont aussi bien représentées.

On peut donc espérer tirer parti de ces modulations pour aider à catégoriser les notes extraites, et éventuellement faire du codage objet de paramètres d'assez haut-niveau, conduisant à des débits extrêmement bas.

4.8.4 Estimation du nombre de sources

Un point faible notable des algorithmes présentés est qu'ils ne gèrent pas les problèmes de parcimonie en hauteur. Sur une trame temporelle, on obtient donc naturellement les premiers atomes qui "expliquent" le mieux le signal, et les suivant sont extraits sur le résiduel ou les lobes secondaires sans qu'on puisse savoir quels sont les atomes physiques et ceux correspondant à l'erreur de modélisation. Il faut alors soit régler les critères de l'algorithme (RSR ou nombre d'atomes maximaux), soit effectuer un post-traitement

sur des décompositions effectuées en utilisant des critères d'arrêt suffisamment non-restrictifs pour que la plus grande partie du signal orignal soit exprimée. La première option nous contraint d'avoir une connaissance *a priori* forte sur les signaux analysés (dynamique du signal, réverbération...).

Nous envisagerons donc des post-traitements adaptés pour faire le tri entre les atomes physiques et les atomes d'erreur de modélisation afin de rendre les applications plus robustes. La prise en compte de la polyphonie pourrait être faite dans l'algorithme en mettant en oeuvre une estimation jointe de la meilleure combinaison d'atomes expliquant le signal à un instant donné, mais ce sujet n'a pas été abordé dans ce travail.

4.9 Post-traitement pour la parcimonie en pitch *a posteriori*

Afin de palier le problème mentionné précédemment, nous introduisons un post-traitement permettant de gérer la parcimonie en hauteur *a posteriori*.

Nous avons évoqué en 4.1 la nécessité de remplir des conditions de parcimonie en pitch pour que le nombre d'atomes extraits à un instant donné corresponde au nombre de sources activées. Or, les algorithmes dérivés du Matching Pursuit ne permettent pas de contrôler à la fois les poids respectifs de l'erreur de modélisation et du nombre d'atomes. En effet, nous ne connaissons pas de critère portant à la fois sur le nombre d'atomes et sur l'erreur de modélisation, strictement monotone et donc qui pourrait être défini comme un critère d'arrêt pour l'algorithme de Matching Pursuit.

On peut néanmoins utiliser des critères de parcimonie *a posteriori* sur des décompositions "quasi-infinies", c'est-à-dire qui extraient beaucoup plus d'atomes que nécessaire, afin de répondre à des critères de parcimonie intéressants pour les applications. Un post-traitement possible consiste à enlever les atomes d'un livre en effectuant des tests pour des instants donnés du signal.

A partir d'un livre supposé infini, et étant donné un instant de la décomposition, la procédure suivante peut être exécutée :

1. L'axe temporel est échantillonné au plus grand commun diviseur entre les échantillonnages temporels Δu des atomes.
2. Pour chaque instant u de cet échantillonnage, on construit l'ensemble $\mathcal{S}$ des atomes dont le support temporel contient l'instant u.
3. Sur cet ensemble, on calcule l'*énergie instantanée* de chacun des atomes, donné par :

$$e_\lambda = (|\alpha_\lambda w(\frac{u - u_\lambda}{s_\lambda})|)^2 \tag{4.28}$$

4. Les atomes sont triés par énergie instantanée décroissante.
5. Une mesure de parcimonie en hauteur *a posteriori* est définie pour les n premiers atomes par :

$$P_n = \frac{\sqrt{\sum_{n'=1}^{n} e_{n'}}}{n^\beta} \tag{4.29}$$

Ce critère est similaire à celui employé pour l'algorithme moléculaire décrit en 4.5.2 : c'est un critère portant sur l'énergie des atomes sélectionnés pénalisé par le nombre d'atome utilisé. Le paramètre β permet de régler l'importance accordée entre le poids des atomes extraits et le nombre d'atomes utilisé pour représenter le signal. La procédure suivante est alors effectuée : tant que P_n augmente, les atomes

h_{λ_n} sont gardés dans le livre. Cette procédure ne garantit pas que le maximum de P_n est atteint, seul un maximum local est garanti. Notons que si $\beta = 0$, (P_n) sera strictement croissante car $e_{n'} > 0$, donc aucun atome ne sera enlevé. Si $\beta = 0.5$, on peut montrer que (P_n) est décroissante car $e_{n'} \geq e_{n'+1}$. Dans ce cas, tous les atomes seront enlevés, sauf le plus énergétique. Il s'agit donc de fixer β entre 0 et 0.5 pour obtenir la parcimonie désirée.

L'effet de cette procédure sur les livres sur la représentation sera mis en valeur dans les sections suivantes. Ce processus est illustré sur la figure 6.6.

On peut imaginer d'autres méthodes de post-traitement faisant intervenir de l'apprentissage : étant donné un livre dont on connait les atomes physiques et les atomes d'erreur de modélisation, on pourrait entrainer un classifieur qui prédirait la nature de l'atome. Ces paramètres pourraient être calculés avant l'extraction de l'atome, et faire intervenir par exemple des critères sur la platitude du spectre dans la zone temps-fréquence de l'extraction, caractéristique des zones bruitées du spectre.

4.10 Bilan

Dans ce chapitre, nous avons présenté des algorithmes dérivés du Matching Pursuit permettant d'extraire les atomes et les molécules que nous avons définis dans le chapitre précédent. Les optimisations que nous avons présentées permettent d'utiliser des atomes avec de nombreux paramètres, sans qu'il soit nécessaire de les échantillonner. L'algorithme moléculaire permet enfin d'extraire des structures longues, qui idéalement seraient proches des notes jouées par les instruments.

Les algorithmes proposés ont l'avantage d'être relativement peu coûteux. Une mise en oeuvre efficace permettrait une exécution temps-réel, avec un échantillonnage suffisant pour obtenir des décompositions pertinentes, et cela même pour un grand nombre d'instruments différents. D'autres variantes seraient intéressantes à développer, comme des Matching Pursuit moléculaires en stéréo et/ou en multi-résolution. Cependant, les algorithmes proposés ont peu de chance de mener à des décompositions optimales, étant donnée la non-orthogonalité forte de certains atomes : si le cas des non-orthogonalités d'atomes successifs a été traité, le problème de l'extraction d'atomes possédant le même support temporel et avec des fréquences fondamentales en rapport harmonique ne l'a pas été. Une perspective intéressante d'étude serait d'extraire des molécules "verticales" d'atomes harmoniques, où des stratégies de sélection jointe pourraient être mises en oeuvre. On peut également penser à utiliser des algorithmes comme FOCUSS ou une modélisation statistique comme celle présentée par Vincent (2004) qui permettrait de traiter le problème d'approximation de façon globale. Ce type d'approche devrait néanmoins augmenter la charge de calcul.

On peut également se demander s'il serait judicieux d'introduire de nouvelles couches au-dessus des molécules à l'intérieur de l'algorithme. On peut penser notamment aux accords, qui seraient constituées d'atomes HSI ou de molécules d'atomes HSI superposé(e)s dans le temps. Dans ce cas de figure, la question intéressante serait de savoir si ces molécules seraient stockées dans un dictionnaire, notamment pour contourner le problème de la non-orthogonalité très forte de certains couples d'atomes, ou si elles seraient calculées par un algorithme d'estimation jointe du meilleur couple d'atome. Dans le premier cas, un compromis entre la charge de calcul et la mémoire à utiliser serait à faire, et il est probable qu'une structuration du dictionnaire en arbre comme l'a

proposée Jost et al. (2006) pourrait faciliter la mise en oeuvre. Un autre type de structuration pourrait être utile : une structuration en lignes mélodiques. Il s'agirait alors de définir une ligne mélodique comme une succession de molécules d'atomes ISH. Ici également, on pourrait choisir entre une approche dictionnaire et une approche *ad hoc*. On peut d'ailleurs mentionner que la représentation des mélodies par un dictionnaire (ou répertoire) a déjà été proposée (Parsons (1975)), et a déjà été abordée dans plusieurs travaux d'extraction de mélodie à partir de données MIDI ou audio (Song et al. (2002)). La sélection de mélodie *ad hoc* pourrait être effectuée en imposant des contraintes ou des pénalités sur des chemins possibles cette fois-ci entre les notes.

Dans le chapitre suivant, nous allons préciser comment apprendre les vecteurs d'amplitudes caractérisant les atomes que nous avons définis.

Chapitre 5

Base de données de sons et apprentissage

Dans ce chapitre, nous présentons les bases de données de sons que nous utiliserons pour l'apprentissage des dictionnaires, et les applications que nous aborderons dans le chapitre 6.

5.1 Bases de données de sons

5.1.1 Notes isolées (ISO)

La base de notes isolées a été construite à partir des trois bases de notes isolées suivantes : RWC Musical Instrument Sound Database (Goto et al., n.d., RWCdb), IRCAM Studio On Line (Ircam, n.d., SOLdb) et University of Iowa Musical Instrument Samples (, auteur inconnu, IOWAdb). Pour les instruments les mieux représentés, on peut avoir jusqu'à 20 échantillons d'une même note car différents modes de jeu et différentes nuances sont présentes dans la base.

Les différentes bases sont annotées en hauteur de note. Cependant, certaines erreurs ont été détectées dans l'annotation IOWA (erreurs de justesse de jeu ou annotation). Les annotations de départ ont alors été modifiées par annotation semi-automatique, en utilisant l'algorithme *yin* (de Cheveigné & Kawahara (2002)) pour produire une annotation ensuite corrigée manuellement. La correction manuelle est faite en exploitant le fait que les notes sont jouées dans des gammes, et occupent donc des hauteurs consécutives. Les instruments pour lesquels nous possédons des échantillons de notes annotés sont présentés dans le Tableau 5.1.

Cette base sera appelée ISO dans la suite.

5.1.2 Phrases solo

Des échantillons de soli d'instruments seront aussi utilisés afin de traiter des signaux plus réalistes, et aussi d'apprendre des atomes sur des enregistrements réels. Cette base de données a été constituée par Slim Essid pour son travail de thèse (Essid (2005)). Les échantillons proviennent pour la plupart de CDs du commerce. D'autres ont été enregistrés en chambre anéchoïque. Pour chaque instrument, la base de soli est composée de deux sous-ensembles disctincts, de durées égales mais de sources disjointes.

Les bases correspondantes seront appelées SOLO1 et SOLO2.

Instrument	Abbréviation	IOWA	SOL	RWC
Contrebasse	Ba	X	X	X
Basson	Bo	X	X	X
Clarinette	Cl	X	X	X
Violoncelle	Co	X	X	X
Cor	Fh	X	X	X
Flute	Fl	X	X	X
Hautbois	Ob	X	X	X
Piano	Pn	X		X
Saxophone alto	Sa	X	X	X
Saxophone soprano	Ss	X	X	X
Tuba	Ta	X	X	X
Trombone	Tb	X	X	X
Trompette	Tr	X	X	X
Violon alto	Va	X	X	X
Violon	Vl	X	X	X

TAB. 5.1 – *Composition de la base ISO. Les croix indiquent que la base en en-tête contient l'instrument considéré.*

5.1.3 Musique d'ensemble

Trois bases contiennent de la musique d'ensemble.

5.1.3.1 DUO

Cette base est composée de duos extraits d'enregistrements du commerce, mettant en jeu les instruments suivants : Violoncelle, Flûte, Violon et Clarinette. Elle a été constituée par Emmanuel Vincent. La composition de la base est la suivante :

Ensemble	Durée en secondes	Nombre de sources
Cl&Fl	400	4
Co&Fl	340	3
Fl&Fl	58	1
Co&Vl	828	7
Total	1626	

TAB. 5.2 – *Composition de la base DUO.*

Les extraits mettent en jeu les deux instruments la majeure partie du temps, mais certaines plages courtes où un seul instrument est actif peuvent être présentes.

5.1.3.2 ENS1

La base de données ENS1 est une base composée de mélanges mono artificiels et instantanés de performances solo provenant de l'ensemble SOLO2 : les signaux sont mélangés en effectuant une simple somme des signaux. Avant d'être mélangés, les signaux sont mis à la même durée et à la même énergie. Les silences de chacune des pistes sont également enlevés avant ces opérations. Dans ces conditions, la polyphonie de ces

signaux est assez strictement contrôlée, et permettra donc de faire des expériences d'estimation du nombre de sources. Les signaux résultants ne seront cependant pas réalistes, étant donné que les instruments joueront de façon désynchronisée et relativement peu en harmonie.

Les mélanges mettent en jeu les instruments suivants : le Basson (Bo), le Violoncelle (Co), la Flûte (Fl), le Hautbois (Ob), le Violon (Vl) et le Violon alto (Va). La base est composée de 52 échantillons de 10 secondes, avec 13 échantillons par cardinal d'ensemble (compris entre 1 et 4).

5.1.3.3 ENS2

La base de données ENS2 est composée de musique d'ensemble en stéréo extraite de CDs du commerce, et a été constituée par David Sodoyer. La version mono de cette base est obtenue en effectuant la moyenne des deux canaux. Les instruments sont les mêmes que ceux présents dans ENS1. Cette base contient également une cinquantaine d'extraits de 10 secondes, chacun provenant de CDs différents, avec le même nombre d'extraits pour chaque cardinal (compris entre 1 et 4).

5.1.4 Autres bases

Deux autres bases seront également utilisées pour les applications, PIANO et COD.

5.1.4.1 PIANO

La base piano est composée de signaux générés à partir de trois fichiers MIDI avec la base de son (*soundfont*) standard utilisée par le logiciel Timidity (T.Toivonen, n.d.). Cette base a été proposée par Adrien Daniel (stagiaire à Télécom Paris) lors de son stage de master (Daniel, 2007). Les sons ne sont pas extrêmement réalistes mais permettent de réaliser une transcription avec une référence contrôlée. Les morceaux utilisés sont :

- Bach, Prélude en do mineur BWV 817 (13 premières secondes)
- Debussy, Suite bergamasque, III. Clair de Lune (20 premières secondes)
- Mozart, Sonate en Ré Majeur, KV 311/ I.Allegro con spirito (13 premières secondes)

5.1.4.2 COD

La base de données COD contient les extraits qui seront testés avec le codeur que nous avons développé. Elle est composée de 5 soli (Cl, Co, Fl, Ob, Vl) et 4 duos (Cl& Fl, Co&Fl, Co&Vl, Fl&Fl). Les 5 soli sont extraits d'une des bases SOLO, et les duos de la base DUO.

Chaque signal dure aux alentours de 10 secondes.

5.2 Apprentissage

5.2.1 Les bases de données utilisées

Pour l'apprentissage d'atomes, nous utiliserons les bases de données ISO et SOLO1. L'apprentissage concerne uniquement les vecteurs d'amplitudes des partiels A.

5.2.2 Apprentissage de vecteurs d'amplitude sur des notes isolées

Dans le cas des notes isolées, les vecteurs d'amplitudes de partiels $\{A_{i,p,k}\}_{k=1...K}$ sont appris pour chaque classe $\mathcal{C}_{i,p}$ sur les trois bases mentionnées dans le chapitre précédent (IOWA, SOL, RWC), constituant la base ISO. La technique d'apprentissage est supervisée : la hauteur de note p est connue car annotée. Il s'agit donc de faire correspondre un atome harmonique h au signal dans une zone de fréquence fondamentale bornée autour de l'annotation. La technique se distingue donc des méthodes moins supervisées présentées par Aharon et al. (2006); Lesage (2007), qui n'imposent pas de structures harmoniques aux atomes appris : les méthodes basées sur les K-SVD permettent d'obtenir des dictionnaires avec lesquels la représentation obtenue par un algorithme de décomposition sera parcimonieuse, sans pour autant que la structure harmonique soit explicite. La contrainte que nous nous fixons d'étudier des signaux provenant de sources quasi-harmoniques, ainsi que la connaissance des hauteurs des notes des signaux d'apprentissage, permet dans notre cas de restreindre la structure, ce qui a comme avantage corollaire de diminuer le nombre de paramètres définissant chaque atome, et ainsi de diminuer le temps de calcul des produits scalaires dans la décomposition.

Pour chaque note isolée, le signal est découpé en trames de la taille de l'échelle de l'atome à apprendre. Afin d'éliminer la partie transitoire des notes, généralement peu harmonique, les atomes seront extraits sur des trames se situant à partir de la trame d'énergie maximale de la note, et jusqu'à ce que l'énergie passe en-dessous d'un ratio $1/20$ de la valeur du maximum. Les amplitudes de partiels sont calculées sur chacune de ces trames d'entraînement grâce à la formule :

$$a_m = \frac{|\langle x, g_{s,u,m\times f_0,m\times c_0}\rangle|}{\left(\sum_{m'=1}^{M} |\langle x, g_{s,u,m'\times f_0,m'\times c_0}\rangle|^2\right)^{1/2}} \tag{5.1}$$

où f_0 et c_0 sont ajustés de sorte à maximiser le RSR sur cette trame, en utilisant l'algorithme du gradient conjugué décrit dans la section 4.6.1. Le vecteur des amplitudes est ensuite associé à la hauteur de note p correspondant à f_0. Les nombres résultants de vecteurs A par instrument et par hauteur de note sont indiqués dans le tableau 5.3.

5.2.3 Réapprentissage sur des soli

Le point faible de l'apprentissage des atomes sur des notes isolées est que les conditions d'enregistrement des bases de sons utilisées sont assez éloignées de conditions standard. Il est donc nécessaire d'introduire des connaissances permettant de tenir compte de conditions d'enregistrement réalistes, notamment des effets de salle.

Nous avons vu que nous disposions de bases de données de soli extraits d'enregistrement commerciaux, dont les conditions d'enregistrements sont plus proches des signaux sur lesquels nos algorithmes seront testés (musique monophonique ou polyphonique dans des enregistrements commerciaux). Cependant, ces enregistrements ne sont pas segmentés et annotés note à note. Réaliser cette opération manuellement serait coûteuse en temps, et le faire grâce à des algorithmes de détection d'onset et de pitch nous soumettrait à leur performances imparfaites.

Nous allons donc tirer parti des algorithmes de décomposition présentés dans le chapitre 4. Pour apprendre des atomes "réalistes" d'un instrument i, nous utilisons un des algorithmes de poursuite avec un dictionnaire contenant uniquement des atomes de l'ins-

Instrument	N_i	N_{ip} moyen
As	29565	629
Ba	33189	722
Bo	20584	479
Cl	33446	712
Co	32029	654
Fl	24314	608
Fh	21756	444
Ob	15804	452
Pn	190774	2120
Ss	16919	498
Ta	13769	313
Tb	22682	493
Tr	26606	665
Va	30128	615
Vl	44142	817

TAB. 5.3 – *Nombre total de trames d'apprentissage par instrument i et nombre moyen par classe d'instrument et de hauteur (i, p) dans la base ISO.*

trument i appris sur des notes isolées. Nous effectuons cependant la poursuite avec les modifications suivantes :

- **à l'étape de sélection** : un vecteur A est calculé en utilisant l'équation (5.1), en fixant f_0 égal à la fréquence fondamentale de l'atome sélectionné, puis stocké afin de constituer un ensemble d'atomes collectés.
- **à l'étape de soustraction** : au lieu de soustraire l'atome sélectionné du signal, le signal est mis à 0 sur la plage temporelle correspondant à l'atome sélectionné, en multipliant le signal par la fonction suivante :

$$\omega(t) = \mathbf{1}_{[u,u+s]} \cos(2\pi(\frac{t-u}{s})) \tag{5.2}$$

  Cette opération empêche l'algorithme d'extraire des atomes sur des zones préalablement perturbées par l'extraction d'autres atomes.

Une fois ce processus achevé, un dictionnaire d'atomes appris peut être construit. On peut remarquer que la base de soli sur laquelle on décompose doit être suffisamment grande pour contenir des notes couvrant tout le registre de chaque instrument. Si aucun atome n'a été trouvé pour une certaine note p, le sous-dictionnaire de la hauteur précédente $p-1$ est utilisé pour le remplacer.

Bien qu'issus d'une approche moins supervisée que sur des notes isolées (les hauteurs de notes dans les soli ne sont pas connues), les atomes appris ont des enveloppes spectrales plus proches de celles qu'on peut trouver dans des signaux réalistes. De plus, concernant les instruments pouvant réaliser des notes doubles, comme les instruments à cordes frottées, la perturbation induite par la présence d'une seconde note peut être incluse dans le modèle appris.

En utilisant ces données, le nombre d'atomes collectés est indiqué dans le tableau 5.4.

Instrument	N_i	N_{ip} moyen
Bo	2457	60
Cl	9048	193
Co	13868	285
Fh	4072	89
Fl	13216	330
Ob	5912	169
Va	10170	216
Vl	37749	700

TAB. 5.4 – *Nombre total de trames d'apprentissage par instrument i et nombre moyen par classe d'instrument et de hauteur (i, p) dans la base SOLO1.*

5.2.4 Réduction des dictionnaires par quantification vectorielle

La taille du dictionnaire varie linéairement en fonction du nombre de vecteurs d'amplitudes. Comme le nombre original de vecteurs est trop grand pour rendre les décompositions efficaces d'un point de vue du temps de calcul, le nombre de vecteurs est réduit par quantification vectorielle, ceci pour chaque classe $\mathcal{C}_{ip}$.

Nous avons choisi d'utiliser l'algorithme des K-moyennes (*K-means*) avec une distance euclidienne, qui permet de représenter les atomes par les centroïdes des principaux amas. Le choix de la distance euclidienne est naturel étant donné que la classification des vecteurs de partiels (3.8) se fait en utilisant une maximisation du produit scalaire canonique dans $\mathbb{R}^M$. L'algorithme de K-Means permet en outre d'éviter le surapprentissage et d'enlever les *outliers*.

On aurait également pu utiliser un algorithme de *K-medoïdes*, permettant de sélectionner des exemples appartenant à l'ensemble d'origine, contrairement à l'algorithme des K-means où les vecteurs retenus n'appartiennent pas à l'ensemble de départ.

5.2.5 Bilan

Nous avons présenté des méthodes permettant d'obtenir des dictionnaires d'amplitudes A à partir de notes isolées annotées en hauteur de note et instrument, et à partir de performances en solo uniquement annotées en instrument. Une quantification vectorielle permet ensuite à la fois de réduire la taille du dictionnaire, d'équilibrer les cardinaux des classes $\mathcal{C}_{ip}$ et de les débruiter.

On peut émettre deux remarques concernant cet apprentissage. La première concerne la technique de quantification : étant donné que la plupart des tâches effectuées feront intervenir une classification des instruments, on peut imaginer mettre en oeuvre une technique de *clustering* discriminatif, exhibant les exemples de chacune des classes qui permettent le mieux de différencier les instruments à partir du produit scalaire. La seconde est que les instruments sont décrits de façon peu compacte : les modèles entre deux hauteurs de notes d'un même instrument n'ont aucune relation explicite. Il pourrait être intéressant de travailler sur une modélisation des instruments unique pour toutes les hauteurs de notes, ce qui permettrait de mieux contrôler l'échantillonnage du modèle. Procéder à une telle opération permettrait également de mettre en oeuvre plus facilement une adaptation du modèle aux signaux étudiés.

Chapitre 6

Applications

Le but de ce chapitre est de montrer que les représentations obtenues grâce aux algorithmes présentés dans le chapitre 4 peuvent servir à un grand nombre d'applications, notamment à la reconnaissance d'instruments de musique dans de la musique multi-instruments, la transcription automatique, le codage objet, et différentes tâches d'édition musicale. Les post-traitements présentés pour les applications étudiées restent ici relativement simples, cependant ils permettent d'illustrer l'applicabilité de nos décompositions aux tâches envisagées. Nous verrons que pour certaines applications, notamment la reconnaissance d'instruments en monophonique et l'estimation de hauteur, les résultats sont comparables à l'état de l'art pour des algorithmes spécifiques à une tâche.

