Séparation de sources audio informée par tatouage pour mélanges linéaires instantanés stationnaires

Mathieu Parvaix

▶ **To cite this version:**

Mathieu Parvaix. Séparation de sources audio informée par tatouage pour mélanges linéaires instantanés stationnaires. Sciences de l'ingénieur [physics]. Institut National Polytechnique de Grenoble - INPG, 2010. Français. NNT : . tel-00558209

HAL Id: tel-00558209

https://theses.hal.science/tel-00558209

Submitted on 21 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE DE GRENOBLE
INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque : |_|_|_|_|_|_|_|_|_|_|

THÈSE

pour obtenir le grade de

DOCTEUR DE L'Université de Grenoble
délivré par l'Institut polytechnique de Grenoble

Spécialité : ***Signal, Image, Parole, Télécoms***

préparée au laboratoire
Gipsa-Lab, Département Parole et Cognition, UMR CNRS 5216

dans le cadre de **l'École Doctorale :**
« **Électronique, Électrotechnique, Automatique et Traitement du Signal** »

présentée et soutenue publiquement

par

Mathieu Parvaix

le 23 Septembre 2010

Séparation de sources audio informée par tatouage pour mélanges linéaires instantanés stationnaires

Directeur de thèse : Laurent Girin (DPC)
Co-directeur de thèse : Jean-Marc Brossier (DIS)

JURY

M.	Christian Jutten,	Président
M.	Philippe Depalle,	Rapporteur
M.	Laurent Daudet,	Rapporteur
M.	Roland Badeau,	Éxaminateur
M.	Laurent Girin,	Directeur de thèse
M.	Jean-Marc Brossier,	Co-directeur de thèse

Résumé

Nous abordons dans cette thèse le problème de la séparation de sources selon un angle novateur à de nombreux niveaux. Ces travaux associent deux domaines du traitement du signal jusqu'alors traités de manière disjointe, la séparation de source et le tatouage numérique. Le procédé mis en place au cours de ces travaux a pour but de permettre à un utilisateur "client" de séparer les différents signaux numériques sources composant un mélange audio à partir de ce seul mélange tatoué. Pour ce faire un marquage du signal est effectué par un utilisateur "fournisseur" avant la fixation du mélange sur son support numérique. Ce marquage consiste en l'insertion sur le signal lui-même d'informations utiles à la séparation, et ceci de façon imperceptible. Le tatouage peut, en principe, être inséré soit sur le mélange, soit sur les signaux sources, qui sont disponibles à l'utilisateur fournisseur. Deux systèmes composent donc ce procédé, un encodeur qui permet à l'utilisateur fournisseur de réaliser la phase de mélange et de marquage, et un décodeur qui permet à l'utilisateur client de contrôler la séparation à partir du mélange. Au cours de cette thèse, il est choisi de tatouer le signal de mélange. Une application cible particulièrement visée est le cas d'un mélange polyphonique (signal de musique) fixé sur un support CD audio. La séparation doit permettre à l'utilisateur client d'effectuer un certain nombre de contrôles (par exemple le volume sonore) sur les différentes composantes de la scène sonore (les différents instruments et voix).

Mots clés

séparation de sources ; informée ; audio ; musique ; sous-déterminé ; codage source-canal ; parcimonie ; tatouage audio ; quantification vectorielle ; QIM ; remixage ; mélange linéaire instantané ; monophonique ; stéréophonique

Abstract

The source separation issue is addressed, in this PhD thesis, with an innovative point of view. This work joint associates two main domains in the signal processing area : digital watermarking and source separation which are most of the time considered unrelated. Our work aims at giving a "client-user" the possibility to separate different digital source signals that have been mixed together by the only use of their single mixture. To enable such a separation, a watermark is embedded by a "provider-user" into the signal before the mixture has been fixed on its digital support. This watermark which has to be imperceptibly inserted into the signal is made of several pieces of information from original signals. The message can be embedded either directly on source signals available to the "provider" before they are mixed or onto the mixture signal. This method is composed of two main parts, a coder where a "provider" can mix signals and embed the watermark, and a decoder where a client can control the separation based on the mixture signal study. In the present work, it was chosen to embed the watermark into the mixture signal. A typical application addressed by the proposed method is the process of Audio-CD polyphonic (stereo) music. The informed separation must enable a client to control several parameters (such as volume) of the different sources (instruments, voices) that compose the audio scene.