6.1 Visualisation

Si l'on représente les livres obtenus à partir des décompositions, nous obtenons des visualisations intéressantes du signal. Les Figures 6.1 et 6.2 montrent des décompositions obtenues à partir d'une performance solo de flûte et une autre de clarinette, avec différents algorithmes : une décomposition avec un Matching Pursuit, sans et avec optimisation des paramètres f_0 et c_0, et un algorithme moléculaire avec optimisation des paramètres.

La Figure 6.3 représente cette fois-ci un duo. Ici, la partition originale (annotée manuellement à l'aide du chromagramme de Sonic Visualizer) permet de souligner la proximité entre les localisations respectives des notes jouées et des atomes dans le plan temps-hauteur.

Enfin, nous pouvons également visualiser une décomposition d'un morceau de piano, avec plusieurs résolutions (4 échelles différentes : 1024, 2048, 4096, 8192 à 44100 Hz) sur la Figure 6.4. La décroissance de l'amplitude des notes est bien visible sur l'amplitude des atomes sélectionnés.

On peut également visualiser la sortie d'un algorithme stéréo en trois dimensions sur la Figure 6.5.

L'effet du post-traitement présenté dans la section précédente sur un livre obtenu avec une décomposition avec un seuil d'arrêt en RSR élevé est présenté sur la Figure 6.6.

Nous remarquons donc que les décompositions présentées permettent d'exhiber un grand nombre d'informations pertinentes pour l'indexation audio : les atomes et molécules s'adaptent aux structures harmoniques présentes dans le signal. Les paramètres

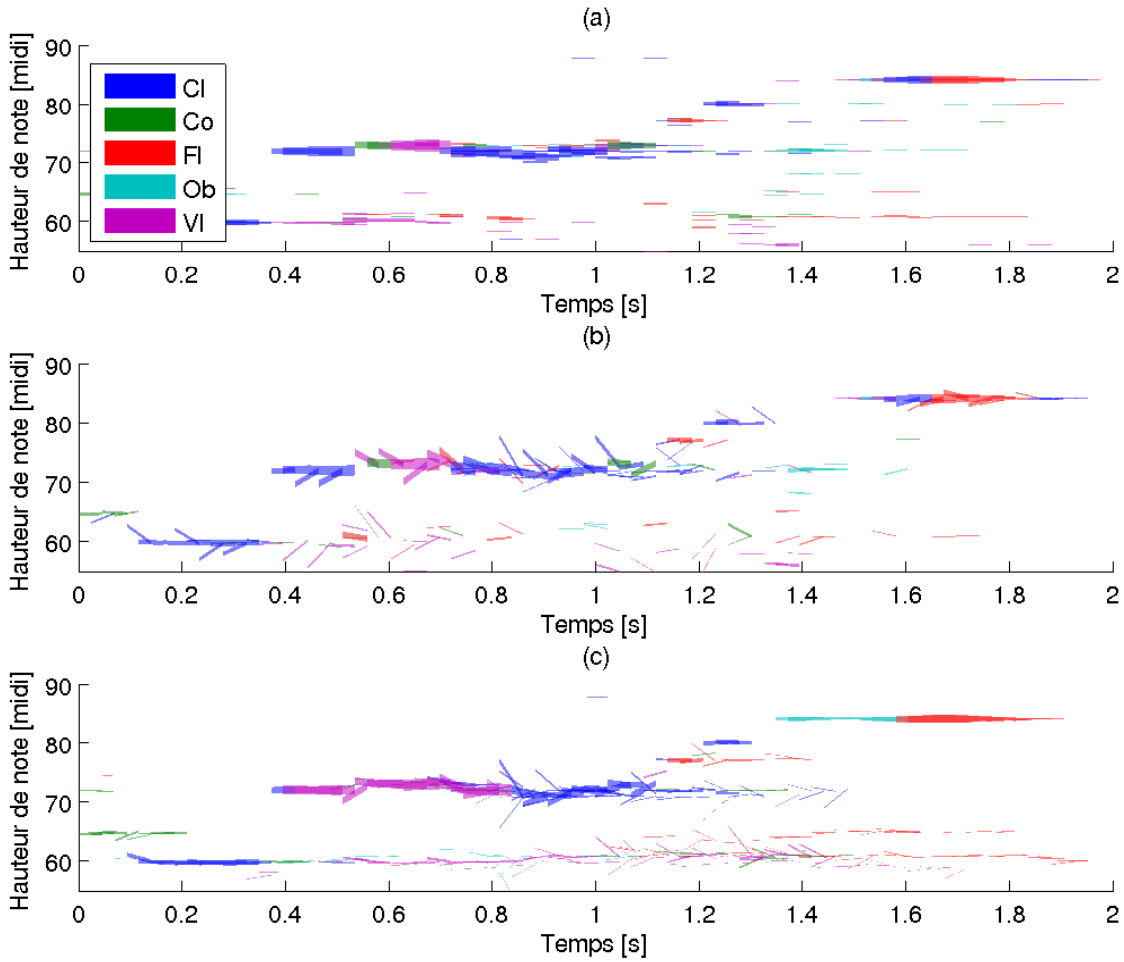


FIG. 6.1 – *Visualisation de décompositions d'un solo de clarinette. Chaque atome est représenté par un parallélogramme centré sur ses coordonnées temps-hauteur (u, f_0), dont la largeur, la hauteur, l'inclinaison sont respectivement proportionnelles à son échelle s, son poids α_λ et taux de modulation de fréquence c_0. (a) Matching Pursuit, (b) Algorithme atomique avec estimation du taux de chirp, (c) Algorithme moléculaire avec estimation du taux de chirp. Pour ces trois algorithmes, le seuil d'arrêt est 20 dB de RSR ou 250 atomes par seconde.*

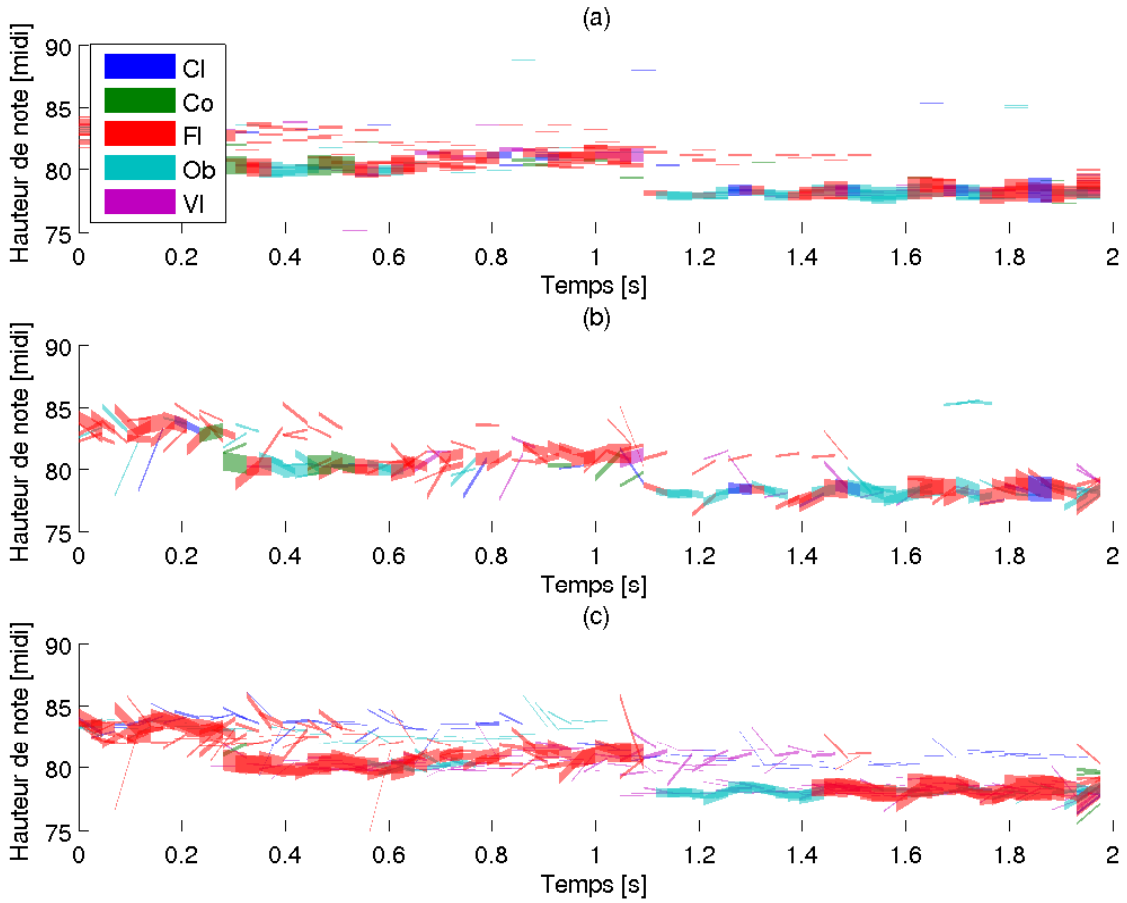
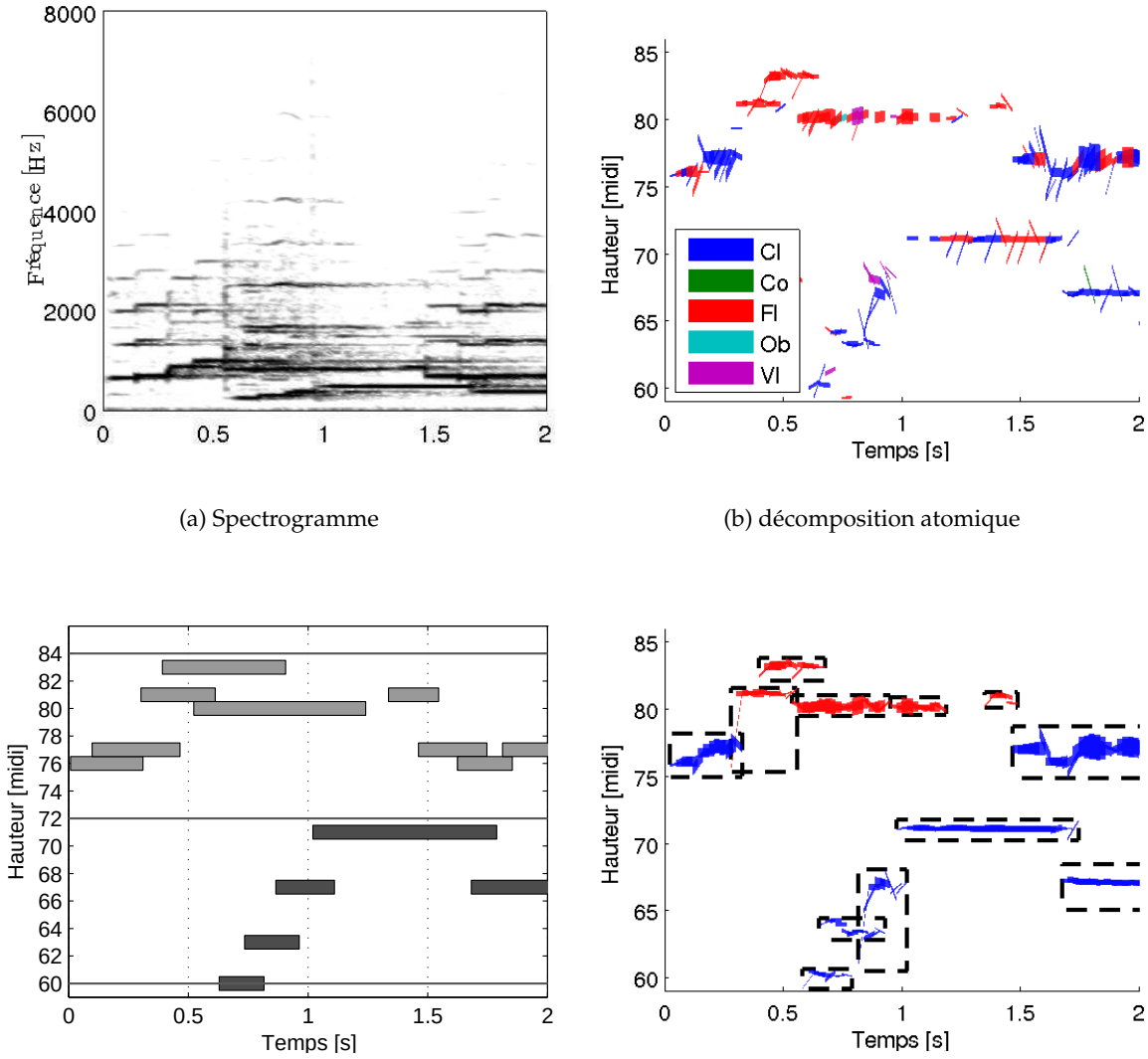


FIG. 6.2 – *Visualisation de décompositions d'un solo de flûte. Chaque atome est représenté par un parallélogramme centré sur ses coordonnées temps-hauteur (u, f_0), dont la largeur, la hauteur, l'inclinaison sont respectivement proportionnelles à son échelle s, son poids α_λ et taux de modulation de fréquence c_0. (a) Matching Pursuit, (b) Algorithme atomique avec estimation du taux de chirp, (c) Algorithme moléculaire avec estimation du taux de chirp. Tous les algorithmes sont arrêtés à 250 atomes par secondes ou 20 dB.*



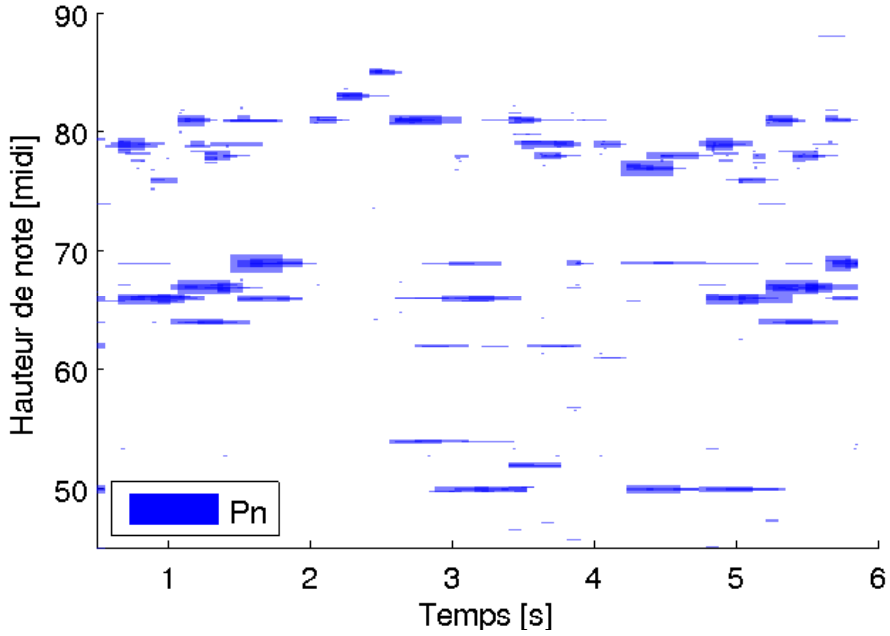
(a) Spectrogramme

(b) décomposition atomique

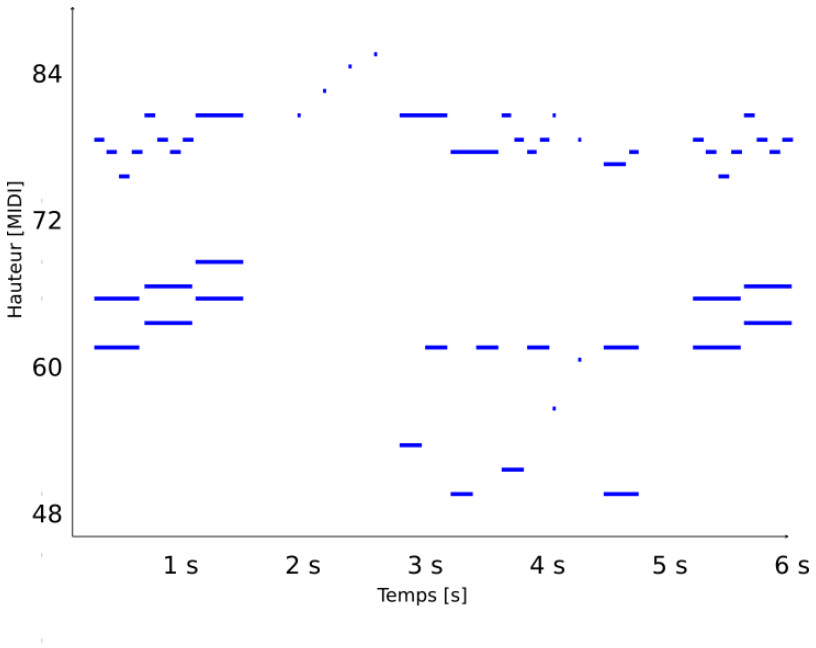
(c) Visualisation Piano Roll

(d) décomposition moléculaire

FIG. 6.3 – *Visualisation d'un duo flûte - clarinette, comparé à la visualisation piano roll annotée à la main. Chaque atome est représenté par un parallélogramme centré sur ses coordonnées temps-hauteur (u, f_0), dont la largeur, la hauteur, l'inclinaison sont respectivement proportionnelles à son échelle s, son poids α_λ et taux de modulation de fréquence c_0. Chaque molécule est représentée par un rectangle en pointillé couvrant plusieurs atomes. L'échelle des gris indique l'instrument associé à chaque atome.*



(a) Visualisation du livre



(b) Piano Roll original

FIG. 6.4 – *Visualisation de la décomposition d'un extrait de piano généra avec Timidity avec un Matching Pursuit (seuil d'arrêt à 20 dB ou 250 atomes par seconde) et du* Piano Roll *du fichier MIDI original. Chaque atome est représenté par un rectangle centré sur ses coordonnées temps-hauteur* (u, f_0)*, dont la largeur, la hauteur sont respectivement proportionnelles à son échelle* s*, et son poids* α_λ*.*

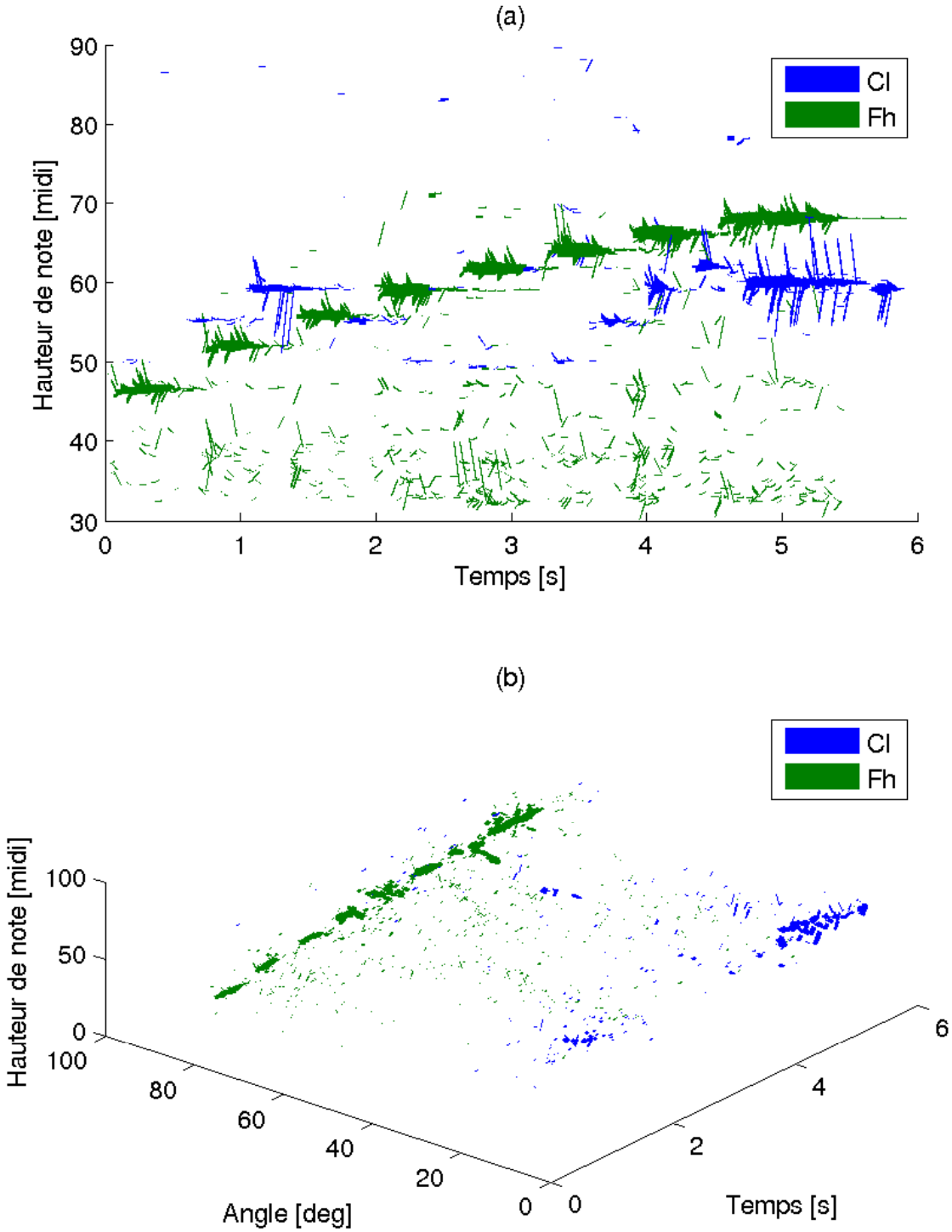


FIG. 6.5 – *Visualisation de décompositions d'un extrait d'un duo stéréo synthétique de Clarinette et Cor (sans post-traitement). Les paramètres de panoramique respectifs sont 10º et 76 º.(a) Vue temps-hauteur, (b) Vue temps-hauteur-angle.*

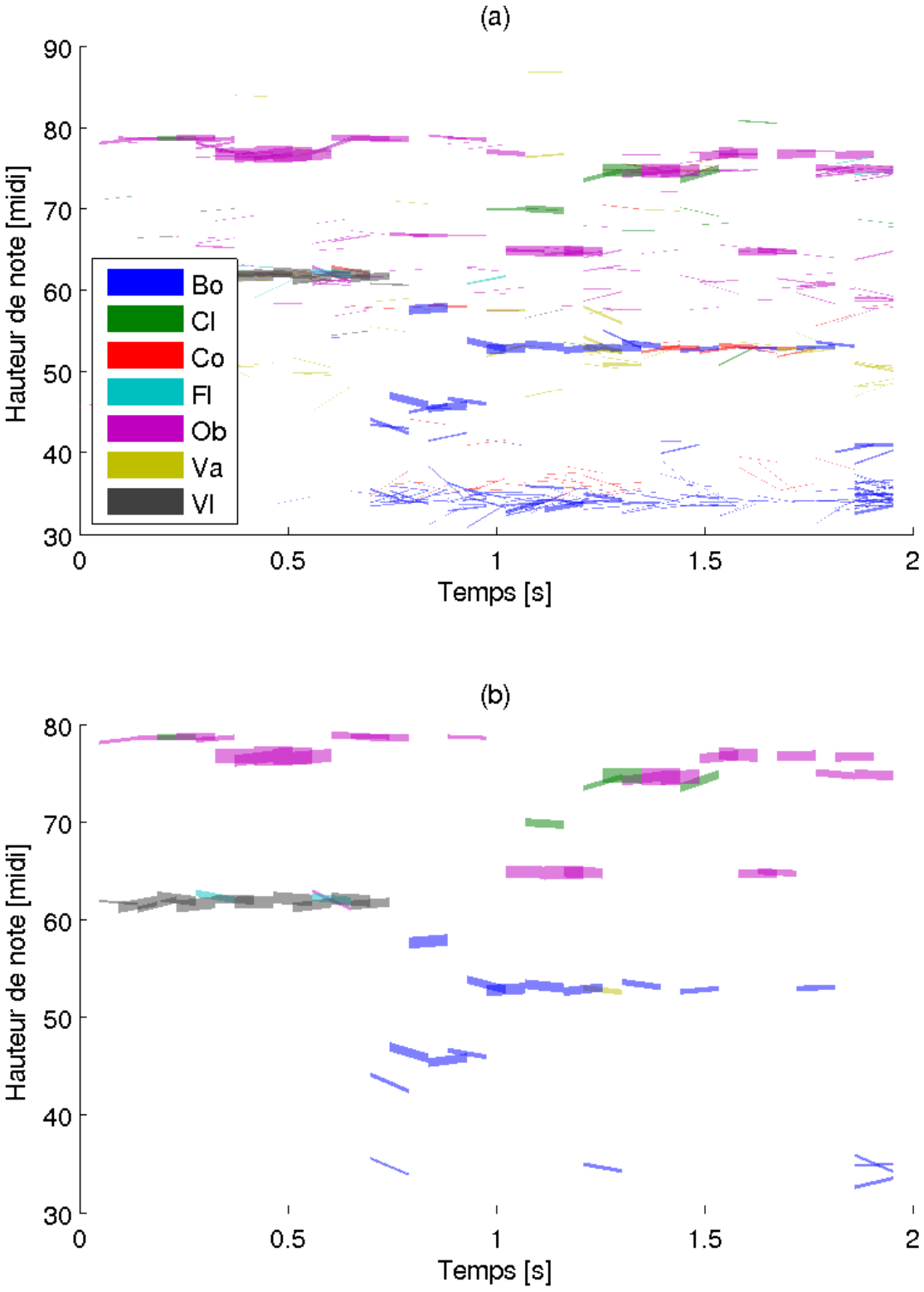


FIG. 6.6 – *Effet du post-traitement pour la parcimonie en hauteur sur un duo Flûte-Basson (mêmes paramètres que précédemment). (a) Livre avant post-traitement (b) Livre après post-traitement pour la parcimonie en hauteur ($\beta = 0.2$).*

présentés sont les hauteurs de notes, les modulations d'amplitude et de fréquence (notamment pour les algorithmes moléculaires) et la source instrumentale d'où provient l'atome.

On peut également souligner les apports des différents algorithmes les uns par rapport aux autres. L'algorithme atomique permettant d'extraire des atomes modulés en fréquence qui permettent de mieux suivre les variations de fréquence fondamentale instantanées des notes analysées, ce qu'on peut observer sur les décompositions de signaux de flûte. Les notes présentant de telles variations sont également représentées avec moins d'atomes en utilisant des atomes modulés en fréquence qu'en utilisant des atomes "plats". Néanmoins, pour ces deux algorithmes, on observe que les amplitudes des atomes ne suivent pas toujours les amplitudes instantanées des notes jouées. Pour le second, certains atomes possèdent des valeurs de taux de modulation de fréquence fondamentale erratiques bien qu'ils soient à la bonne fréquence fondamentale. Ces deux aspects sont dus à la nature gloutonne de l'algorithme employé, comme nous l'avons mentionné en 4.2.3. On peut observer que l'algorithme moléculaire permet d'éliminer cet artefact : les paramètres de chaque atome d'une molécule sont estimés avant la soustraction, et suivent donc fidèlement les structures harmoniques sous-jacentes.

On peut signaler que toutes ces informations peuvent être extraites grâce à d'autres méthodes, par exemple en mettant en jeu des processus de séparation de sources, de reconnaissance d'instruments sur les sources séparées ou de transcription automatique. Cependant, les décompositions que nous présentons permettent d'obtenir toutes ces informations à la fois, avec une précision moindre qu'une approche par estimation jointe des sources activées (Vincent (2006)) mais avec une complexité nettement moindre.

Nous allons maintenant évaluer la précision de nos algorithmes concernant les paramètres haut-niveau qui peuvent être extraits à partir des résultats des décompositions.

6.2 Reconnaissance des instruments : le cas mono-instrument

La sélection d'atomes d'un instrument donné est un indice sur les instruments mis en jeu dans un signal musical. Avant d'évaluer des taux de reconnaissance à partir de décompositions sur des signaux réels, nous allons tout d'abord examiner le pouvoir discriminant des vecteurs d'amplitudes quant aux classes d'instruments.

Dans toutes les expériences présentées dans cette section, cinq instruments seront mis en jeu : la Clarinette (Cl), le Violoncelle (Co), la Flûte (Fl), le Hautbois (Ob) et le Violon (Vl).

6.2.1 Classification des atomes sans décomposition

Comme nous l'avons vu, l'algorithme de Matching Pursuit consiste à sélectionner à chaque itération l'atome qui maximise la multiplication du produit scalaire entre les enveloppes des partiels normalisées par la racine carrée d'une somme spectrale. A sommes spectrales égales, ce qui est le cas pour deux atomes de même fréquence fondamentale, le choix de l'instrument se fait sur la valeur de ce produit scalaire.

Nous pouvons évaluer l'efficacité de ce type de classification en l'évaluant sur le dictionnaire des vecteurs d'amplitudes. Etant donné une hauteur p, un vecteur B de classe $\mathcal{C}_{ip}$ (instrument i et hauteur de note p) est bien classifié si le vecteur d'amplitudes $\hat{A}$ tel

que :

$$\hat{A} = \arg\max_{j \in \mathcal{I}} \{\langle A, B\rangle | A \in \mathcal{C}_{jp}, A \neq B\} \tag{6.1}$$

appartient à la classe $\mathcal{C}_{ip}$.

Les tests suivant ont été réalisés : pour chaque pitch p, un test en prenant l'ensemble ISO pour l'apprentissage et SOLO1 pour le test, et un autre en faisant l'inverse. Nous travaillerons sur les dictionnaires déjà quantifiés par l'algorithme des K-moyennes.

Comme la sélection d'une classe d'instrument équivaut à un algorithme des K plus proches voisins (K-PPV) avec $K = 1$, nous avons testé l'algorithme avec K plus grand. Les résultats sont légèrement meilleurs en utilisant $K = 2$, cependant il ne semble pas indispensable de réfléchir à une adaptation de cet algorithme à la décomposition.

Algorithme	ISO → SOLO1	SOLO1 → ISO
1-PPV	56.9	56.7
2-PPV	57.7	56.2
4-PPV	57.5	55.7
8-PPV	56.7	53.9

TAB. 6.1 – *Classification des dictionnaires en utilisant l'algorithme des K-PPV*

Algorithme	0 → 1	1 → 0
dB(A)	56.0	56.5
dB(B)	56.8	56.7
dB(C)	56.9	56.7
dB(D)	57.8	55.8
formants	56.9	56.7

TAB. 6.2 – *Influence de la pondération sur la classification des dictionnaires ($K = 16$)*

Nous avons également étudié l'influence de différents types de pondération psycho-acoustique standard sur les résultats de classification. La Figure 6.7 montre l'allure des pondérations utilisées : elles visent à augmenter l'influence des parties fréquentielles auxquelles l'oreille est plus sensible. Une restriction du domaine fréquentiel a aussi été testée : elle consiste à ne garder que la zone contenant les principaux formants des instruments de musique, c'est à dire la zone de fréquences 200-4000Hz. Utiliser ces pondérations consiste à multiplier point à point les vecteurs d'amplitudes par les valeurs des courbes aux fréquences des partiels correspondants. Si l'on utilise ces pondérations dans une décomposition, il faut de plus normaliser les vecteurs d'amplitude à 1 et appliquer le filtrage au signal.

Comme le montre le Tableau 6.2, les pondérations n'ont pas une grande influence sur les résultats de classification d'enveloppes. Cependant, il peut être intéressant de noter que l'identification des instruments reste correcte si l'on prend une bande passante assez faible (pondération formants). Cet aspect pourra être utilisé pour obtenir des décompositions plus rapides, en restreignant le nombre de composantes à multiplier lors des produits scalaires.

La Figure 6.8 montre le taux de réussite de la classification en fonction de la hauteur de note. Les résultats de la classification sont comparés au hasard, obtenu en prenant l'inverse du nombre de classe d'instrument pouvant jouer la note p. On remarque que

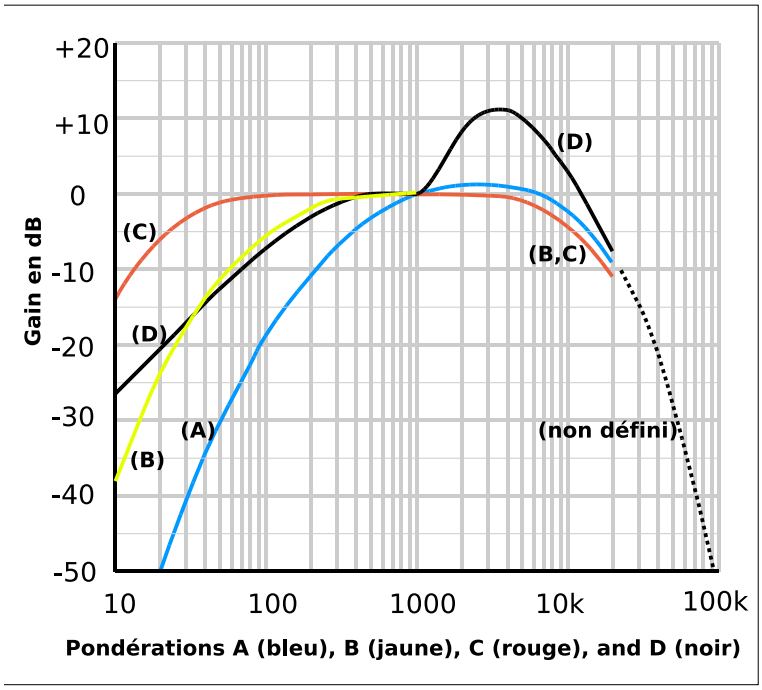


FIG. 6.7 – *Courbes de pondération psychoacoustique standard (source :Wikipedia).*

le taux de reconnaissance est plus bas lorsque le nombre d'instruments pouvant jouer la note est plus important (hasard à 20 %), ce qui est tout à fait normal. On en conclut que les atomes en marge de cette région seront très importants pour déterminer l'instrument qui joue.

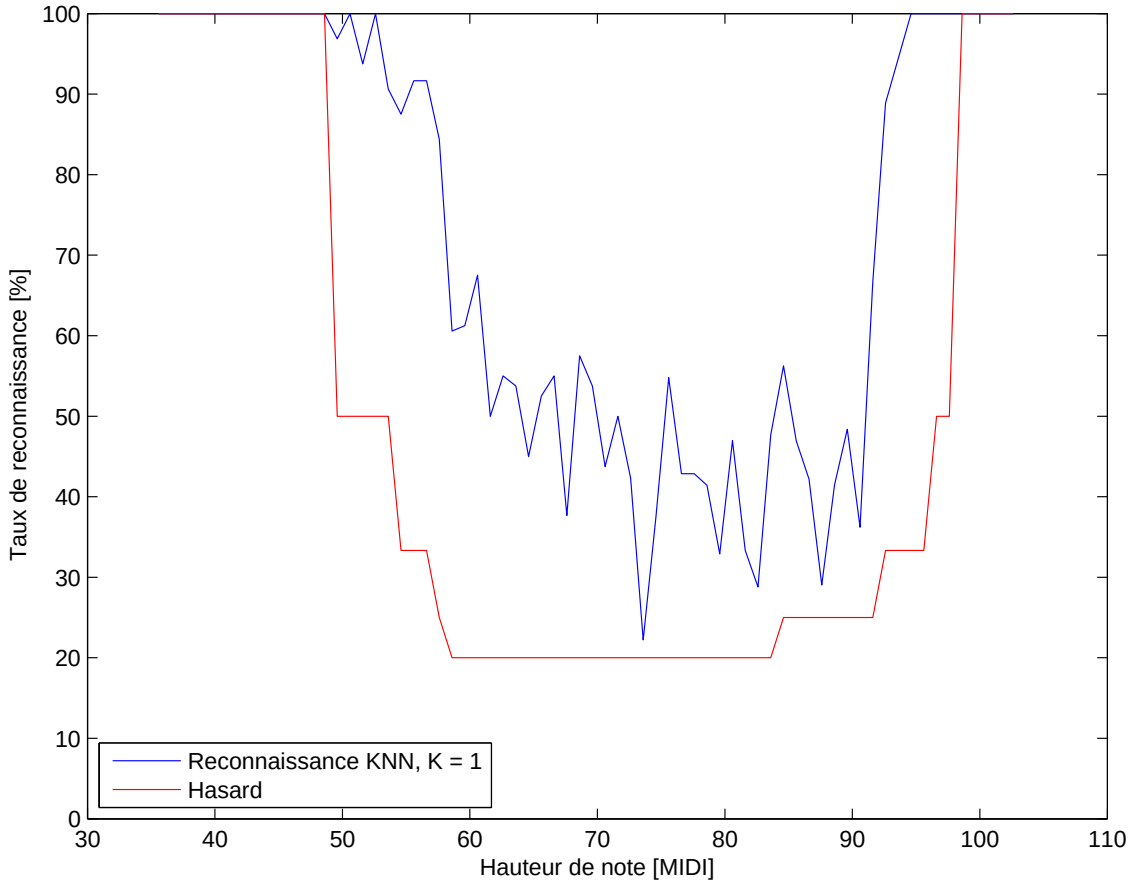


FIG. 6.8 – *Reconnaissance de vecteurs d'amplitudes sans décomposition : influence de la hauteur de note sur la discrimination des vecteurs d'amplitudes. Test sur cinq instruments (Cl, Co, Fl, Ob, Vl)*

6.2.2 Reconnaissance de segments de performances solo

La reconnaissance des instruments de musique sur des phrases solo a été abordée par Brown (1999); Martin (1999); Eronen & Klapuri (2000); Essid et al. (2006b) avec des approches par sac de trames. Les performances sont maintenant proches de ce que des musiciens experts peuvent réaliser. Dans (Martin, 1999), des musiciens experts ont réalisé des scores d'indentification de 67 % pour l'identification d'extraits de 10-s parmi 27 instruments, alors que le système complet décrit dans (Essid, 2005) atteint 70 % pour un cas de figure similaire.

Comme nous l'avons présenté dans la section 4, les algorithmes de décomposition permettent d'effectuer une classification itérative du signal analysé. Le passage à l'analyse d'un signal induit deux difficultés par rapport à la classification d'enveloppes pré-

sentée dans la section précédente :

1. Les peignes du signal qui sont analysés lors de la décomposition, notés B dans la section 3.8, sont paramétrés par la fréquence fondamentale des atomes du dictionnaire. Or f_0 est échantillonnée sur une grille, et donc le vecteur B n'est pas nécessairement strictement aligné avec la structure harmonique du signal la plus pertinente. Il faudra donc veiller à ce que la fréquence fondamentale f_0 soit échantillonnée suffisamment finement pour que les vecteurs B évalués soient proches des meilleures structures harmoniques du signal. On peut noter que le défaut d'alignement peut être compensé par la largeur de lobe fréquentiel des atomes de Gabor utilisés pour modéliser les partiels. Il y a donc un compromis à effectuer : un lobe trop fin ne permet pas de compenser les erreurs d'alignement, et un lobe trop large implique un mauvais critère de quasi-orthogonalité entre les partiels.
2. Il faut fusionner les décisions prises au niveau de chaque atome sélectionné afin d'obtenir un score sur le segment de signal analysé, ce que nous détaillons dans la suite de cette section.

Dans un souci de cohérence et de robustesse, nous cherchons idéalement des méthodes qui conviennent à des décompositions quasi-infinies (définies en 4.9), c'est-à-dire qui produisent des scores pour chaque instrument qui convergent vers une valeur finale lorsque le nombre d'atomes n tend vers l'infini. Bien entendu, en pratique, les décompositions auront un critère d'arrêt pour lequel on supposera que la majeure partie de l'information harmonique du signal a été extraite. Etant donné un ensemble d'atomes sélectionnés, plusieurs stratégies peuvent être envisagées :

- *vote majoritaire (VM)* : Cette méthode consiste à compter, pour chaque instrument du dictionnaire, les atomes extraits dont le vecteur A vient de l'instrument considéré. L'instrument choisi est alors l'instrument qui possède le plus d'atomes. Cette méthode a l'inconvénient de dépendre de la profondeur de la décomposition car elle attribue la même importance à tous les atomes, quelle que soit leur énergie.
- *vote majoritaire pondéré (VMP)* : Pour contourner le problème évoqué ci-dessus, on peut mettre en place une pondération des votes : plus l'atome est énergétique, plus il a une influence sur le score final. Le score obtenu est donc la somme des poids portés par les atomes de chaque classe. Afin de pondérer l'importance de l'amplitude dans la décision, on définit un coefficient γ tel que le poids $\mathcal{W}(h_\lambda)$ s'exprime :

  $$\mathcal{W}(h_\lambda) = |\alpha_\lambda|^\gamma \tag{6.2}$$

  Le score $\mathcal{S}_i$ d'un instrument est alors donné par :

  $$\mathcal{S}_i = \sum_{\lambda \in \Lambda} \mathcal{W}(h_\lambda) \tag{6.3}$$

  Pour γ = 2, on a alors une pondération par l'énergie portée par l'atome. Ainsi, des valeurs de gamma faibles agissent comme une compression de la dynamique sur le poids des atomes d'un segment et réduisent donc l'influence de l'amplitude des atomes. Les valeurs élevées de γ correspondent à une expansion, et permettent donc de priviligier l'influence relative des poids de forte énergie sur le score final.
- *scores probabilisés avec hypothèse d'indépendance (Prob)* : Dans les deux méthodes précédentes, on considère que chaque atome porte une décision "dure" concernant l'instrument qu'il représente. On peut également ne pas prendre de décision pour

chaque atome en gardant des *saillances d'instrument* qui peuvent être extraites dès l'étape de décomposition. En effet, comme nous l'avons indiqué en 3.8, le produit scalaire entre les atomes et le signal peut être vu comme une saillance de fréquence fondamentale et d'instrument. Pour garder une indécision au niveau de la sélection de l'atome, il suffit alors de garder les valeurs des produits scalaires entre les atomes de la même fréquence fondamentale et de même support temporel et les enveloppes du signal. Par exemple, supposons qu'un atome de flûte soit extrait du signal à la hauteur 65 et avec une amplitude de 2,5. Dans ce cas, on retiendra que la saillance de flûte pour cet atome est 2,5, et on gardera en mémoire les valeurs des produits scalaires entre le signal et les meilleurs atomes des autres instruments à cette hauteur et cette localisation temporelle (par exemple 2,3 ; 1,5 ; 1,6 ; 0,7).
Plus formellement, étant donné un atome sélectionné à la fréquence fondamentale $f_{0\lambda}$, à l'échelle s_λ et à la localisation u_λ, on définit alors la saillance de fréquence fondamentale et d'instrument par :

$$S_{\lambda,i} = \max_{A \in \mathcal{C}_{i,p(f_{0\lambda})}} \left\{ |\langle x, h_{s_\lambda,u_\lambda,f_{0\lambda},A,\Phi} \rangle| \right\} \tag{6.4}$$

Si un instrument i ne peut pas jouer à la hauteur $p(f_{0\lambda})$, c'est-à-dire $\mathcal{C}_{i,p} = \varnothing$, les saillances correspondantes sont mises à 0. Ainsi, bien que ces saillances ne soient pas requises pour la décomposition, toutes les saillances d'instrument pour chaque atome sélectionné sont gardées pour l'étape de post-traitement. Afin de pondérer l'influence de l'amplitude de l'atome sur la saillance, on peut comme précédemment mettre les saillances à la puissance γ avant de les sommer sur tout le segment. Le score $\mathcal{S}_i$ d'un instrument i sera donc la somme de toutes les saillances correspondantes contenues dans tous les atomes de la décomposition, ce qui est similaire à une fusion de scores probabilisés en prenant une hypothèse d'indépendance entre les décisions :

$$\mathcal{S}_i = \sum_{\lambda \in \Lambda} S_{\lambda,i}^{\gamma} \tag{6.5}$$

En pratique, l'ensemble d'atomes sur lequel les scores sont calculés peut-être choisi de plusieurs façons. Soit tous les atomes de la décomposition sont gardés, soit ils sont sélectionnés pour leur propension à être des atomes physiques grâce au post-traitement présenté en 4.9. Dans les expériences qui suivront, on effectuera le post-traitement avec $\beta = 0,5$, de telle sorte qu'on garde un atome par trame temporelle, pour un échantillonnage temporel donné. Cette méthode rejoint alors les techniques utilisées dans les approches par sac de trames (Eggink & Brown (2004), Kitahara et al. (2005)). Si tous les atomes sont gardés, nous n'évaluerons pas la méthode par vote majoritaire (VM) car les scores ne peuvent pas être prouvés comme convergents quand n augmente. Par contre, les autres méthodes de vote apparaissent comme converger vers une valeur donnée lorsque le nombre d'atomes extraits devient grand. Ce résultat est à mettre en relation avec la décroissance quasi-exponentielle de l'énergie des atomes que l'on observe quand n augmente, mais qui n'a cependant pas été prouvée. La convergence des scores quand n augmente est illustrée sur la figure 6.9.

Dans les paragraphes suivants, deux ensembles d'atomes seront utilisés pour déterminer les scores de chacun des segments. Le premier sera le livre entier ($\beta = 0$), le second celui obtenu en ne gardant qu'un atome par trame ($\beta = 0.5$). Dans les tableaux indiquant les pourcentages d'identification correcte des instruments, la partie du haut indique les

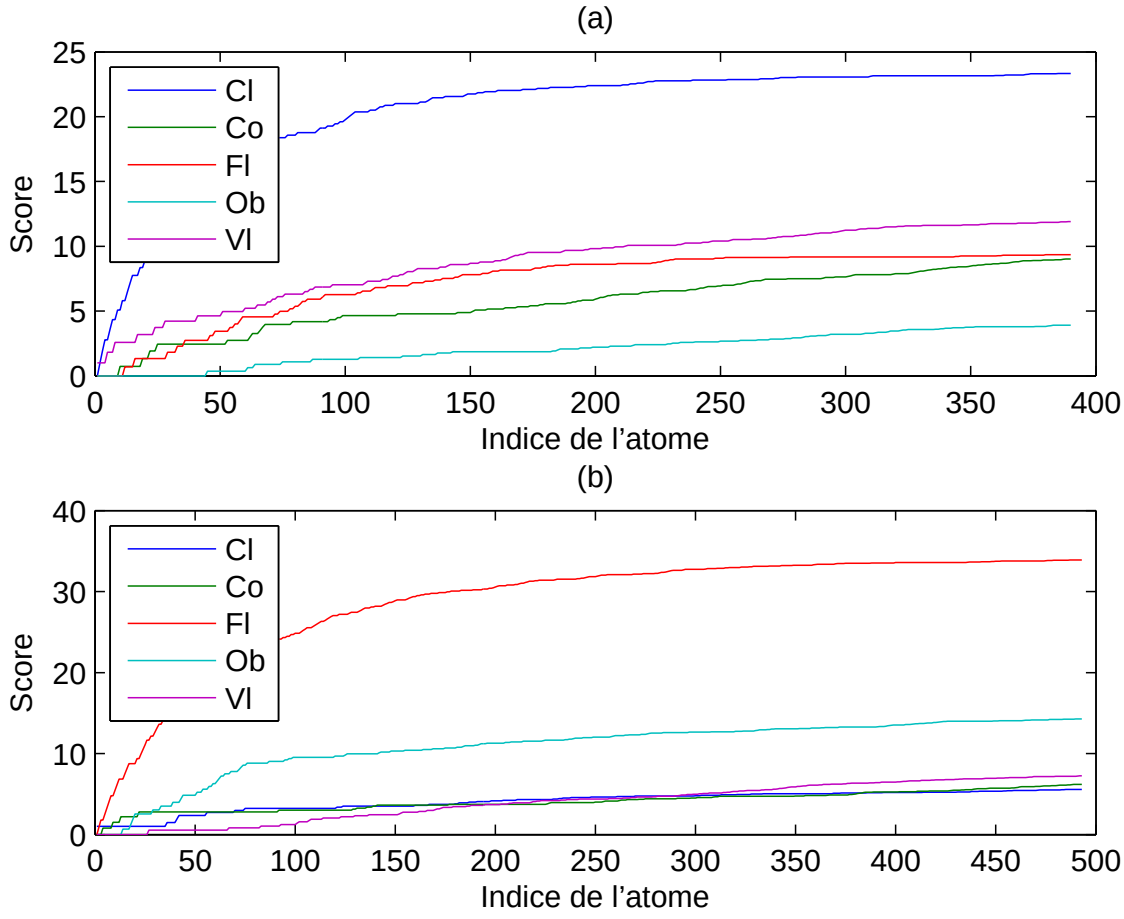


FIG. 6.9 – *Convergence du score $\mathcal{S}_i^l$ de reconnaissance d'instrument sur un segment de 2 secondes de clarinette (a) et un autre de flûte (b) en fonction de la profondeur de la décomposition en nombre d'atomes. Algorithme de vote : vote majoritaire pondéré (VMP).*

résultats sur le premier ensemble, celle du bas le résultat sur le second.