Keywords

source separation ; informed ; audio ; music ; under-determined ; source-canal coding ; sparcity ; audiowatermarking ; vectorial quantization ; QIM ; remixing ; linear instantaneous mixtures ; mono ; stereo

Remerciements

Je tiens avant tout à remercier mes directeurs de thèse Messieurs Laurent Girin et Jean-Marc Brossier pour m'avoir proposé ce sujet de thèse original qui m'a permis de cotoyer les domaines de la séparation de sources et du tatouage audio durant ces trois années. Je les remercie pour leur encadrement, leurs conseils, et leur très grande disponibilité. Ils ont su me guider dans mes travaux tout en me laissant la liberté d'explorer mes propres voies.

Je souhaite également remercier l'ensemble des membres du jury qui m'ont fait l'honneur de lire ce manuscrit, d'apporter leur expertise et leur caution scientifique : Messieurs Laurent Daudet et Philippe Depalle pour avoir rapporté ce manuscrit ainsi que pour la qualité et la pertinence de leurs remarques, Monsieur Christian Jutten pour avoir accepté de présider le jury et enfin Monsieur Roland Badeau, pour la justesse de ses remarques et de ses questions.

Je tiens à remercier tous les membres du projet DReaM, parmi lesquels Messieurs Antoine Liutkus et Sylvain Marchand. Cette collaboration, même de courte durée, a été pour moi très enrichissante.

Mes remerciements vont également à Monsieur Gérard Bailly, directeur du département Parole et Cognition du GIPSA-lab qui m'a permis de réaliser cette thèse au sein de son laboratoire. Je tiens également à remercier l'ensembre des membres du DPC, personnels, permanents et thésards pour m'avoir accepté parmi eux pendant ces trois années. Je remercie mes compagnions de labeur, les déjà ou bientôt docteurs Olivier, Viet Anh, Xavier, Aymeric, Lucille, Clément, Anahita, Zu Heng et les DISsiens Julien, Bertrand, Nico. Merci également à Maeva, Thomas et Nico pour les discussions animées autour de la machine à café. Sans oublier ma "co-burotte" Amélie pour m'avoir supporté pendant les deux dernières années, ainsi que les discussions enflammées que j'ai pu avoir avec mon directeur de thèse lors de corrections d'articles ou de manuscrit. Merci aussi à ceux qui n'ont fait qu'un passage au sein du DPC mais ont largement contribué à égayer ces trois années : Ronald, Emre, Martin, Keigo et j'en oublie certainement.

Je souhaite également remercier Cléo Baras pour m'avoir permis de faire mes premiers pas dans le monde de l'enseignement lors de vacations à l'IUT1 de Grenoble. Un

grand merci également à Jonathan Pinel pour son travail sur le tatouage audio haute capacité qui représente une composante très importante de l'approche de séparation de sources informée.

Merci également à Monsieur Guillaume Boyer, du cabinet de conseils en propriété industrielle Casalonga, et à Madame Isabelle Chery, chargée de valorisation au sein du groupe Grenoble INP, pour avoir pris part à la valorisation de la première partie de ces travaux de thèse.

Car je pense que l'existence est faite d'opportunités, de la rencontre des bonnes personnes, au bon endroit et au bon moment, je tiens à remercier Cornel Ioana qui m'a permis de faire mes premiers pas de "chercheur" au cours de mon stage de Master, m'a donné l'opportunité de découvrir la richesse et l'étendue du monde, et a été à l'origine, avec Laurent Girin et Jean-Marc Brossier, de ma candidature à cette thèse. Pour m'avoir orienté vers quelques uns de ces contacts industriels qui m'ont permis de postuler à mon premier emploi de docteur, je tiens également à remercier Monsieur Gaël Richard de Telecom Paris Tech, sans qui je ne connaîtrais pas les bienfaits de la météo californienne.

Et bien sûr un très grand merci à ma famille qui m'a apporté le soutien nécessaire durant les périodes difficiles. Merci pour leur patience, et leur écoute de tous les instants. Merci de m'avoir toujours encouragé à pousser plus loin mes études, et à franchir les frontières.
Et enfin pour l'énergie et le bonheur qu'elle m'apporte au quotidien, je n'ai pas assez de mots pour remercier Mariko.