6.2.3 Expériences sur des performances soli

Les paramètres par défaut pour les décompositions utilisées dans ces expériences sont les suivants :

- La fréquence fondamentale f_0 est échantillonnée logarithmiquement à un dixième de ton,
- L'échelle s est fixé à 46 ms (1024 échantillons à 22010 Hz), Δu à 23 ms,
- L'échantillonnage du dictionnaire d'amplitudes K est fixé à 16 éléments par classe $\mathcal{C}_{ip}$,

Les expériences sur les soli ont été réalisées en apprenant les dictionnaires sur les ensembles ISO ou SOLO1, et en effectuant les tests sur l'ensemble SOLO2 (sources distinctes de SOLO1). Les durées de décision sont de 2 secondes. Pour cette durée, le nombre total d'échantillons de test est d'environ 1300. Pour un score d'environ 85 %, l'intervalle de confiance à 95 % est d'un peu moins de ± 1 point, sous une hypothèse abusive d'indépendance entre les éléments du tests. En effet, on ne peut pas considérer deux extraits d'une même pièce comme décorrélés, étant donné que les instruments et les conditions de jeu sont les mêmes pour toute la pièce.

Afin d'avoir un élément de comparaison avec un bon algorithme existant, nous avons effectué une expérience avec un algorithme de reconnaissance de type sac-de-trames. Il s'agit de l'algorithme d'Essid et al. (2006b). La méthode s'appuie sur une sélection de caractéristiques appelé *Fisher-clustering*, avec laquelle 40 caractéristiques sont retenues parmi 543. Ensuite une classification utilisant des Machines à Vecteurs Supports (SVM) est utilisée, en utilisant une stratégie de décision par paire. Les caractéristiques extraites couvrent une plus grande partie du timbre instrumental que la classification opérée par notre algorithme. En effet, les caractéristiques permettent d'obtenir des informations temporelles à long-terme, comme les modulations d'amplitude et de fréquence, ainsi que des caractéristiques portant sur les parties bruitées ou impulsives du signal. Les résultats obtenus avec cette méthode s'élèvent à 83.9 %.

6.2.4 Influence de l'ensemble d'apprentissage

Etudions tout d'abord l'influence de l'ensemble d'apprentissage sur les résultats de classification. Le Tableau 6.2.4 montre les résultats de classification en utilisant le dictionnaire d'amplitudes appris sur les notes isolées ISO et ceux en utilisant les dictionnaires appris sur des soli (base SOLO1).

On retrouve ici un résultat bien connu en classification statistique : les résultats d'une classification sont très dépendants des similarités entre les ensembles d'apprentissage et de test. Ici l'apprentissage d'atomes sur des solos, bien que moins contrôlé et supervisé que celui sur les atomes de notes isolées, permet d'obtenir des atomes qui sont plus proches des signaux analysés. On peut d'ailleurs noter que si l'on utilise un dictionnaire appris sur l'ensemble SOLO1 et que les tests sont effectués sur le même ensemble, les résultats d'identification dépassent 95 %. Bien que ce résultat n'ait pas un grand intérêt pratique, il montre que l'algorithme de décomposition suivi du post-traitement adéquat peuvent être considérés comme un classifieur valide.

Apprentissage → Test	ISO → SOLO2	SOLO1 → SOLO2
VMP	75	84
Prob	72	82
VM	78	83
VMP	77	82
Prob	76	81

TAB. 6.3 – *Pourcentage d'identification correcte des instruments sur des performances solo : influence de l'ensemble d'apprentissage. Taille des segments de décision : 2 secondes. Haut du tableau : résultats basés sur les livres entiers. Bas du tableau : résultats basés sur les livres avec un atome par trame temporelle. VMP : vote majoritaire pondéré, VM : vote majoritaire, Prob : score probabilisé.*

Dans la suite, nous effectuerons les décompositions en utilisant uniquement les dictionnaires appris sur SOLO1. Elles seront appliqués sur les sons de SOLO2.

6.2.5 Influence des paramètres de décomposition

Nous allons maintenant étudier l'influence de chacun des paramètres pris séparément. Une recherche en grille de l'ensemble des meilleurs paramètres n'aurait pas été possible techniquement étant donnés leur nombre et la durée des décompositions. Nous supposerons donc qu'ils influent de façon décorrélée sur le résultat de classification.

6.2.5.1 Influence de la quantification de la fréquence fondamentale

Le Tableau 6.2.5.1 montre l'influence de la quantification de f_0 sur les résultats de reconnaissance.

Δf_0	1/2 ton	1/6 ton	1/10 ton	1/14 ton
VMP	74	85	84	84
Prob	70	81	82	82
VM	75	82	83	83
VMP	75	81	82	83
Prob	73	80	81	82

TAB. 6.4 – *Influence de la quantification de la fréquence fondamentale sur les résultats de classification. Taille des segments de décision : 2 secondes. Haut du tableau : résultats basés sur les livres entiers. Bas du tableau : résultats basés sur les livres avec un atome par trame temporelle. VMP : vote majoritaire pondéré, VM : vote majoritaire, Prob : score probabilisé.*

Nous voyons que l'échantillonnage de la fréquence fondamentale est important pour la bonne classification des instruments de musique : un échantillonnage à un demi-ton n'est pas suffisant. A partir d'un échantillonnage à $1/6$ de ton, les résultats sont proches d'une valeur asymptotique, quelque soit la méthode de vote choisie.

6.2.5.2 Influence de la quantification du dictionnaire (K)

Le Tableau 6.2.5.2 montre l'influence de la quantification du dictionnaire de vecteurs d'amplitudes A, c'est-à-dire du paramètre K dans l'algorithme des K-moyennes.

K	1	2	4	8	16
VMP	79	81	83	85	84
Prob	79	82	82	83	82
VM	75	78	81	82	83
VMP	74	77	80	82	82
Prob	74	78	80	81	81

TAB. 6.5 – *Influence de la quantification du dictionnaire sur les résultats de classification. Taille des segments de décision : 2 secondes. Haut du tableau : résultats basés sur les livres entiers. Bas du tableau : résultats basés sur les livres avec un atome par trame temporelle. VMP : vote majoritaire pondéré, VM : vote majoritaire, Prob : score probabilisé.*

Comme précédemment, la quantification du dictionnaire en vecteurs d'amplitudes influe sur les résultats de classification. Ici, l'optimum semble être atteint vers $K = 8$ si on utilise le Vote Majoritaire Pondéré. La valeur $K = 16$ donne les meilleurs résultats pour la méthode de vote majoritaire. Cependant, l'utilisation d'un nombre d'atomes plus important nécessite plus de calcul. La valeur $K = 8$ semble donc être un bon compromis.

6.2.5.3 Influence des échelles utilisées

Le Tableau 6.2.5.3 montre l'influence du choix des échelles pour la classification d'instruments. La plus petite échelle utilisée sera 1024 ($46ms$). Prendre une échelle plus petite rendrait les partiels trop corrélés pour les fréquences fondamentales les plus basses. Les

s	1024	2048	$1024, 2048, 4096, 8192$
VMP	84	85	83
Prob	82	81	77
VM	83	82	80
VMP	82	81	80
Prob	81	79	74

TAB. 6.6 – *Influence de la quantification du dictionnaire sur les résultats de classification. Taille des segments de décision : 2 secondes. Haut du tableau : résultats basés sur les livres entiers. Bas du tableau : résultats basés sur les livres avec un atome par trame temporelle. VMP : vote majoritaire pondéré, VM : vote majoritaire, Prob : score probabilisé.*

résultats montrent que l'utilisation de grandes échelles, même combinées à de petites, dégradent les résultats en classification. L'utilisation de plusieurs échelles ne permet pas d'améliorer les résultats par rapport à l'utilisation de la plus grande uniquement. En effet, étant donné que les notes durent généralement plus longtemps que les atomes les plus longs utilisés, ceux-ci sont sélectionnés de façon prioritaire dans les décompositions et capturent de plus la plus grande partie de l'énergie du signal, même si la structure

harmonique dans le signal possède des modulations de fréquence. Les atomes longs ont toutefois un désavantage majeur par rapport à des atomes de taille plus courte (50 ms) : bien qu'ils soient localisés aux bonnes hauteurs, leur enveloppe temporelle ne s'adapte pas bien aux enveloppes temporelles du signal, et ils ne peuvent pas suivre les modulations de fréquence apparaissant dans les vibratos. On pourrait néanmoins compenser ce phénomène en sous-pondérant légèrement les atomes de longue échelle, c'est-à-dire en fixant leur énergie à une valeur légèrement plus petite que 1.

Cependant, une critique peut être faite sur ces expériences : les amplitudes du dictionnaire ont été apprises sur une seule échelle, et sont utilisées pour les atomes de toutes les échelles. Or, rien ne garantit que les amplitudes des partiels sont les mêmes si on observe le spectre des instruments sur des échelles différentes, même si la condition de quasi-orthogonalité est respectée. Les différences peuvent apparaitre essentiellement pour les instruments présentant des modulations de fréquence. Dans ce cas, étant donné que les modulations de fréquences ont une amplitude proportionnelle à l'ordre des partiels, les partiels d'ordre élevé risquent d'avoir leur lobe principal écrasé si on les observe avec des atomes de Gabor de grande échelle. Des expériences supplémentaires mettant en jeu un apprentissage de vecteurs d'amplitudes sur différentes échelles permettraient de conclure plus rigoureusement sur l'utilité ou non de plusieurs échelles pour la reconnaissance des instruments. La classification d'instruments ayant un registre plus grave permettrait également de conclure définitivement sur l'apport de la multi-résolution pour cette tâche.

6.2.5.4 Influence du type d'algorithme

Nous allons maintenant examiner si le type d'algorithme utilisé influe sur la décision. Les algorithmes examinés seront donc le Matching Pursuit (MP), l'algorithme atomique avec réestimation des paramètres (Atom.) et l'algorithme moléculaire avec pénalisation de la longueur des chemins (Moleculaire 1). Le paramètre β pour l'algorithme moléculaire 1 a été réglé manuellement à 0,3. Un réglage rigoureux nécessiterait des tests de classification sur une ensemble de développement. Les résultats sont présentés dans le tableau 6.2.5.4.

Algorithme	MP	Atom.	Moléculaire 1
VMP	84	85	87
Prob	82	81	82
VM	83	83	83
VMP	82	82	83
Prob	81	81	81

TAB. 6.7 – *Influence de l'algorithme sur les résulats de classification. MP : Matching Pursuit, Atom : algorithme atomique avec réestimation des paramètres, . Taille des segments de décision : 2 secondes. Haut du tableau : résultats basés sur les livres entiers. Bas du tableau : résultats basés sur les livres avec un atome par trame temporelle. VMP : vote majoritaire pondéré, VM : vote majoritaire, Prob : score probabilisé.*

L'algorithme moléculaire par pénalisation de la longueur de chemin se démarque des autres concernant la classification. Il permet de faire porter le poids sur des décisions plus

robustes. Les résultats obtenus grâce à cet algorithme sont significativement au-dessus de l'algorithme par sac de trames.

6.2.5.5 Bilan sur l'influence des paramètres sur la reconnaissance des instruments

Les expériences que nous avons effectuées permettent de tirer les conclusions suivantes :

- La quantification fine de la fréquence fondamentale est indispensable à une bonne classification,
- L'utilisation d'échelles plus grandes n'a pas d'intérêt pour la classification avec l'ensemble d'instruments étudié,
- La quantification du dictionnaire d'amplitudes est elle-aussi importante, 8 atomes par classe $\mathcal{C}_{ip}$ semble la valeur optimale en terme de rapport score de classification/coût de calcul,
- Dans pratiquement tous les cas de figure, le Vote Majoritaire Pondéré sur les décompositions complètes apparait comme la méthode de décision la plus robuste.
- L'algorithme moléculaire donne les résultats les plus élevés, supérieurs à ceux donnés par l'algorithme de classification par sac de trames.

La matrice de confusion obtenue grâce à l'algorithme moléculaire est affichée dans le tableau 6.2.5.5. On remarque que les confusions sont assez équilibrées entre les classes, à l'exception de la confusion du Violoncelle en Violon et de la Clarinette en Flûte.

	Cl	Co	Fl	Ob	Vl
Cl	84.8	1.8	8.1	2.2	3.1
Co	1.6	89.8	0.8	0.0	7.8
Fl	3.4	2.2	87.5	5.2	1.7
Ob	3.9	1.0	1.3	88.9	4.9
Vl	1.9	3.8	6.2	5.6	82.5

TAB. 6.8 – *Matrice de confusion pour l'algorithme moléculaire.*

6.2.6 Perspectives

D'autres paramètres pourraient être calculés avant la soustraction des atomes, comme la proportion d'énergie enlevée dans la zone temps fréquence de l'atome. Cela reviendrait à extraire un rapport "partie harmonique à bruit" local, caractéristique qui a fait ses preuves pour la reconnaissance des instruments de musique avec des méthodes par sacs de trames.

Il est enfin nécessaire d'évaluer l'algorithme sur des plus grandes bases d'instruments, afin de valider ou non sa validité sur des cas réalistes. On peut néanmoins s'attendre à une baisse globale de performance dans des contextes plus ouverts (par exemple 40 instruments), comme pour tous les algorithmes de classification. Une classification hiérarchique devrait cependant fonctionner, en cherchant à identifier les familles d'instruments mis en jeu plutôt que l'instrument exact. Pour de tels contextes, on pourrait également mettre en oeuvre des pondérations des atomes en fonction de leur probabilité *a priori* d'être joués. Par exemple, les parties extrêmes des registres des instruments de musique sont beaucoup moins exploitées que le milieu. On pourrait réaliser cette opération grâce à un Matching Pursuit Pondéré (Escoda et al., 2006).

6.3 Reconnaissance d'ensembles

Nous avons vu que la décomposition pouvait être effectuée indépendamment du nombre d'instruments présents dans le mélange. On peut alors mettre en oeuvre des post-traitements permettant d'expliquer le mélange d'instruments, en se basant sur le fait que la reconnaissance d'instruments sur des performances solo fonctionne de manière correcte et comparable à l'état de l'art. Dans les expériences qui suivront nous supposerons que chacun des instruments étudiés émet au maximum une note à la fois. Le traitement d'instruments polyphoniques nécessite une réflexion plus approfondie qui n'a pas été menée.

6.3.1 Expérience préliminaire : Reconnaissance de duos

Une première série d'expériences a été réalisée sur des duos réels (ensemble DUOS), et a été présentée dans (Leveau et al., 2008). Dans cette étude, le nombre maximal d'instrument contenu dans le signal est donc connu et égal à deux. L'ensemble d'instruments parmi lesquels les duos seront choisis est le même que dans la section précédente (Cl, Co, Fl, Ob et Vl). Les paramètres utilisés sont les paramètres par défaut de l'algorithme présenté précédemment.

L'algorithme permettant de prendre une décision est le suivant. Un échantillonnage temporel est effectué au plus grand commun diviseur des Δu du dictionnaire. Chaque instant ainsi obtenu est associé à un ou deux instruments en sélectionnant les deux atomes ayant les énergies instantanées $(|\alpha_\lambda| w(\frac{u-u_\lambda}{s_\lambda}))^2$ (définies en 4.9) les plus élevées, ou un atome seulement si un seul instrument est présent à cet instant. Ensuite, le label du duo du segment entier est décidé par vote majoritaire pondéré par la somme de la valeur absolue des poids sur chaque trame. Cette méthode ne fait pas intervenir de méthodes de suivi de lignes mélodiques contrairement à d'autres approches (Klapuri & Davy, 2006), mais est relativement simple à mettre en oeuvre.

Nous calculons ensuite plusieurs scores permettant d'évaluer notre méthode :

- Le score A mesure la précision de la reconnaissance du duo, ou du solo si seulement un instrument est détecté. Par exemple, si le label à détecter est Co&Fl, les détections correctes sont Co, Fl et Co&Fl. La tolérance des soli dans ce score permet d'éviter l'annotation manuelle de toute la base de données, qui consisterait à indiquer à chaque instant si un seul instrument joue ou si ce sont les deux instruments qui sont actifs à la fois.
- Le score B compte une bonne détection lorsque tous les instruments détectés appartiennent au duo annoté. Dans notre exemple, Les labels Co&Co et Fl&Fl sont alors aussi acceptés.
- Le score C compte une bonne détection lorsque au moins un instrument du duo a été trouvé. Dans notre exemple, les labels Co&Vl, Co&Ob, Co&Cl, Cl&Fl, Ob&Fl et Fl&Vl sont ajoutés.

Les scores développés surévaluent les performances de l'algorithme : il prend en compte le fait que les deux instruments ne jouent pas nécessairement simultanément sur toutes les trames du signal, même si les extraits ont été sélectionnés de telle sorte que cette situation soit majoritaire.

Les scores obtenus en utilisant un tirage aléatoire seraient égaux à 5%, 25% et 55% pour un duo avec deux instruments différents, et 10%, 10% et 30% respectivement pour les duos de deux instruments identiques, en considérant tous les labels équiprobables.

Les scores obtenus pour la décomposition atomique et la décomposition moléculaire sont présentés respectivement dans les Tableaux 6.9 et 6.10. Les paramètres utilisés pour ces expériences sont les paramètres par défaut de la décomposition. Les algorithmes testés sont l'algorithme atomique avec réestimation des paramètres et l'algorithme moléculaire par délimitation de la zone de recherche (avec $\mu_0 = 0,03$ et $\mu_e = 0,2$). Le critère d'arrêt des décompositions est un SRR de 15 dB ou un nombre maximal de 250 atomes par seconde.

%	Nombre d'extraits	A	B	C
Cl&Fl	200	55	78	97
Co&Fl	170	40	81	98
Fl&Fl	29	48	48	93
Co&Vl	414	23	70	96
Total	813	35	74	97

TAB. 6.9 – *Résultats de reconnaissance des instruments de musique sur des duos en utilisant la décomposition atomique (A : vrai duo ou solo, B : bons instruments présents, C : au moins un instrument trouvé).*

%	Nombre d'extraits	A	B	C
Cl&Fl	200	58	87	98
Co&Fl	170	55	79	100
Fl&Fl	29	69	69	90
Co&Vl	414	24	59	89
Total	813	41	71	94

TAB. 6.10 – *Résultats de reconnaissance des instruments de musique sur des duos en utilisant la décomposition moléculaire (A : vrai duo ou solo, B : bons instruments présents, C : au moins un instrument trouvé).*

L'algorithme moléculaire possède ici des performances nettement supérieures à l'algorithme atomique concernant le score A, et du même ordre pour les scores B et C. Une interprétation possible est que la prise de décision plus lissée qu'ils effectuent permet de mieux discriminer les notes que les décisions faites sur les atomes lorsque les sources se perturbent entre elles. Ces résultats nécessitent cependant une validation plus robuste, car un biais peut être introduit par l'utilisation d'un ensemble de test restreint à 4 types de duos (Cl&Fl, Co&Fl, Fl&Fl et Co&Vl) au lieu des 15 possibles.

6.3.2 Reconnaissance d'ensembles

L'expérience présentée dans le paragraphe précédent ne permet pas de déterminer le nombre d'instruments présents dans le mélange à un instant donné à partir d'une décomposition profonde, et ne fonctionne que parce que le seuil d'arrêt de la décomposition a été choisi soigneusement. Nous proposons donc dans cette partie de travailler sur des décompositions profondes, avec beaucoup plus d'atomes que nécessaire. De plus, nous travaillerons sur l'ensemble de test ENS1, qui possède une polyphonie plus contrôlée que l'ensemble DUO : tous les instruments jouent en même temps, et sans plage de silence.

Dans cette expérience, il s'agit donc de déterminer le nombre de sources sur des segments en utilisant un critère de parcimonie en hauteur *a posteriori*. Le critère défini en 4.9 peut être utilisé pour sélectionner un certain nombre d'atomes par trame, chacun représentant une source : dans ce cas, la méthode de vote présentée dans la section 6.3.1 peut être mise en oeuvre. On peut également garder une incertitude sur les décisions des atomes, et faire un fusion tardive en utilisant des saillances d'ensemble. C'est la méthode que présenterons dans la suite, et qui a déjà été introduite dans (Leveau et al., 2007).

6.3.2.1 Saillances d'ensemble

Afin de déterminer le score d'un ensemble pour une trame donnée, on peut tout d'abord penser à prendre les labels des atomes extraits à un instant donné après un post-traitement de parcimonie en hauteur, puis effectuer un vote entre les trames afin d'obtenir un score global pour le segment étudié et ainsi prendre une décision, de façon similaire à l'expérience sur les duos.

On peut également garder un degré de liberté au niveau des décisions par trame en définissant des saillances d'ensemble, de façon similaire à 6.2.2. Les saillances d'instruments permettent d'obtenir des *saillances d'ensemble* pour des instants donnés : le score d'une classe d'ensemble dépend des saillances de hauteur et d'instruments portés par les atomes extraits dont le support temporel contient ces instants.

Les instants de calcul de saillance sont le résultat de l'échantillonnage du temps avec un pas correspondant au plus grand commun diviseur entre les Δu correspondant à chaque échelle. Ainsi, la contribution de chaque atome h à un instant donné est égal à la valeur à l'instant u de la fenêtre de pondération w de l'atome commençant à u_a multipliée par le poids de l'atome. Etant donné un instant u et un label d'ensemble e, sa saillance d'ensemble est la suivante[1] :

$$\mathcal{S}_e(u) = \frac{\max_{C_e \in \mathcal{C}_e} \sum_{a \in C_e} \mathcal{S}_{ia}(u)}{N_e^\beta} \tag{6.6}$$

où $\mathcal{C}_e$ est l'ensemble de toutes les combinaisons de saillance d'instrument dont le support temporel contient u. Par exemple, si deux atomes sont présents à l'instant u, la saillance de l'ensemble Co&Fl (Violoncelle et Flûte) est le maximum entre la somme de la saillance de Fl du premier atome et de la saillance de Co du second, et de la somme de la saillance de Co du premier atome et de celle de Fl du second, divisé par 2^β. Un exemple de livre et de la représentation en saillance correspondante est affichée sur la Figure 6.10.

Le paramètre β est un paramètre de parcimonie en pitch *a posteriori*, similaire à celui présenté en 4.9, et permet donc de pondérer le nombre de sources activées. Sa valeur doit être ajustée sur un ensemble de développement.

6.3.2.2 Vote

Comme dans le cas monophonique, une stratégie de vote doit être mise en place. On peut utiliser les même stratégies de pondération des votes que celles mentionnées en 6.2.2 : vote majoritaire, vote majoritaire pondéré, ou score probabilisé. Des expériences

[1] L'utilisation de la norme $L2$ $\sqrt{\sum_{a \in C_e} (\mathcal{S}_{ia}(u) w(\frac{u-u_a}{s_a}))^2}$ à la place de la norme $L1$ $\sum_{a \in C_e} \mathcal{S}_{ia}(u) w(\frac{u-u_a}{s_a})$ serait plus cohérente avec le critère d'optimalité de la décomposition, mais mène à des résultats plus faibles dans l'application étudiée

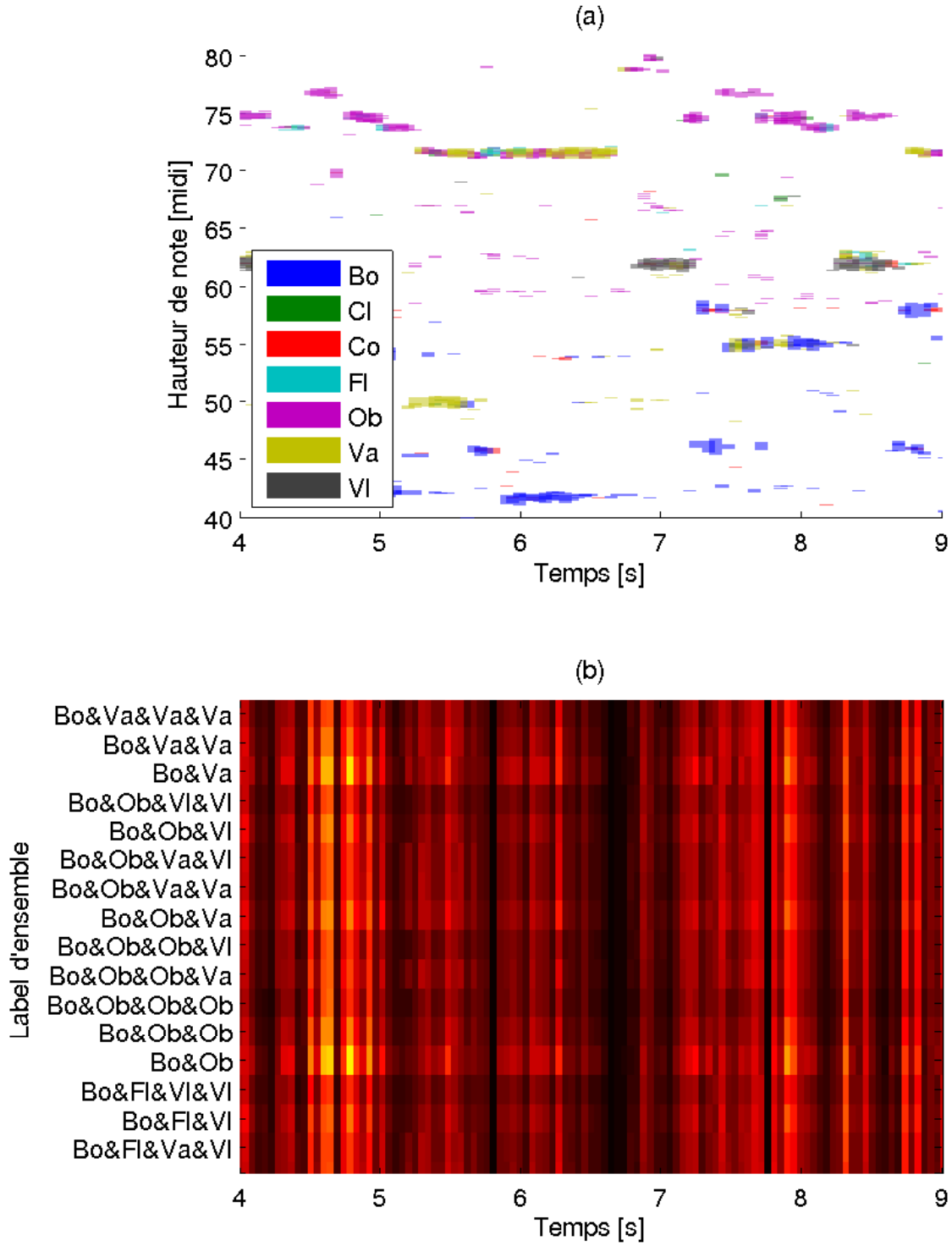


FIG. 6.10 – *duo de Basson (Bo) et de Hautbois (Ob) (mélange synthétique) : (a) Représentation des livres dans le plan temps-pitch : les atomes sont représentés par des rectangles, pour chacun desquels la largeur est l'échelle de l'atome et la hauteur son amplitude, (b) Saillances d'ensemble pour un sous-ensemble des labels d'ensembles (les hautes saillances sont plus claires).*

préliminaires ont montré que le score probabilisé donnait de meilleurs résultats, nous ne nous intéresserons donc qu'à cette méthode.

6.3.2.3 Expériences

L'ensemble étudié est cette fois-ci composé du Basson (Bo), du Violoncelle (Co), de la Clarinette (Cl), du Hautbois (Ob), du Violon alto (Va) et du Violon (Vl). Les paramètres utilisés pour les décompositions sont $s = 93ms$, $\Delta = 46ms$. L'échelle des atomes est choisie plus longue que dans les expériences sur les soli car le Basson, mis en jeu dans l'expérience, possède des notes plus graves que le violoncelle. f_0 est échantillonnée logarithmiquement avec un pas d'$1/10$ ton. Les décompositions sont effectuées jusqu'à ce que le RSR atteigne 20 dB.

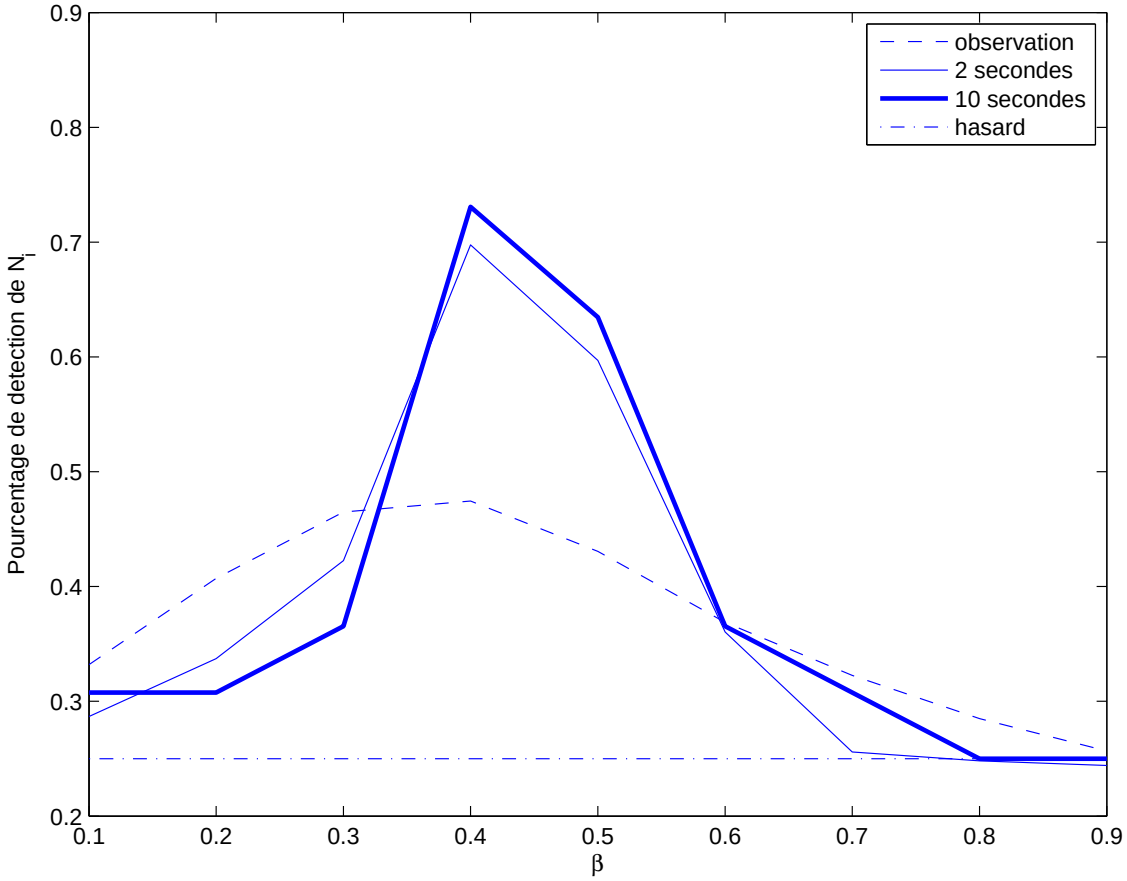


FIG. 6.11 – *Précision de l'estimation du nombre d'instruments en fonction de β pour des décisions par trame, par segment de deux secondes et segments de 10 secondes.*

L'ensemble de développement est ENS1. Les paramètres β et γ ont été réglés de façon à maximiser la précision de l'estimation du nombre d'instruments qui est requise pour estimer le bon ensemble. L'optimisation de ces paramètres pour la précision des labels d'instruments impliquerait un surapprentissage sur la reconnaissance des soli car il est plus facile d'identifier le bon label dans ce cas (moins de possibilités). Dans nos expériences, les coefficients γ optimaux sont indépendants de la fenêtre de décision : la

valeur $\gamma = 0.8$ donne les meilleurs résultats. β est également indépendant de la longueur de décision, et sera donc choisi à $\beta = 0.4$.

Pour ces valeurs, les taux de reconnaissance d'instruments sur des segments de 10 secondes sont affichés sur la Figure 6.12. Contrairement aux expériences présentées en 6.3.1, aucune tolérance n'est considérée quant au nombre d'instruments annoté, censé être le nombre exact d'instruments mis en jeu. Il montre que le problème de trouver un instrument dans un mélange est correctement traité quand le nombre d'instruments est connu (de 70 % à 100% selon le type d'ensemble), mais moins quand le nombre d'instruments de l'ensemble est inconnu (de 54 % à 84 %). Cependant, lorsque le nombre d'instruments à reconnaître augmente, il est plus difficile de les reconnaître précisément. Le cas des ensembles de trois instruments demande des techniques plus sophistiquées à la fois au niveau de la décomposition et du post-traitement, car le problème devient très difficile (le résultat d'un tirage aléatoire est de moins de 1 %).

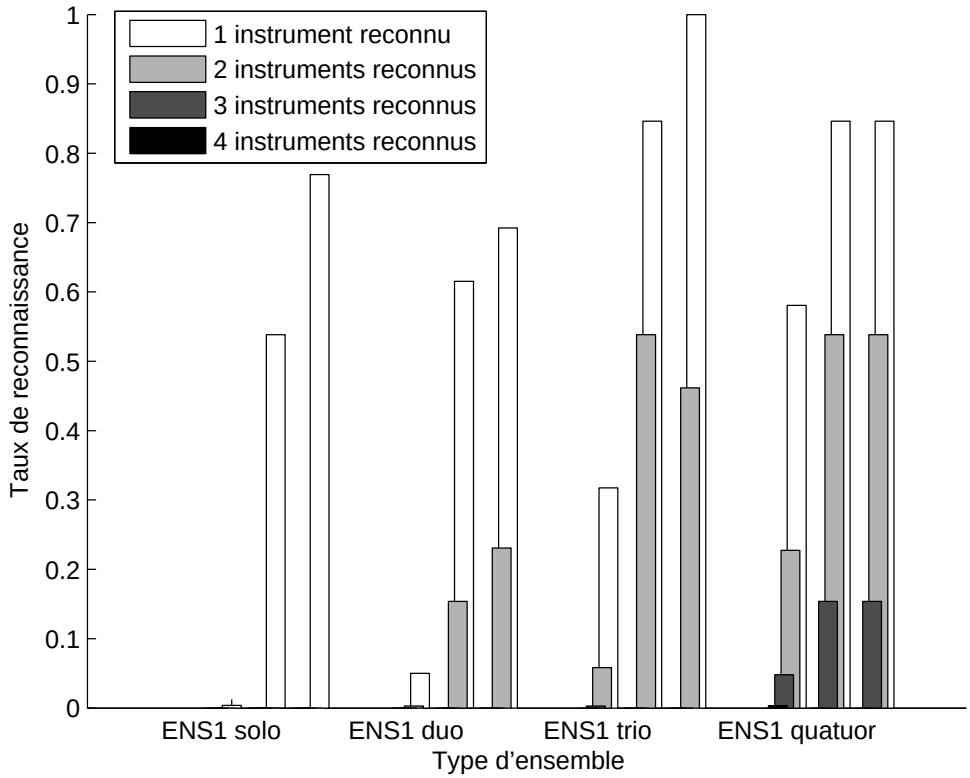


FIG. 6.12 – *Reconnaissance d'ensemble pour chaque sous-ensemble de ENS1 (soli, duos, trios, quatuors). Pour chaque ensemble, les trois groupes de barres représentent respectivement les résultats d'un tirage aléatoire, de l'algorithme sans information a priori sur le nombre d'instruments qui jouent, et de l'algorithme avec une connaissance du nombre d'instruments qui jouent.*

Des tests sur des mélanges réels ont été également été effectués (ensemble ENS2), mais dans ce cas les résultats sont très faibles et parfois à peine meilleurs que le hasard : bien que certaines notes soient souvent bien identifiées en hauteur et instrument, le cardinal de l'ensemble d'instruments est souvent mal estimé. En effet, le traitement par trames proposé ne convient pas à ces signaux, dans lesquels il est fréquent que les instruments

ne jouent pas tous ensemble. De plus les instruments jouent la plupart du temps en harmonie, ce qui perturbe l'extraction des sources, d'autant plus lorsque des algorithmes de décomposition itératifs comme les notres sont utilisés. Par conséquent, de nombreuses améliorations sont nécessaires afin de mieux identifier les sources mises en jeu dans des mélanges complexes et réalistes. Si un algorithme de décomposition permettant d'estimer conjointement les sources activées semble être indispensable, le problème de l'identification des ensembles doit également être mieux posé, et les évaluations plus élaborées. Un autre point manquant pour aborder ce problème est qu'on ne dispose pas d'étude montrant quelles sont les performances humaines dans l'identification d'ensemble.

6.4 Localisation spatiale et reconnaissance des sources : le cas stéréophonique

6.4.1 Position du problème

Nous allons maintenant dresser des perspectives quant à l'application des décompositions atomiques dans le cas stéréo. Le travail a été présenté dans (Sodoyer et al., 2007), et a été effectué en collaboration avec David Sodoyer, post-doctorant à l'Institut Jean Le Rond d'Alembert.

L'ajout d'une information de panoramique dans la paramétrisation des atomes permet de donner des indications sur la localisation des sources sonores sur l'axe de la panoramique. Cette information pourrait être exploitée dans un post-traitement, en associant des atomes extraits possédant une proximité dans l'espace paramétré par leur timbre, leur localisation temporelle, leur hauteur de notes et leur angle.

Dans le cas anéchoïque, le signal stéréo d'un instrument $x_m(t)$ peut être modélisé comme une paire de signaux :

$$x_{st}(t) = [\cos(\theta)x_m(t) \quad \sin(\theta)x_m(t-\tau)]^T \tag{6.7}$$

où θ est le paramètre de panoramique de la source et τ un paramètre de délai entre les canaux. Dans cette modélisation, nous ne tenons pas compte de la directivité des micros et de la source qui modifient le spectre du signal orginal. Dans le cas d'un enregistrement acoustique avec un effet de salle, le signal stéréo est le résultat de la convolution des signaux sources avec des réponses impulsionnelles spécifiques aux couples sources-mircrophones :

$$x_{st}(t) = [\psi_l(t) * x_m(t) \quad \psi_r(t) * x_m(t-\tau)]^T \tag{6.8}$$

Si l'on utilise des considérations d'acoustique des salles, ces réponses impulsionnelles peuvent être considérées comme le résultat de la contribution d'un grand nombre de sources virtuelles I, avec pour chacune un paramètre de panoramique θ_i, et qui émettent avec un délai t_i par rapport à l'émission originale. Nous ne tiendrons pas compte ici des éventuels filtrages effectués par les parois en supposant qu'elles n'effectuent qu'une absorption, introduisant un facteur multiplicatif a_i sur l'amplitude de la source.

Ainsi, on peut écrire la réponse impulsionnelle du filtre stéréo $\psi_{st}(t)$:

$$\psi_{st}(t) = \left[\sum_{i=1}^{I} a_i \cos(\theta_i)\delta(t-t_i) \quad \sum_{i=1}^{I} a_i \sin(\theta_i)\delta(t-t_i-\tau_i)\right]^T \tag{6.9}$$

Si l'on identifie toutes les sources, virtuelles ou non, les sources les plus énergétiques sont de bonnes candidates pour être de "vraies" sources (c'est-à-dire les sources directes), les moins énergétiques étant plus vraisemblablement issues de réflexions.

Dans l'application envisagée, on considère que certaines sources, réelles ou virtuelles, ont été résolues, c'est-à-dire identifiées comme des sources uniques, et que d'autres ont été fusionnées, formant alors des sources plus étalées le long de l'axe de la panoramique, et dont le paramètre de panoramique est intermédaire entre ceux des sources fusionnées. Cet aspect est illustré sur la figure 6.13(c).

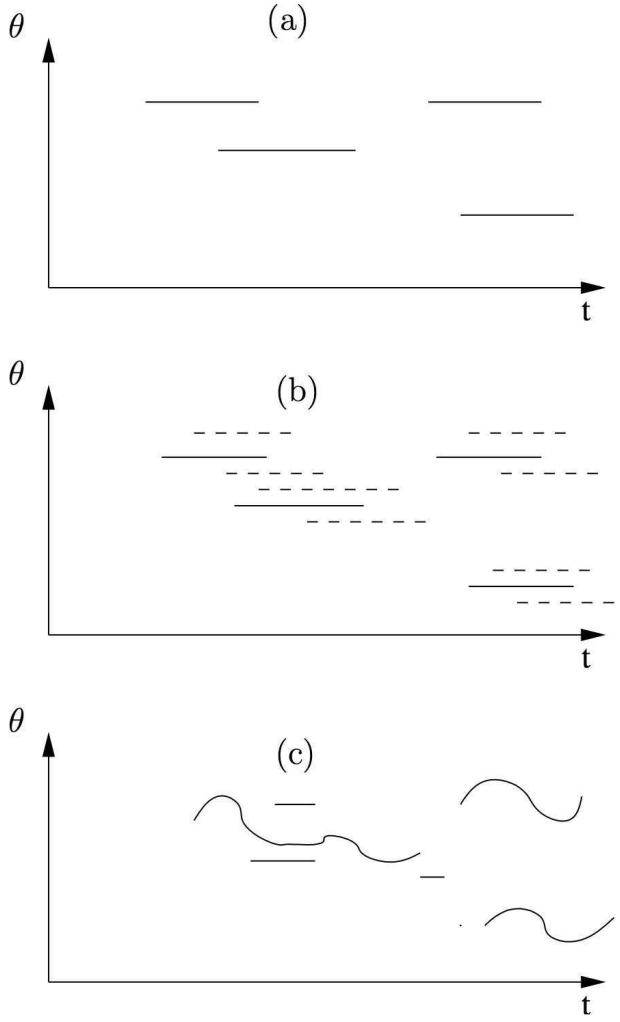


FIG. 6.13 – *Représentation de l'activité de trois sources réelles dans le plan temps-panoramique. (a) Sources dans le cas anéchoïque. (b) Sources réelles et virtuelles dans le cas convolutif lorsque toutes les sources sont séparées. (c) Sources observées avec un paramètre de panoramique global et variable lorsque certaines sources sont fusionnées.*

6.4.2 Résultats préliminaires

L'algorithme est testé sur un mélange stéréo synthétique et un mélange stéréo réel. Le mélange synthétique est un mélange linéaire et instantané composé à partir de perfor-

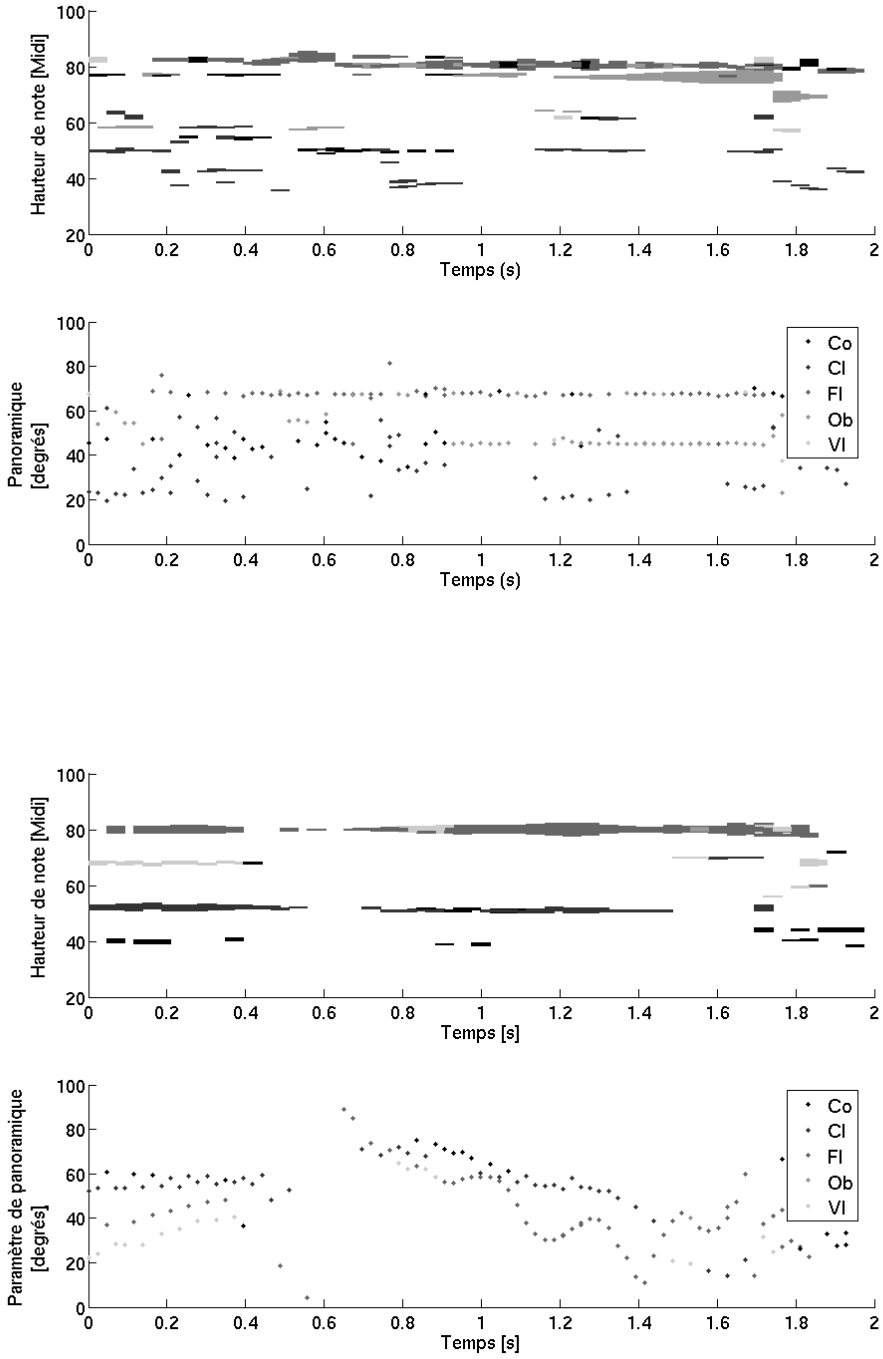


FIG. 6.14 – *Paramètres des atomes pour le mélange instantané (premier groupe de figures) et le mélange réel (deuxième groupe de figures). Pour chaque groupe :* ***Haut*** *: Représentation des atomes dans le plan temps-hauteur. Chaque atome est représenté par un rectangle, dont la largeur et la hauteur est son amplitude. Chaque couleur correspond à l'instrument d'où provient le vecteur d'amplitudes.* ***Bas :*** *paramètre de panoramique en fonction du temps.*

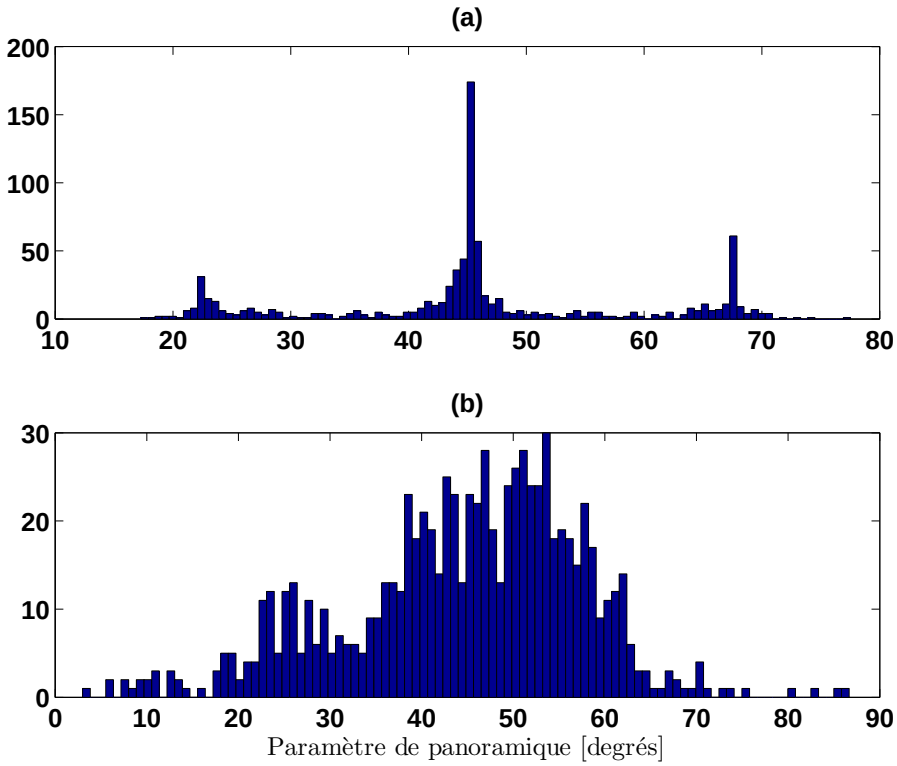


FIG. 6.15 – *Histogramme des paramètres de panoramique, pour des échantillons de 10 secondes. (a) Mélange instantané. (b) Mélange réel.*

mances solo de trois instruments (Vl, Co et Fl) avec des paramètres de panoramique de $\theta_{Vl} = 22°$, $\theta_{Co} = 45°$ et $\theta_{Ob} = 67°$. L'enregistrement réel stéréo est un trio issu d'un CD commercial, composé des mêmes instruments.

Les paramètres utilisés pour cette décomposition sont $s = 46ms$, $\Delta = 23ms$. f_0 est échantillonné à $1/10$ de ton. Les décompositions sont effectuées jusqu'à ce que le rapport signal à résiduel atteind 20 dB.

Les décompositions obtenues contiennent des atomes qui proviennent de différentes classes $\mathcal{C}_{ip}$. Bien que l'instrument i ne soit pas toujours correctement identifié, la hauteur p est correcte dans une large majorité des cas.

La Figure 6.15(a) montre l'histogramme des paramètres de panoramique des atomes. Les sources aparaissent comme clairement séparées. En effet, trois pics peuvent être observés : ils sont centrés sur les angles mentionnés précédemment, ce qui montre que la décomposition est cohérente avec le mélange réalisé. Ainsi, la décomposition mène une représentation utile pour localiser, identifier et compter les sources dans ce cas de figure, un post-traitement simple permettant de réaliser ces tâches.

Dans le cas d'un mélange réel et non-instantané, aucun pic clair n'apparaît sur l'histogramme des paramètres de panoramique présenté sur la Figure 6.15(b). Cependant, comme le montre la Figure 6.14, les sources semblent être représentées par des atomes dont le paramètre de panoramique $\theta(n)$ évolue au cours du temps sur pratiquement tout son axe pour une source donnée. Dans ce cas, la détection du nombre d'instruments ne peut pas être realisée aussi simplement que précédemment. Cependant, ces atomes ne sont pas distribués au hasard selon cet axe : ils semblent suivre une trajectoire nette dans le plan temps-panoramique. Une technique de post-traitement consisterait à suivre ces trajectoires dans l'espace temps-hauteur-panoramique, et à allouer chacune à des sources. Les positions des sources pourraient être trouvées en prenant le paramètre de panoramique du début de ces trajectoires, les premiers atomes des trajectoires ne sont en effet pas perturbés par les réflexions de la salle et ne proviennent que de la source directe.

6.5 Transcription

Le problème de la transcription automatique de la musique est souvent considéré comme le Graal de l'indexation automatique de la musique. Si ce problème était résolu, une conversion "WAV to MIDI" permettrait d'utiliser toute une série algorithmes déjà traitant les données symboliques de type MIDI (similarité mélodique, reconnaissance de genre, extraction de grilles d'accord, de mélodie...).

Jusqu'à présent, ce problème est encore loin d'être résolu dans un contexte ouvert, c'est-à-dire avec peu de connaissances *a priori* sur le signal à analyser, et dans la musique polyphonique mettant en jeu des instruments réels et variés. Klapuri & Davy (2006) fait l'inventaire des méthodes qui ont été proposées dans ce but.

Les algorithmes que nous proposons ne sont pas optimisés pour réaliser cette tâche. Nous allons néanmoins montrer que l'algorithme développé permet d'extraire avec précision la hauteur d'une note. L'extraction des débuts et des fins de notes (*onsets*, *offsets*) n'a cependant pas été évaluée.

6.5.1 Evaluation sur des notes isolées

Afin de valider notre algorithme pour la transcription automatique, les décompositions sont effectuées sur une base de données de notes isolées (RWC). Les dictionnaires

utilisés sont construits à partir d'une base de données de notes isolées distinctes (IOWA + SOL), et contenant des amplitudes de tous les instruments pouvant être mis en jeu dans le signal. Le sous-ensemble d'instruments testé est composé de Cl, Co, Fl, Ob et Vl. La base de test est de 1920 notes. Il faut également signaler que le dictionnaire est le même pour tous les sons, et contient tous les cinq instruments précédemment cités.

Les décompositions sont effectuées jusqu'à un RSR de 5dB, sans pré-traitement du signal à analyser. Les paramètres de la décomposition sont les paramètres par défaut mentionnés en 6.2.3. Une fois la décomposition effectuée sur un signal, la hauteur de la note est inférée en prenant la valeur médiane des fréquences fondamentales des atomes extraits. Une détection est considérée comme correcte si la différence entre la hauteur détectée et la vraie valeur est inférieure à un quart de ton.

Nous utiliserons les paramètres par défaut des expériences de reconnaissance des instruments, aux exceptions près de la taille des atomes et de l'échantillonnage temporel ($s = 92ms, \Delta u = 46ms$). Les expériences ont été effectuées avec différentes configurations :

1. Paramètres par défaut,
2. Paramètres par défaut avec algorithme moléculaire,
3. Paramètres par défaut avec pondération dBA.

Les algorithmes sont comparés avec l'algorithme yin (de Cheveigné & Kawahara (2002)), considéré comme l'un des plus performants pour l'estimation de pitch en conditions monophoniques.

Les résultats des expériences 1, 2, 3 et de yin sont respectivement de 98,1 %, 97,4 %, 97,5 % et 99,0 %. Bien que moins performant que l'algorithme YIN, les algorithmes permettent d'avoir une précision correcte. Comme pour la classification, la pondération psycho-acoustique n'apporte rien à l'estimation de hauteur pour l'ensemble considéré.

Ainsi, ces résultats valident *a posteriori* le réapprentissage que nous avons effectué sur les soli (5.2.3) : le taux d'erreur est faible, de plus les erreurs sont majoritairement effectuées au niveau des limites du registre des instruments, qui sont plus rarement utilisées dans des conditions de jeu réelles que les notes en milieu de registre.

D'autres évaluations seraient nécessaires pour valider l'estimation de hauteur, comme par exemple l'évaluation de l'estimation de fréquences fondamentales multiples.

6.5.2 Evaluation sur une tâche de transcription de piano

6.5.2.1 Post-traitement pour la conversion livre/MIDI

Les algorithmes de décomposition peuvent être utilisés afin de réaliser une transcription automatique des signaux analysés. Pour réaliser cette tâche, il s'agit de convertir la décomposition objet du son en un fichier MIDI. La décomposition objet obtenue à partir du signal exhibe déjà des propriétés intéressantes :

- Les atomes possèdent un poids, qui peut être mis en rapport avec la vélocité MIDI,
- Les atomes sont liés à une hauteur déjà codée en MIDI,
- Les atomes ont une localisation temporelle u et une durée s, qu'on peut lier aux paramètres MIDI correspondants.

Dans le cas de la décomposition atomique, les échelles s des atomes sont échantillonnées en un faible nombre de valeurs, la plupart du temps courtes par rapport aux durées des notes mises en jeu. Il est donc nécessaire de grouper les atomes entre eux afin de

former les notes. Dans le cas du piano, le critère retenu pour grouper des atomes deux à deux sera simplement qu'ils doivent correspondre à la même hauteur de note et avoir des supports temporels voisins : considérant deux atomes h_1 et h_2 tels que $u_1 < u_2$, il suffit que $p_1 = p_2$ et $u_1 + s_1 > u_2$. Dans le cas de la décomposition moléculaire, le même traitement peut être effectué, en considérant le livre obtenu comme celui d'une décomposition atomique.

Comme nous l'avons souligné précédemment, les algorithmes de décomposition proposés ne permettent pas de gérer convenablement la parcimonie en hauteur. Le post-traitement proposé en 4.9 est effectué avant de réaliser ces opérations.

6.5.2.2 Visualisations de transcriptions

La décomposition des sons de piano a été effecuée avec les paramètres suivants : $Fs = 22050Hz$, $(\Delta u, s) \in \{(512, 256), (1028, 2048), (1024, 4096), (1024, 8192)\}$, et un seuil d'arrêt à 50 atomes par seconde. L'utilisation de longues échelles est nécessaire afin que les partiels soient quasi-orthogonaux pour les notes les plus graves du piano.

La Figure 6.16 montre un exemple de transcription. De nombreuses erreurs sont présentes, notamment des notes supplémentaires dans les très basses fréquences fondamentales.

6.5.2.3 Evaluation subjective et objective

Un algorithme complet de transcription de piano (décomposition et conversion livre/ MIDI) a donc été proposé pour une évaluation menée par Adrien Daniel (stagiaire à Télécom Paris). L'évaluation a été effectuée à l'aide de tests perceptifs, où les auditeurs devaient juger de la gêne provoquée par les erreurs de transcriptions de plusieurs algorithmes, et une autre à l'aide de paramètres objectifs (Daniel, 2007).

Les résultats figurent en annexe de ce document. Ceux donnés par les paramètres objectifs sont présentés pour deux pièces de piano : Debussy, Suite bergamasque, III. Clair de Lune (20 premières secondes), et Mozart, Sonate en Ré Majeur, KV 311/ I.Allegro con spirito (13 premières secondes). Il y apparaît que l'algorithme est assez loin d'être le plus performant, mais qu'il arrive à la hauteur d'autres algorithmes sur certains critères présentés dans le travail. L'évaluation subjective montre que l'algorithme est évalué de façon comparable à d'autres sur certains morceaux.

6.5.2.4 Améliorations possibles

L'algorithme de transcription proposé ne prend pas en compte certaines caractéristiques temporelles du piano, comme le transitoire bref et la décroissance caractéristique de l'amplitude des notes.

On pourrait donc définir un dictionnaire plus adapté à cet instrument, en faisant intervenir des sinusoïdes amorties pour capturer les transitoires, éventuellement en rapport quasi-harmonique.

D'un point de vue algorithmique, un algorithme de High Resolution Matching Pursuit permettrait également de mieux modéliser les phénomènes transitoires avec un dictionnaire multi-résolution, en éliminant le phénomène de pré-écho qui a porté préjudice à la bonne localisation temporelle des notes. On peut également penser utiliser un algorithme moléculaire complexe, qui poserait des contraintes sur la succession des ampli-

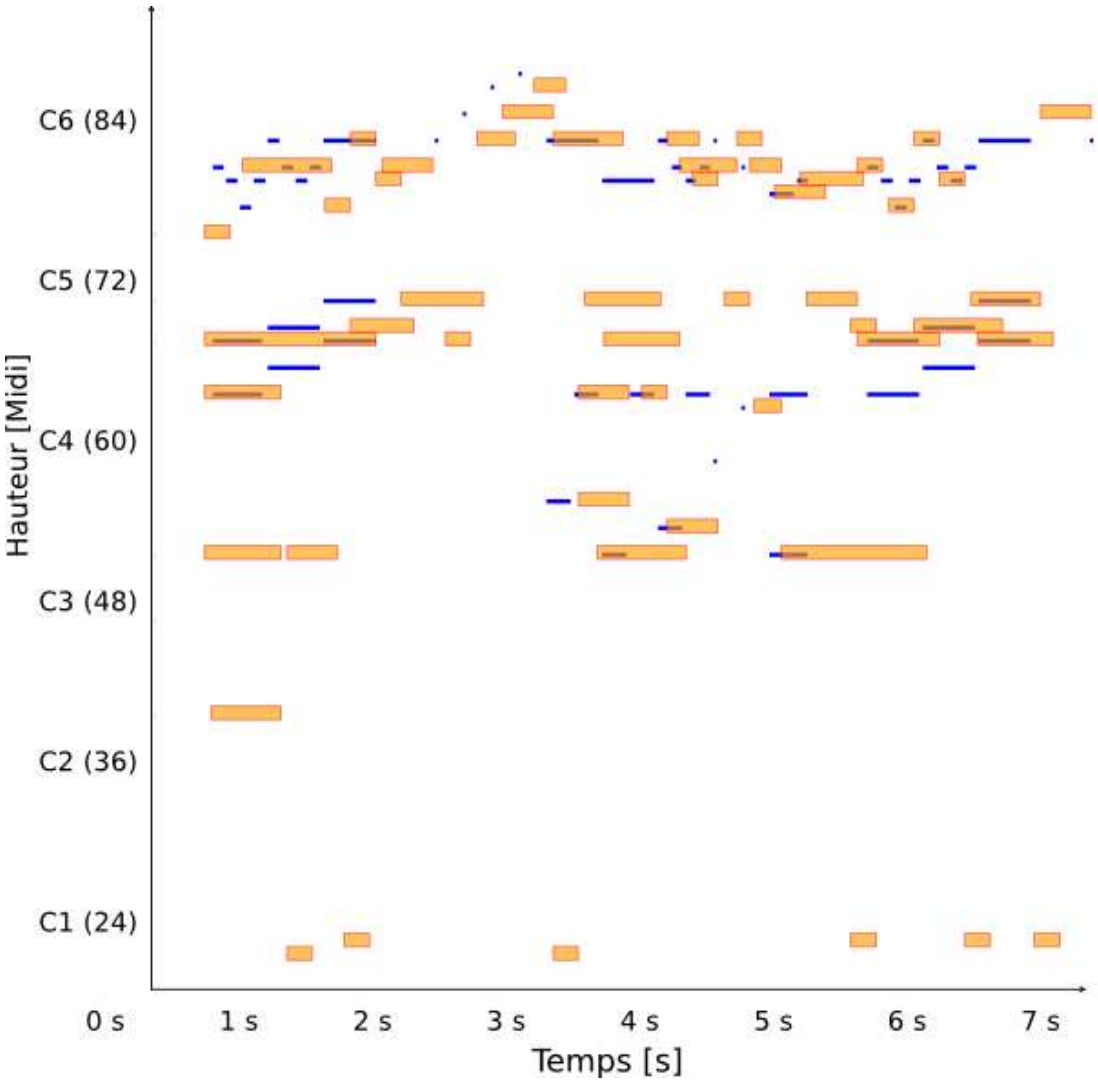


FIG. 6.16 – *Visualisation d'une transcription de piano. Rectangles bleus : partition originale, Rectangles oranges : transcription réalisée.*

tudes des atomes : on disposerait alors d'un dictionnaire d'enveloppes temporelles dont les atomes seraient corrélés avec les projections du signal sur les atomes IHSI.

6.5.3 Bilan sur la transcription automatique

Nous avons vu que les décompositions proposées sont pertinentes comme pré-traitement pour la transcription automatique de musique. Les atomes et les molécules sélectionnées sur des notes isolées ont quasi-exclusivement une fréquence fondamentale correspondant à la note jouée. Concernant une tâche plus compliquée comme la transcription de piano, un meilleur choix du dictionnaire et de l'algorithme de décomposition permettrait sans doute d'améliorer sensiblement les performances. L'utilisation de règles musicologiques simples (Klapuri & Davy, 2006) permettrait sans doute d'améliorer les résultats.

Une autre perspective pour ce travail serait d'évaluer ces algorithmes sur la transcription de pièces mettant en jeu des instruments différents, de type musique de chambre par exemple.

6.6 Codage objet très bas-débit

Pour des signaux mettant en jeu les sources modélisées par les dictionnaires, il est possible de mettre en oeuvre un codage objet efficace, pourvu que la décomposition ait extrait les structures de signal pertinentes. Le travail présenté a été effectué conjointement avec Grégory Cornuz (stagiaire à l'Institut Jean le Rond d'Alembert) et Emmanuel Ravelli (doctorant à l'Institut Jean le Rond d'Alembert), et est publié dans (Cornuz et al., 2007).

Le codage audio peut mettre en jeu différentes techniques, suivant le débit qui est visé. Si le débit visé est assez haut, le codage audio est effectué à l'aide de transformées du signal dont on ne retient que les coefficients les plus pertinents (par exemple le MPEG-4 AAC et le MPEG4-TwinVQ (ISO/IEC 14496-3 :2001, 2001)). Si les débits visés sont plus bas, les codeurs par transformée ne donnent pas des résultats satisfaisants. On utilise donc des codeurs paramétriques, où l'on utilise des modèles dont on estime des paramètres. Par exemple, le codeur MPEG4-SSC (den Brinker et al., 2002), basé sur des modélisations des sinusoïdes, des transitoires et du bruit, fournit de meilleurs résultats que le MPEG4-AAC à 24kits. Cependant, il ne fonctionne pas à des débits plus bas. Un autre codec, le MPEG4-HILN (Purnhagen & Meine, 2000) qui combine les modélisations des structures harmoniques, des sinusoides et du bruit fonctionne à des débits plus bas mais ses performances sont très dépendantes du signal et est comparable en moyenne au MPEG4-AAC à 16kbit/s et au MPEG4-TwinVQ à 6 kbit/s. L'avantage apporté par le codage HILN est qu'il permet des opérations de changement de durée et de hauteur de notes à la synthèse.

L'étape suivante permettant d'atteindre des débits encore plus bas est le passage à des codages *objet*. Dans ce cas, on n'estime pas les paramètres de structures simples commme des sinusoides, mais on considère que le signal est un mélange d'objets sonores comme des notes ou des accords qu'il s'agit d'identifier. Ensuite, les paramètres de ces objets sont quantifiés et encodés. Une approche permettant de réaliser cette tâche a été proposée par Vincent & Plumbley (2007) : les objets possèdant une hauteur sont modélisés par une somme de partiels en rapport harmonique et sont estimés en utilisant une approche statistique. Le codeur en résultant fonctionne mieux que les codeurs par transformée et les codeurs paramétriques sur des soli ou des duos d'instruments harmoniques à 8kbit/s et

2 kbit/s. Cependant, cette approche demande une grande charge de calcul, ce qui la rend inapplicable pour la plupart des applications pratiques.

Nous allons donc mettre en oeuvre une quantification des molécules obtenues avec l'un des algorithmes présentés afin d'évaluer le potentiel de nos représentations pour le codage objet de la musique. Le contexte sera bien sûr restreint aux instruments présents dans le dictionnaire.

6.6.1 Codage des paramètres

Après une étape de décomposition utilisant l'algorithme moléculaire présenté en 4.5.3, un codage des paramètres des atomes ou molécules est effectué. Un décodeur permet ensuite de regénérer le livre puis de synthétiser le signal. Dans la suite, la base sur laquelle les amplitudes sont apprises est la base ISO.

Deux propriétés de la représentation permettent un codage efficace à des taux de compressions très bas. Tout d'abord, l'algorithme moléculaire renvoie des objets composés d'une succession d'atomes. Les paramètres des atomes qui appartiennent à une même molécule sont fort corrélés et donc peuvent être codés efficacement. Ensuite, comme le dictionnaire est déjà échantillonné avant l'étape de codage, certains paramètres peuvent être encodés sans perte par codage entropique.

Voici le codage des paramètres mis en oeuvre :

- L'échelle s_n est constante et donc n'est pas codée
- La localisation temporelle u_n est sur une grille avec un pas de $s_n/2$. Seule la position absolue du premier atome d'une molécule est codée, les positions des atomes suivants sont donc les valeurs consécutives sur la grille. Les seuls paramètres additionnels requis par le décodeur sont le nombre d'atomes qui appartiennent à la molécule.
- La fréquence fondamentale f_{0_n} de chaque atome est codée de façon brute (avant l'optimisation des paramètres 4.6.1), sur 9 bits. Pour les atomes d'une molécule (excepté le premier), les différences entre les valeurs consécutives de la fréquence fondamentale sont calculées, et les valeurs résultantes sont codées par codage entropique.
- Le poids α du premier atome d'une molécule est codé en utilisant un quantifieur uniforme standard et une approche par codage entropique (Gersho & Gray (1991)). Les poids des atomes suivants sont codés en utilisant un codage différentiel et une quantification uniforme.
- Les amplitudes de partiels A_n sont déjà quantifiées vectoriellement (cf section 5.2.4). On transmet donc l'indice du vecteur correspondant dans le dictionnaire. L'indice est composé par : la classe de hauteur p (version brute de la fréquence fondamentale, déjà codée) + la classe d'instruments (codée une fois par molécule) i + l'indice K dans la classe $\mathcal{C}_{ip}$. Cet indice est codé par codage entropique.
- Le taux de modulation du fondamental c_{0_n} n'est pas codé car nous avons trouvé qu'il n'était pas assez significatif perceptivement parlant par rapport au budget à y allouer pour le coder.
- Les phases ne sont pas codées. Une approche alternative est utilisée, où les phases sont interpolées au niveau du décodeur pour assurer la continuité entre les partiels des atomes successifs. Cette opération est utilisée couramment en codage de la parole.

6.6.2 Evaluation

Le codeur est évalué sur la base COD composée de 5 soli et 4 duos (aucun recouvrement avec la base ISO).

Les deux étapes du processus de codage, la décomposition du signal et le codage des paramètres, ont été effectuées avec ces paramètres :

- **Paramètres d'échantillonnage :** pour notre application, le choix d'une seule échelle s correspondant à une durée de 46 ms est suffisant. Cette échelle est assez longue pour avoir une bonne résolution fréquentielle. Concernant le pas de localisation Δu, elle est fixée à la moitié de l'échelle, suffisamment courte pour suivre les variations pertinentes d'amplitude et de fréquence du signal qui correspondent à des caractéristiques pertinentes perceptuellement, comme le tremolo ou le vibrato (entre 4 et 10 Hz). La fréquence fondamentale est échantillonnée à un pas de $1/10$ de ton.
- **Paramètres de décomposition :** Le seuil général pour les décompositions a été fixé à 15 dB ou 250 atomes par seconde. Pour la formation des chemins d'atomes, la différence entre deux fréquences fondamentales consécutives est fixée au pas d'échantillonnage de $f_0 : 1/10$ de ton.
- **Paramètres de quantification :** Le poids du premier atome de la molécule est quantifié sur 6 bits, et les poids des atomes suivants sur 4 bits. L'ordre du quantifieur différentiel DPCM est mis à un. Le codeur entropique utilisé pour tous les paramètres est le codeur arithmétique adaptatif développé par Witten et al. (1987).

Avec ces paramètres, les temps de calcul sont d'environ 10 fois le temps réel sur un ordinateur équipé d'un processeur à 3GHz, largement dominé par l'algorithme de décomposition.

6.6.2.1 Codec complet et codec réduit

Durant l'étape d'analyse, à la fin de la décomposition, l'algorithme moléculaire a tendance à produire des molécules dont les paramètres ne correspondent pas à des notes réellement jouées (molécules d'erreur de modélisation, 4.1.1). Par conséquent, la décomposition doit être stoppée avant l'apparition de telles molécules[2]. Dans le cadre de l'étude menée, nous avons donc préféré tronquer la décomposition "manuellement" à un certain niveau de RSR pour chacun des signaux audio. Une interface graphique Matlab a donc été implémentée (Fig. 6.17) : un opérateur peut écouter le signal synthétisé avant la quantification et la suppression des taux de chirp et des phases en fonction du nombre d'itérations de l'algorithme moléculaire et choisir le nombre selon lui optimal de molécules. De tels optima ont été réglés par un seul opérateur. Pour deux fichiers (Solo de Co, et duo Co&Vl), le critère d'arrêt original de l'algorithme moléculaire donne le meilleur résultat, ce qui indique que la décomposition n'a pas été poussée assez loin. Nous appelons le codeur basé sur cette manipulation "codeur réduit" ; tandis que le codeur qui encode toute la décomposition est appelé le "codec complet".

6.6.2.2 Tests d'écoute

Pour évaluer nos codecs, nous avons effectués plusieurs tests d'écoute en utilisant la méthode standard MUSHRA (ITU (2003)). 15 personnes ont pris part aux tests d'écoute

[2] On pourrait également appliquer un critère de parcimonie en hauteur *a posteriori* comme dans l'application 6.3.2, mais cela n'a pas été testé.

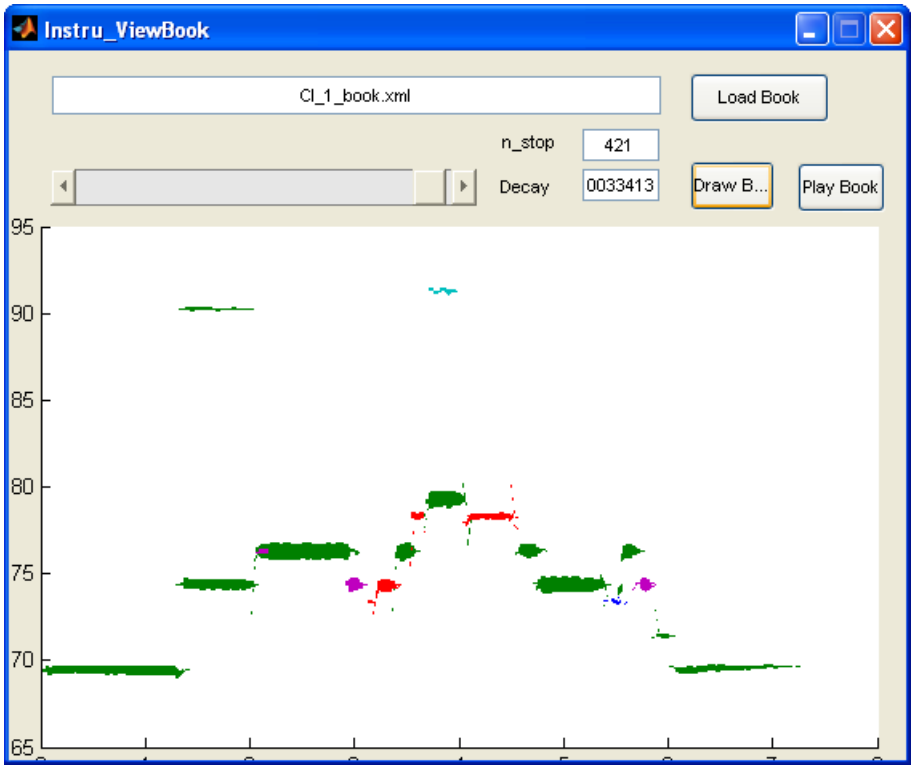


FIG. 6.17 – *Interface graphique Matlab permettant à un opérateur de visualiser et d'écouter la représentation du signal et de sélectionner le seuil optimal pour la décomposition. Des couleurs différentes indiquent des instruments différents.*

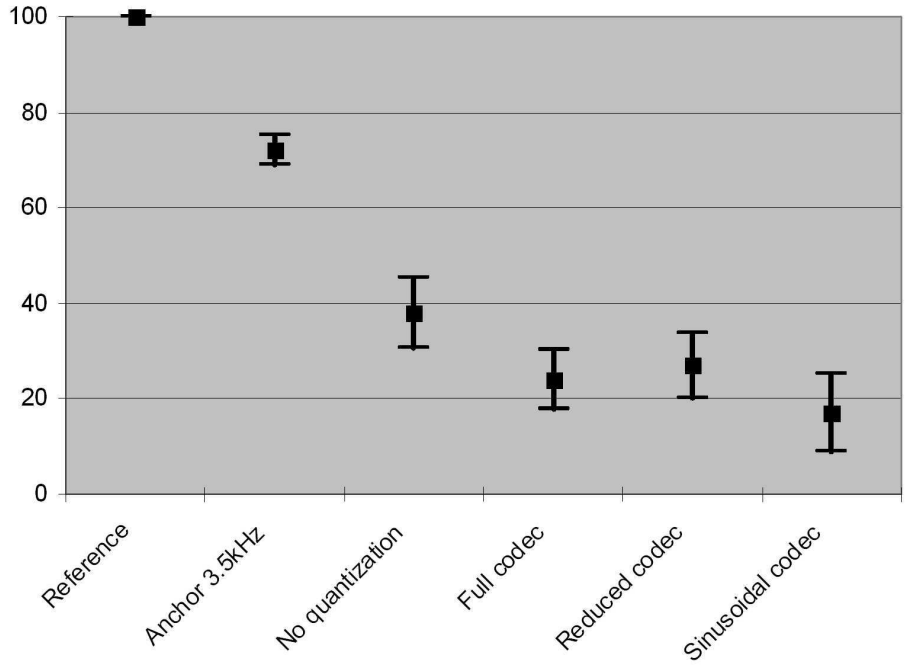


FIG. 6.18 – *Scores MUSHRA moyens pour les 6 types de signaux.*

	Codec complet (FC)	Codec réduit (RC)
Cl	1.3	1.1
Co	3.8	3.8 (*)
Fl	1.3	1.1
Ob	2.6	1.0
Vl	4.4	2.3
Cl & Fl	2.4	1.9
Co & Fl	1.6	1.5
Co & Vl	3.9	3.9 (*)
Fl & Fl	4.6	2.6

TAB. 6.11 – *Débits (en kbit/s par seconde) pour chaque fichier de test et deux variantes du codec. Pour les deux fichiers marqués d'une astérisque (*), le codec réduit est égal au codec complet*

	HR	AN	NQ	FC	RC	SC
Cl	100	69	44	29	21	29
Co	100	81	32	20	20	6
Fl	100	76	29	31	34	30
Ob	100	70	40	12	20	18
Vl	100	66	62	33	33	14
Cl & Fl	100	74	41	36	42	36
Co & Fl	100	74	25	15	21	8
Co & Vl	100	70	43	13	13	6
Fl & Fl	100	68	30	23	21	3

TAB. 6.12 – *Moyenne des scores MUSHRA scores pour chaque version de chaque signal test*

pour comparer 5 versions de chaque signal : une référence cachée (HR), un signal ancre (AN) (à passe-bas 3,5 kHz), le signal synthétisé (NQ) obtenu à la fin de l'algorithme moléculaire (sans quantification), le codec complet (FC), le codec réduit (RC), et un simple codeur sinusoïdal par trame utilisé comme codeur paramétrique de référence (SC, adapté de Jensen & Heusdens (2003)). Le débit moyen des codecs est d'environ 3kbit/s par seconde pour le codec complet (FC) et 2kbit/s pour le codec réduit (RC) (voir Table 6.11). Ces débits sont extrêmement bas, étant donné qu'un codeur paramétrique comme l'AAC nécessite 64 kbit/s pour une qualité correcte. Pour le codeur sinusoïdal (SC), le débit a été fixé à 2 kbit/s par seconde. Les moyennes des scores obtenus pour chaque version de chaque signal sont dans la Table 6.12. Les moyennes générales sont dans la figure Fig. 6.18. Ces résultats montrent tout d'abord que le codeur réduit a des performances similaires ou meilleures que le codec complet, sauf pour deux fichiers (Cl et Fl&Fl) : utiliser plus de bits diminue paradoxalement la qualité. Les résultats montrent également que le codeur réduit donne des résultats similaires ou meilleurs que le codeur sinusoïdal (excepté pour la clarinette). Il faut noter que la qualité du signal décodé est très variable selon les fichiers : certains ont une qualité très faible, d'autres ont une qualité acceptable.

6.6.2.3 Conclusion sur le codage

Le codage objet de la musique polyphonique simple est envisageable en utilisant un dictionnaire d'atomes harmoniques proches des sources à encoder. Le codec développé

montre des performances en général meilleures qu'une approche sinusoïdale. Cependant, des améliorations conséquentes sont à apporter pour que le codec soit réellement exploitable, notamment dans le cas polyphonique. Dans ce cas, une estimation jointe des sources en présence pourrait présenter un grand intérêt, comme pour les tâches précédemment évoquées. L'étape de quantification pourrait aussi être améliorée afin de réduire la perte de qualité qu'elle introduit. Enfin, le codage objet de la musique est très lié à la transcription automatique : les améliorations apportées à cette tâche concernant notamment le comptage et l'identification des sources en hauteur et en instrument rendraient le résultat du codage plus valide perceptivement.

Il serait aussi intéressant de rendre le codage plus générique, en n'imposant pas une position strictement harmonique des partiels, et en introduisant d'autres couches de codage à l'aide de modélisations supplémentaires, comme des représentations des transitoires rapides et du bruit (Verma & Meng (1998)).

6.7 Autres applications potentielles

Les applications décrites dans cette section sont des pistes de travail qui n'ont pu être explorées par manque de temps.

6.7.1 Extraction de tempo

Nous avons vu que les algorithmes moléculaires permettent d'obtenir des objets proches des notes. Etant donnés les onsets de ces objets, il est alors possible d'en déduire le tempo. L'utilisation d'une représentation objet pour la détection de tempo est de contourner un problème inhérent aux méthodes actuelles de détection d'onset : elles sont généralement basées sur la détection de changements dans le signal. Cela a pour conséquence une sensibilité aux modulations à l'intérieur des notes, et également de mauvaises performances lorsque les attaques ne sont pas franches.

On peut ainsi espérer que les algorithmes moléculaires rendent possible l'estimation de tempo en musique classique, qui reste le style musical le plus problématique pour cette tâche.

6.7.2 Algorithmes traitant de données symboliques

Les représentations obtenues semblent pertinentes pour d'autres post-traitements mettant en jeu des données symboliques. On peut par exemple adapter des algorithmes qui permettent de traiter des fichiers MIDI, comme par exemple les algorithmes de détermination de clé musicale (Chew (2001)) ou de similarité mélodique (Hu et al. (2002)).

6.7.3 Edition musicale

La représentation objet du son peut servir de base à une manipulation intuitive du son, en modifiant les attributs des objets extraits.

On peut penser à des effets classiques d'étirement temporel (*time-stretching*), de changement de hauteur (*pitch-shifting*). On peut également modifier l'instrument qui interprète le morceau, en exploitant à la fois les caractéristiques portées par les enveloppes spectrales et les variations temporelles de l'amplitude et de la fréquence fondamentale à l'intérieur des molécules.

6.7.4 Séparation de source / *remixing*

L'algorithme a pour but d'identifier les sources en présence dans le signal musical. Etant donné qu'on peut attribuer chaque objet (atome ou molécule) à une source, on peut construire des livres spécifiques à chacune des sources, puis les resynthétiser séparément afin d'effectuer une séparation de sources. Des opérations de remixage peuvent également être effectuées, en réhaussant une des sources puis en les réadditionnant.

Comme pour les opérations de codage, la réussite de ces deux processus seront très dépendants de la qualité de la transcription effectuée par l'algorithme.

6.8 Bilan

Dans ce chapitre, nous avons montré que les représentations objet du son que nous avons extraites peuvent être appliquées à un nombre important de tâches moyennant des post-traitements assez simples : elles sont suffisamment génériques pour que l'informations utile en indexation soit conservée.

Conclusion

Bilan

Dans ce travail, nous avons choisi d'approcher l'indexation audio en adoptant une représentation objet du son. Les décompositions du son que nous avons mises en oeuvre s'appuient sur le domaine des représentations parcimonieuses, dont la flexibilité a permis d'une part de définir des atomes d'une structure relativement complexe, et d'autre part de lier l'espace de représentation à des connaissances sur des sources analysées. Différentes variantes de structures harmoniques ont été proposées, permettant de couvrir une grande partie des sons possédant une hauteur. Des algorithmes efficaces ont été élaborés, permettant d'obtenir des représentations pour des durées d'exécution qui peuvent atteindre le temps réél si une implémentation adéquate est effectuée. A partir de ces représentations, différents traitements ont été effectués : la reconnaissance d'instruments sur les solos permet d'obtenir des scores proches de l'état de l'art, la reconnaissance de hauteur de notes atteint des résultats satisfaisants, et les perspectives de traitement sur des mélanges réels sont encourageantes, en mono et en stéréo.

Contributions

Nous avons proposé une nouvelle approche pour l'indexation des signaux musicaux, ou plus précisément leur représentation comme un ensemble structuré d'objets : à l'aide d'un dictionnaire contenant des exemples paramétrisés des sources, le problème de représentation objet du signal musical peut être vu comme un problème d'optimisation où il s'agit de trouver quelle est la combinaison linéaire des sources qui permet de minimiser l'erreur de modélisation sous une contrainte de parcimonie.

La définition des modèles de sources (du dictionnaire) est un point critique que nos avons abordé afin de trouver un compromis entre l'adaptabilité du modèle au signal, la compacité de la paramétrisation et la représentativité des sources étudiées. Les atomes et molécules proposés représentent de façon adéquate les parties stationnaires des sources instrumentales visées par notre étude, et permettent de lier de façon originale la décomposition à des connaissances sur les sources mises en jeu en effectuant un apprentissage assez supervisé au préalable.

La partie algorithmie a demandé la recherche d'algorithmes rapides pour l'extraction de groupements pertinents d'atomes (molécules), permettant de s'adapter à la fois à la structure harmonique et aux variations à long terme du signal, aspects souvent difficiles à concilier dans les décompositions du son. L'approche par pénalisation de la longueur des chemins permet d'obtenir des objets proches des notes de musique, et constituent ainsi une base pertinente pour des traitements ultérieurs. Cette approche pourrait être étendue

à d'autres types de dictionnaires, comme par exemple à des dictionnaires de Gabor pour l'extraction de partiels, et plus généralement à tout dictionnaire à partir desquels on veut construire des molécules avec une contrainte de compacité de représentation. On peut également signaler que les optimisations de paramètres proposées permettent de mieux faire coïncider les atomes extraits au signal analysé. Là aussi, ce principe peut être appliqué en général aux décompositions pour lesquelles on espère extraire une grande partie du signal pendant les premières itérations en calculant explicitement le gradient de la projection du signal sur les atomes par rapport aux paramètres à optimiser.

Enfin, du point de vue applicatif, les décompositions ont été tout d'abord validées sur des expériences où une source instrumentale était activée. Nous avons montré que le fait d'apprendre des atomes sur des classes spécifiques pouvait servir à effectuer des classifications de signaux inconnus, une fois une décomposition effectuée, même sous-optimale. La possibilité de faire varier tous les paramètres des atomes a été appliquée sur différentes tâches, et permet d'explorer de nouvelles voies par rapport aux approches existantes basées sur des modèles. Des décompositions avec plusieurs résolutions ont par exemple été employées, néanmoins sans succès significatif par rapport aux approches monorésolution dans les signaux traités (comportant peu de transitoires). Nous avons également montré qu'il était possible, avec des post-traitements simples, d'aborder des problèmes encore peu traités en indexation et traitement du signal audio, comme la localisation, le dénombrement et l'identification d'instruments dans des mélanges. Nous avons aussi proposé une approche pour identifier des ensembles où plusieurs instruments de la même classe peuvent être mis en jeu, ce qui constitue un cas de figure qui n'a jamais été traité. D'autres problèmes ont été abordés : la transcription et le codage objet des sons musicaux polyphoniques à extrêmement bas débit, problèmes qui apparaissent très liés. Cependant, des améliorations sont à apporter afin d'obtenir des résultats plus probants sur ces tâches, aussi bien au niveau de la décomposition qu'au niveau des post-traitements.

Perspectives

Dressons maintenant les perspectives issues de ce travail.

Dictionnaires

Concernant les modèles de signaux, nous avons utilisé un modèle relativement supervisé, où la position des partiels est assez rigide (harmonicité stricte ou inharmonicité donnée par un paramètre). Il pourrait donc être intéressant de permettre une certaine liberté dans le positionnement des partiels afin de traiter les sources instrumentales qui ne sont pas abordées dans ce travail. L'utilisation de méthodes non-supervisées pour l'apprentissage des vecteurs d'amplitudes, comme la factorisation en matrices non-négatives, pourrait être intéressante. Les atomes utilisés pourraient également être adaptés sur les signaux analysés, en exploitant par exemple des portions des pièces musicales où les sources jouent isolément pour rendre les mélanges moins ambigus.

On peut ensuite exploiter la possibilité offerte par le cadre des représentations parcimonieuses de pouvoir introduire des formes d'onde de nature différente, notamment des dictionnaires permettant de bien représenter les phénomènes impulsifs. On peut notamment penser à des ondelettes dyadiques, ou alors des sinusoïdes amorties en rapport quasi-harmonique qui permettraient, par exemple, de modéliser les transitoires du piano.

Concernant les molécules, tous les cas n'ont pas été traités, on peut penser notamment au cas de la multi-résolution et celui des atomes stéréo, où seuls des algorithmes atomiques ont été proposés. On peut également réfléchir à l'ajout de couches au dessus des molécules proposées (ou *méta-molécules*), qui permettraient d'extraire directement des lignes mélodiques ou des accords. La question qui reste en suspend serait de décider si ces structures seraient à placer dans un dictionnaire d'accord ou de mélodie, ou si elles seraient calculées *ad hoc* sur le signal en utilisant des contraintes de longueur de description comme nous l'avons fait dans le cas des molécules.

Algorithmes

L'approche moléculaire pourrait également s'étendre à la polyphonie : il s'agirait cette fois-ci de sélectionner des molécules composées d'atomes harmoniques ayant le même support temporel. Dans ce cas, l'orthogonalité variable des atomes entre eux, selon qu'ils soient en rapports harmoniques ou non, poserait des problèmes de complexité car il faudrait cette fois évaluer les poids *a posteriori* pour un grand nombre de combinaisons d'atomes. On pourrait cependant aborder le problème en examinant un nombre restreint de combinaisons d'atomes, en les choisissant dans un sous-ensemble d'atomes pour lesquels les corrélations avec le signal sont au-dessus d'un certain seuil.

Quelques algorithmes gloutons intéressants n'ont pas été essayés, comme le High Resolution Matching Pursuit qui permettrait de définir les phénomènes transitoires de façon plus précise avec un dictionnaire multi-résolution. Il serait intéressant d'essayer d'autres algorithmes utilisant des optimisations plus globales (comme FOCUSS, Basis Pursuit) pour traiter le problème, néanmoins au prix d'une complexité plus élevée.

Concernant le passage à des échelles réalistes, c'est-à-dire par exemple 40 instruments, le principe du Matching Pursuit peut être gardé, en mettant en oeuvre une organisation hiérarchique du dictionnaire.

Applications

Les applications pourraient bénéficier des améliorations que nous venons de proposer au niveau des modèles de signaux et des algorithmes. Concernant la reconnaissance des instruments de musique, tous les paramètres utilisés dans l'état de l'art n'ont pas été exploités, notamment les paramètres de modulation d'amplitude et de fréquence, ainsi que ceux sur la partie inharmonique du son. L'estimation de hauteur sur des mélanges polyphoniques profiterait également d'une meilleure gestion de la polyphonie au niveau de l'algorithme de décomposition. Une amélioration des post-traitements afin de tenir compte de considérations musicologiques serait également envisageable. Des modélisations statistiques sur les représentations objets permettraient également de tenir compte d'informations provenant de l'organisation des atomes. On peut notamment penser à l'utilisation de l'extraction de lignes mélodiques ou d'a priori sur les accords et instrumentations probables. Ces aspects pourraient également être introduits par une pondération préalable des atomes du dictionnaire. Enfin, une voie intéressante à explorer serait l'adaptation du dictionnaire au signal analysé. En effet, la musique est rarement d'une complexité uniforme : les sources sont parfois présentes seules sur certains segments. Exploiter ces zones peu ambigües serait sans doute fructueux.

Conclusion générale

Nous avons proposé une approche pour l'indexation et le traitement du signal audio basée sur l'extraction d'objets sonores. Si les objets sonores que nous pouvons extraire semblent suffisamment valides pour un certain nombre d'applications comme la reconnaissance des instruments et l'estimation de hauteur de note, leur extraction dans des situations complexes comme les mélanges polyphoniques atteint un degré de performance qui n'est pas encore satisfaisant pour des cas pratiques. Comme nous l'avons souligné, l'amélioration des performances des applications pourrait venir d'algorithmes nécessitant une complexité accrue et donc rendant caduque l'applicabilité de notre méthodologie. L'utilisation de connaissances musicologiques en post-traitement pourrait également améliorer les performances, mais cela suppose que toutes les informations utiles soient encore présentes dans le résultat de la décomposition, ce qui n'est pas garanti dans des situations polyphoniques complexes où les sources se perturbent entre elles.

Le paradigme d'analyse que nous proposons est constitué de deux étapes assez classique : une étape de représentation où le signal est exprimé dans un espace où il exhibe des propriétés plus significatives que dans sa forme originale, et une étape d'interprétation consistant à un traitement effectué sur la représentation afin d'en extraire l'information haut-niveau désirée. L'originalité de notre approche est d'injecter très tôt de la connaissance dans la représentation en liant chacun des éléments de la représentation à un modèle de source. Une question intéressante à explorer est de savoir si l'on peut aller plus loin dans ce processus, par exemple en incorporant des informations d'enveloppes temporelles pour les sources, puis éventuellement sur des lignes mélodiques ou des accords probables etc. Si ce processus devrait être possible techniquement et devrait améliorer les résultats, il présenterait l'inconvénient de *spécialiser* le système : traiter toute la combinatoire serait extrêmement coûteux, il faudrait donc restreindre l'ensemble des possibilités.

Ainsi, il semble peu réaliste de définir un système composé d'un seul algorithme et muni d'un dictionnaire unique permettant de traiter toute la musique existante, et de la décomposer en objets sonores pertinents. Par contre, des algorithmes tels que nous les avons définis peuvent convenablement s'insérer dans un système qui définirait un *contexte* d'analyse pour le signal, à l'aide de données annexes au contenu, ou alors d'un système mettant en jeu d'autres agents qui permettraient d'extraire des informations grossières, comme le tempo, l'harmonie, les motifs qui se répètent, les ensembles d'instruments probables avec d'autres algorithmes. D'ailleurs, des algorithmes de décompositions avec des dictionnaires adaptés et généraux pourraient jouer le rôle de ces agents. Une fois l'univers des possibles ainsi restreint, des algorithmes avec des dictionnaires plus spécifiques comme nous les avons définis pourraient être mis en jeu afin d'obtenir une description précise du signal. Finalement, si l'on veut obtenir la décomposition objet d'un signal musical quelconque, un point clé est de définir quelle est la bonne articulation entre les processus *bottom-up* (décomposition du signal à l'aide d'un dictionnaire, avec éventuellement une adaptation du dictionnaire) et *top-down* (construction du dictionnaire adapté, lié à des connaissances, avant la décomposition) afin de rendre l'exécution réalisable dans un contexte ouvert.

Bibliographie personnelle

Journal

– Leveau, P., Vincent, E., Richard, G. & Daudet, L. (2008). Instrument-specific harmonic atoms for mid-level music representation, *to appear in IEEE Trans. on Audio, Speech and Language Processing*.

Conférences

– Sodoyer, D., Leveau, P. & Daudet, L. (2007). Using stereo information for instrument identification in polyphonic mixtures, *IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.
– Leveau, P., Sodoyer, D. & Daudet, L. (2007). Automatic Instrument Recognition in a Polyphonic Mixture using Sparse Representations, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.
– Cornuz, G., Ravelli, E., Leveau, P. & Daudet, L. (2007). Object coding of harmonic sounds using sparse and structured representations, *Proc. of COST G-6 Conference on Digital Audio Effects (DAFX)*.
– Richard, G., Leveau, P., Essid, S., Daudet, L. & David, B. (2007). Towards Polyphonic Instrument Recognition, *Proc. of Int. Congress on Acoustics (ICA)*.
– Leveau, P., Vincent, E. , Richard, G. & Daudet, L. (2006). Mid-level sparse representations for timbre identification : design of an instrument-specific harmonic dictionary, *1st Workshop on Learning the Semantics of Audio Signals (LSAS)*.
– Leveau, P. & Daudet, L. (2006). Multi-resolution partial tracking with modified matching pursuit, *Proc. of European Signal Processing Conference (EUSIPCO)*.
– Leveau, P., Daudet, L., Krstulovic, S. & Gribonval, R. (2005). Model-Based Matching Pursuit - Estimation of Chirp Factors and Scale of Gabor Atoms with Iterative Extension, *In Proc. Signal Proc. with Adaptive Sparse Structured Representations (SPARS'05)*.
– Krstulovic, S., Gribonval, R., Leveau, P. & Daudet, L. (2005). A Comparaison of two Extensions of the Matching Pursuit Algorithm for the Harmonic Decomposition of Sounds, *IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.
– Essid, L., Leveau, P., Richard, G., Daudet, L. & David, B. (2005). On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments, *Proc. AES 118th Convention*.
– Leveau, P., Daudet, L. & Richard, G. (2004). Methodology and tools for the evaluation of automatic onset detection algorithms in music, *Proc. of Int. Conf. on Music*

Information Retrieval (ISMIR).

Bibliographie

Aharon, M., Elad, M. & Bruckstein, A. (2006). K-svd : an Algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Trans. on Signal Processing* **54**(11) : 4311–4322.

Aucouturier, J.-J. (2006). *Ten Experiments on the Modelling of Polyphonic Timbre*, PhD thesis, University Pierre et Marie Curie, Paris, France.

(auteur inconnu) (n.d.). Iowa database. Available from : `http://theremin.music.uiowa.edu/MIS.html`.

Bello, J. P. & Pickens, J. (2005). A robust mid-level representation for harmonic content in music signals, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 304–311.

Brown, J. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features, *Journal of the Acoustical Society of America* **105** : 1933.

Bultan, A. (1999). A four-parameter atomic decomposition of chirplets, *IEEE Trans. on Signal Processing* **47**(3) : 731–745.

Casey, M. & Westner, A. (2000). Separation of mixed audio sources by independent subspace analysis, *Proc. of Int. Conf. on Computer Music (ICMC)*.

Chen, S. S., Donoho, D. L. & Saunders, M. A. (2001). Atomic decomposition by basis pursuit, *SIAM review* **43**(1) : 129–159.

Chétry, N. & Sandler, M. (2006). Linear predictive models for musical instrument identification, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Chew, E. (2001). Modeling Tonality : Applications to Music Cognition, *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* pp. 206–211.

Coifman, R. & Wickerhauser, M. (1992). Entropy-based algorithms for best basis selection, *IEEE Trans. on Information Theory* **38**(2 Part 2) : 713–718.

Cornuz, G., Ravelli, E., Leveau, P. & Daudet, L. (2007). Object coding of harmonic sounds using sparse and structured representations, *Proc. of COST G-6 Conference on Digital Audio Effects (DAFX)*.

Daniel, A. (2007). *Évaluation des Méthodes de Transcription Automatique de la Musique*, Master's thesis, Université Pierre et Marie Curie.

Daudet, L. (2006). Sparse and structured decompositions of signals with the molecular matching pursuit, *IEEE Trans. on Audio, Speech and Language Processing* pp. 1808–1816.

Daudet, L. & Sandler, M. (2004). MDCT analysis of sinusoids : exact results and applications to coding artifacts reduction, *IEEE Trans. on Speech and Audio Processing* **12**(3) : 302–312.

de Cheveigné, A. & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America* **4**(111) : 1917 – 1930.

den Brinker, A., Schuijers, E. & Oomen, A. (2002). Parametric coding for high-quality audio, *Proceedings of the 112th AES Convention*, Munich, Germany.

Depalle, P., Garcia, G. & Rodet, X. (1993). Tracking of partials for additive sound synthesis using hidden markov models, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Duda, R., Hart, P. & Stork, D. (2000). *Pattern Classification*, Wiley-Interscience.

Eggink, J. & Brown, G. J. (2003). Application of missing feature theory to the recognition of musical instruments in polyphonic audio, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.

Eggink, J. & Brown, G. J. (2004). Instrument recognition in accompanied sonatas and concertos, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Ellis, D. & Poliner, G. (2006). Classification-based melody transcription, *Machine Learning* **65**(2) : 439–456.

Ellis, D. P. W. & Rosenthal, D. F. (1998). Mid-level representations for computational auditory scene analysis, *in* D. F. Rosenthal & H. G. Okuno (eds), *Computational auditory scene analysis*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 257–272.

Emiya, V., David, B. & Badeau, R. (2007). A parametric method for pitch estimation of piano tones, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Eronen, A. & Klapuri, A. (2000). Musical instrument recognition using cepstral cofficients and temporal features.

Escoda, O., Granai, L. & Vandergheynst, P. (2006). On the use of a priori information for sparse signal approximations, *IEEE Trans. on Signal Processing* **54** : 3468–3482.

Essid, S. (2005). *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique (in French)*, PhD thesis, Université Pierre et Marie Curie.

Essid, S., Leveau, P., Richard, G., Daudet, L. & David, B. (2005). On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments, *Proc. AES 118th Convention*.

Essid, S., Richard, G. & David, B. (2006a). Instrument recognition in polyphonic music based on automatic taxonomies, *IEEE Trans. on Audio, Speech and Language Processing* **14**(1) : 68–80.

Essid, S., Richard, G. & David, B. (2006b). Musical instrument recognition by pairwise classification strategies, *IEEE Trans. on Audio, Speech and Language Processing*.

Févotte, C., Daudet, L., Godsill, S. & Torrésani, B. (2006). SPARSE REGRESSION WITH STRUCTURED PRIORS : APPLICATION TO AUDIO DENOISING, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Vol. 3.

Févotte, C., Godsill, S. & Wolfe, P. (2004). Bayesian approach for blind separation of underdetermined mixtures of sparse sources, *Proc. of Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*, Springer, pp. 398–405.

Forney Jr, G. (1973). The Viterbi algorithm, *Proc. of the IEEE* **61**(3) : 268–278.

Gabor, D. (1947). Acoustical quanta and the theory of hearing, *Nature* pp. 591–516.

George, E. B. & Smith, M. J. T. (1992). Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones, *Journal of the Audio Engineering Society* **40**(6) : 497–516.

Gersho, A. & Gray, R. M. (1991). *Vector Quantization and Signal Compression*, Springer.

Goodwin, M. & Vetterli, M. (1999). Matching pursuit and atomic signal models based on recursive filter banks, *IEEE Trans. on Signal Processing* **47**(7) : 1890–1902.

Goodwin, M. M. (2001). Multiscale overlap-add sinusoïdal modeling using matching pursuit and refinements, *in* IEEE (ed.), *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 207–210.

Gorodnitsky, I. & Rao, B. (1997). Sparse signal reconstruction from limited data using FOCUSS : are-weighted minimum norm algorithm, *IEEE Trans. on Signal Processing* **45**(3) : 600–616.

Goto, M., Hashiguchi, H., Nishimura, T. & Oka, R. (n.d.). RWC Musical Instrument Sound Database. Distributed online at http ://staff.aist.go.jp/m.goto/RWC-MDB/.

Gribonval, R. (1999). *Approximations non-linéaires pour l'analyse des signaux sonores (in French)*, PhD thesis, UNIVERSITÉ DE PARIS IX DAUPHINE.

Gribonval, R. (2001). Fast matching pursuit with a multiscale dictionary of Gaussian chirps, *IEEE Trans. on Signal Processing* **49**(5) : 994–1001.

Gribonval, R. (2002). Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Gribonval, R. & Bacry, E. (2003). Harmonic decomposition of audio signals with matching pursuit, *IEEE Trans. on Signal Processing* **51**(1) : 101–111.

Gribonval, R. & Nielsen, M. (2003). *Highly sparse representations from dictionaries are unique and independent of the sparseness measure*, Department of Mathematical Sciences, Aalborg University.

Gribonval, R., Bacry, E., Mallat, S., Depalle, P., Rodet, X. & IRCAM, P. (1996). Analysis of sound signals with high resolution matching pursuit, *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on* pp. 125–128.

Heusdens, R., Vafin, R. & Kleijn, W. B. (2002). Sinusoidal modeling using psychoacoustic-adaptive matching pursuits, *IEEE Signal Processing Letters* **9**(8) : 262–265.

Hu, N., Dannenberg, R. & Lewis, A. (2002). A Probabilistic Model of Melodic Similarity, *Proc. of Int. Conf. on Computer Music (ICMC)*.

Ircam (n.d.). Studio online database. Available from : `http://forumnet.ircam.fr/402.html?L=1`.

ISO/IEC 14496-3 :2001 (2001). Information technology - Coding of Audio-Visual Objects - part 3 : Audio.

ITU (2003). ITU-R BS.1534-1 : Method for the subjective assessment of intermediate quality levels of coding systems.

Jensen, J. & Heusdens, R. (2003). A comparison of differential schemes for low-rate sinusoidal audio coding, *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 205–208.

Jinachitra, P. (2004). Polyphonic instrument identification using independent subspace analysis, *Proc. of Int. Conf. on Multimedia and Expo (ICME)*.

Jost, P., Vandergheynst, P. & Frossard, P. (2006). Tree-based pursuit : Algorithm and properties, *IEEE Trans. on Signal Processing* **54**(12) : 4685–4697.

Kashino, K. & Murase, H. (1999). A sound source identification system for ensemble music based on template adaptation and music stream extraction, *Speech Communication* **27** : 337–349.

Kitahara, T., Goto, M., Komatani, K., Ogata, T. & Okuno, G. (2006). Instrogram : A new musical instrument recognition technique without using onset detection nor f0 estimation, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Vol. 5, pp. 229–232.

Kitahara, T., Goto, M., Komatani, K., Ogata, T. & Okuno, H. (2005). Instrument identification in polyphonic music : feature weighting with mixed sounds, pitch-dependent timber modeling, and use of musical context, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.

Klapuri, A. & Davy, M. (eds) (2006). *Signal Processing Methods for Music Transcription*, Springer, New York, NY.

Klapuri, A. P. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.

Krstulovic, S. & Gribonval, R. (2006). Mptk : matching pursuit made tractable, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Lagrange, M., Marchand, S. & Rault, J.-B. (2004). Using linear prediction to enhance the tracking of partials, *in* IEEE (ed.), *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Lee, D. & Seung, H. (2001). Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems* **13** : 556–562.

Lesage, S. (2007). *Apprentissage de dictionnaires structurés pour la modélisation parcimonieuse de signaux multicanaux*, PhD thesis, Université de Rennes 1.

Leveau, P. (2004). *Paramètres adaptés pour la reconnaissance automatique des instruments de musique*, Master's thesis, Universitéé Pierre et Marie Curie (Paris 6).

Leveau, P. & Daudet, L. (2006). Multi-resolution partial tracking with modified matching pursuit, *Proc. of European Signal Processing Conference (EUSIPCO)*.

Leveau, P., Sodoyer, D. & Daudet, L. (2007). Automatic Instrument Recognition in a Polyphonic Mixture using Sparse Representations, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, Vienne, Autriche.

Leveau, P., Vincent, E., Richard, G. & Daudet, L. (2008). Instrument-specific harmonic atoms for mid-level music representation, *to appear in IEEE Trans. on Audio, Speech and Language Processing*.

Li, T. & Ogihara, M. (2005). Music genre classification with taxonomy, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Logan, B. T. (2000). Mel frequency cepstral coefficients for music modeling, *Proc. of Int. Symp. on Music Information Retrieval (ISMIR)*.

Mallat, S. (2000). *Une exploration des signaux en ondelettes*, Les Editions de l'Ecole Polytechnique.

Mallat, S. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries, *IEEE Trans. on Signal Processing* **41**(12) : 3397–3415.

Marolt, M. (2006). A mid-level melody-based representation for calculating audio similarity, *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.

Martin, K. (1999). *Sound-Source Recognition : A Theory and Computational Model*, PhD thesis, Massachusetts Institute of Technology.

McAdams, S., Winsberg, S., de Soete, G. & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : common dimensions, specficities and latent subject classes., *Psychological Research* (58) : 177–192.

McAulay, R. & Quatieri, T. (1986). Speech analysis/Synthesis based on a sinusoidal representation, *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on* **34**(4) : 744–754.

Molla, S. (2003). *Signaux audiophoniques : modelisation hybride et schema de codage*, PhD thesis, Université de Provence, Marseille.

Olshausen, B. & Field, D. (1997). Sparse coding with an overcomplete basis set : A strategy employed by V1, *Vision Research* **37**(23) : 3311–3325.

Ozerov, A., Philippe, P., Gribonval, R. & Bimbot, F. (2005). One microphone singing voice separation using source-adapted models, *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* pp. 90–93.

Parsons, D. (1975). *The Directory of Tunes and Musical Themes*, Spencer Brown.

Pati, Y., Rezaiifar, R. & Krishnaprasad, P. S. (1993). Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition, *Proc. 27th Asilomar Conf. on Signals, Systems and Computers*.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project, *Technical report*, IRCAM, Paris.

Peeters, G. & Rodet, X. (2003). Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments Databases, *Proc. of COST G-6 Conference on Digital Audio Effects (DAFX)*.

Purnhagen, H. & Meine, N. (2000). HILN-the MPEG-4 parametric audio coding tools, *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2000)*, Vol. 3, pp. 201–204.

Rodet, X. (1984). Time-domain formant-wave-function synthesis., *COMP. MUSIC J.* **8**(3) : 9–14.

Rossing, T. & Fletcher, N. (1998). *The Physics of Musical Instruments*, Springer.

Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, PhD thesis, Stanford University.

Smaragdis, P. & Brown, J. (2003). Non-negative matrix factorization for polyphonic music transcription, *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180.

Sodoyer, D., Leveau, P. & Daudet, L. (2007). Using stereo information for instrument identification in polyphonic mixtures, *Proc. of IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Song, J., Bae, S. & Yoon, K. (2002). Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System, pp. 133–139.

Temlyakov, V. (2000). Weak greedy algorithms, *Advances in Computational Mathematics* **12**(2,3) : 213–227.

Tropp, J. (2004). Greed is good : algorithmic results for sparse approximation, *IEEE Trans. on Information Theory* **50**(10) : 2231–2242.

T.Toivonen (n.d.). Timidity. Available at http ://timidity.sourceforge.net.

Tzanetakis, G. & Cook, P. (2002). Musical Genre Classification of Audio Signals, *IEEE Trans. on Speech and Audio Processing* **10**(5) : 293.

Verma, T. & Meng, T. (1998). An analysis/synthesis tool for transient signals that allows aflexible sines+ transients+ noise model for audio, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Vol. 6.

Verma, T. & Meng, T. (1999). Sinusoidal modeling using frame-based perceptually weighted matching pursuits, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Vincent, E. (2004). *Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux*, PhD thesis, University Pierre et Marie Curie.

Vincent, E. (2006). Musical source separation using time-frequency source priors, *IEEE Trans. on Audio, Speech and Language Processing* **14**(1) : 91–98.

Vincent, E. & Plumbley, M. (2007). Low bitrate object coding of musical audio using bayesian harmonic models, *IEEE Trans. on Audio, Speech and Language Processing*.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, *IEEE Trans. on Audio, Speech and Language Processing* **15**(3) : 1066–1074.

Vogel, B. K., Jordan, M. I. & Wessel, D. (2005). Multi-instrument musical transcription using a dynamic graphical model, *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*.

Vos, K., Vafin, R., Heusdens, R. & Kleijn, W. (1999). High-quality consistent analysis-synthesis in sinusoidal coding, *17th Audio Engineering Society International Conference* pp. 244–250.

Wikipedia (2007). Rasoir d'occam. Available from : `http://fr.wikipedia.org/wiki/Rasoir_d%27Occam`.

Witten, I. H., Neal, R. M. & Cleary, J. G. (1987). Arithmetic coding for data compression, *Commun. ACM* **30**(6) : 520–540.

Zils, A. (2004). *Extraction de descripteurs musicaux : une approche évolutionniste*, PhD thesis, Université Pierre et Marie Curie.

Annexe

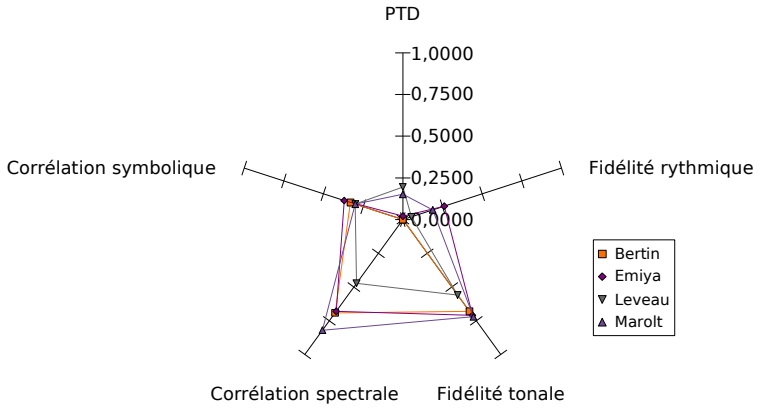
Résultats du stage d'Adrien Daniel sur la transcription automatique de piano (Se référer au rapport de Master pour plus de détails). L'algorithme présenté est nommé "Leveau".

1ère Page : Evaluation sur des critères calculés. 5 critères sont évalués :
- PTD : distance par transfert de poids.
- Corrélation symbolique : corrélation des paramètres des données symboliques.
- Corrélation spectrale : corrélation des spectres reconstitués de la transcription et de l'original.
- Fidélité tonale : évalue dans quelle mesure les notes extraites appartiennent à la tonalité jouée.
- Fidélité rythmique : évalue dans quell mesure les accents rythmiques de la transcription corresponde à ceux de l'original.

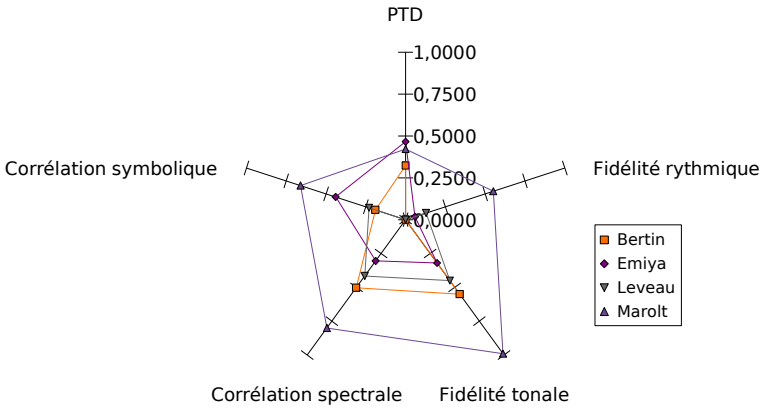
2ème Page : Evaluation perceptive. Chaque barre de couleur représente l'évaluation d'un algorithme différent.

Évaluation

	PTD	Corrélation symbolique	Corrélation spectrale	Fidélité tonale	Fidélité rythmique
Bertin	0,0000	0,3300	0,6908	0,6800	0,0000
Emiya	0,0206	0,3700	0,6805	0,7100	0,2615
Leveau	0,1956	0,3000	0,4718	0,5600	0,0543
Marolt	0,1513	0,3000	0,8195	0,7200	0,1884



	PTD	Corrélation symbolique	Corrélation spectrale	Fidélité tonale	Fidélité rythmique
Bertin	0,3233	0,1900	0,5029	0,5500	0,0000
Emiya	0,4657	0,4400	0,3049	0,3200	0,0591
Leveau	0,0000	0,2300	0,4159	0,4500	0,1284
Marolt	0,4222	0,6600	0,8006	0,9900	0,5505



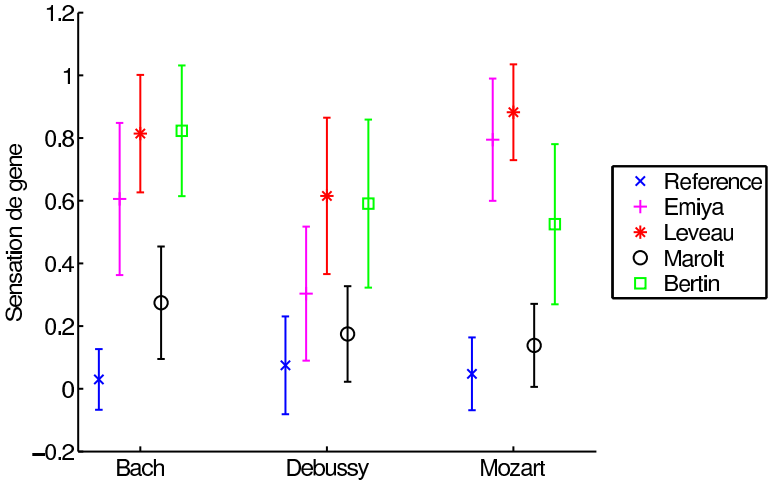


FIG. 6.4 – Résultats de l'étape 1 (sujets musiciens et non-musiciens confondus).

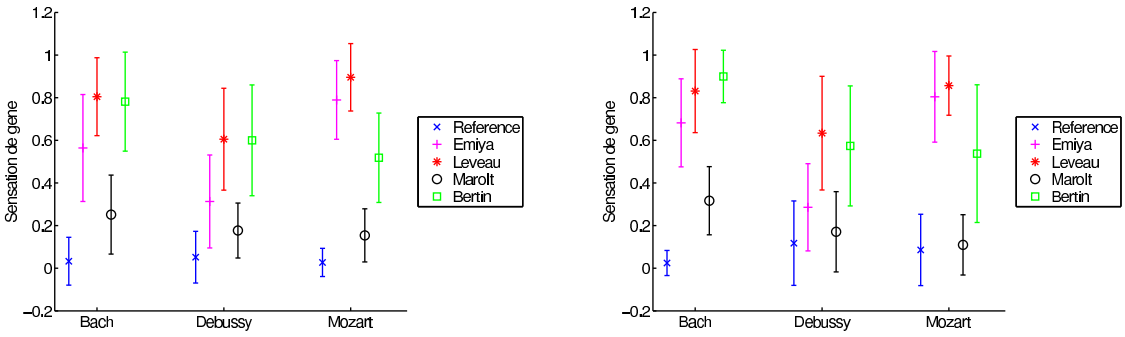


FIG. 6.5 – Résultats de l'étape 1 (sujets musiciens à gauche et non-musiciens à droite).