Table des matières

Table des figures

Liste des tableaux

Acronymes

AAC	Advanced Audio Coding.
BF	Basses Fréquences.
BZ	(Méthode) de Bofill et Zibulevski.
CASA	Computational Auditory Scene Analysis.
CD	Compact Disk.
DOA	Direction of Arrival.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DRM	Digital Right Management.
DWT	Discrete Wavelet Transform.
EMD	Empirical Mode Decomposition.
HF	Hautes Fréquences.
HRIR	Head Related Impulse Response.
HRTF	Head Related Transfer Function.
ICA	Independent Component Analysis (ACI en français).
ILD	Interaural Time Difference.
ITD	Interaural Time Difference.
IMDCT	Inverse Modified Discrete Cosine Transform.
ISA	Independent Subspace Analysis (Analyse en Sous-Espaces Indépendants).
ISNR	Improvement of Signal to Noise Ratio.
KBD	Kaiser-Bessel Derived.
LBG	Linde Buzo Gray.
LIS	(Mélange) Linéaire Instantané Stationnaire.
LISM	(Mélange) Linéaire Instantané Stationnaire Monophonique.
LISS	(Mélange) Linéaire Instantané Stationnaire Stéréophonique.
LSB	Least Significant Bit.
MDCT	Modified Discrete Fourier Transform.
MGF	Moyenne Gain Forme.
MP	Matching Pursuit.
MPA	Modèle Psycho-Acoustique.
MMP	Molecular Matching Pursuit.
MPEG	Moving Pictures Experts Group.
ODG	Objective Difference Grade.
PCM	Pulse Coded Modulation.

QIM	Quantization Index Modulation.
QSU	Quantification Scalaire Uniforme.
SSI	Séparation de Sources Informée.
TDAC	Time Domain Aliasing Cancellation.
TIMIT	Texas Instrument Massachusetts Institute of Technology Base de données de signaux de parole.
SAC	Spatial Audio Coding.
SAR	Signal to Artifact Ratio.
SDR	Signal to Distortion Ratio.
SDG	Subjective Difference Grade.
SIR	Signal to Interference Ratio.
SMR	Signal to Mask Ratio.
SNR	Signal to Noise Ratio.
SNRout	Signal to Noise Ratio en sortie.
SNRin	Signal to Noise Ratio en entrée.
SSI-C	Séparation de Sources Informée par Codage.
SSI-I	Séparation de Sources Informée par Indexation.
SSI-C-I	Séparation de Sources Informée par Codage et Indexation.
TFCT	Transformée de Fourier à Court Terme.
TF	Temps-Fréquence.

Chapitre 1

Introduction

1.1 Le principe de la séparation de sources informée

1.1.1 À la croisée de deux domaines du traitement du signal

Le travail proposé au cours de cette thèse est à la croisée de deux grands domaines du traitement du signal : la séparation de sources d'une part et le tatouage des signaux d'autre part. La séparation de sources consiste à estimer des signaux sources à partir de l'observation d'un certain nombre de mélanges de ces mêmes signaux sources. L'objectif est généralement de rehausser, voire si possible d'extraire complètement un signal cible noyé dans un ensemble de signaux considérés comme parasites. Quant au tatouage numérique, aussi appelé *watermarking* en anglais, il consiste à insérer sur un signal une information binaire de façon imperceptible. Un des champs d'application principaux du tatouage est la protection des droits d'auteur pour des œuvres sur support numérique, et plus généralement la traçabilité d'informations sur ce type de support par ajout d'une information aidant à identifier un média. L'imperceptibilité du tatouage est un facteur fondamental : il ne doit cependant pas altérer la qualité perceptive du média sur lequel il est inséré.

Ces deux domaines du traitement du signal sont *a priori* disjoints et sont caractérisés par des techniques et des applications spécifiques. La combinaison de techniques issues de ces deux domaines offre donc un champ tout à fait original et des perspectives nouvelles pour la séparation de sources, notamment dans le cas difficile dit "sous-déterminé" où l'on dispose d'un nombre d'observations du mélange inférieur au nombre de sources présentes dans le mélange. L'extraction d'un signal source est dans ce cas très difficile, voire impossible en raison de la faible quantité d'information disponible dans ces mélanges par rapport à celle présente dans les sources. Un exemple flagrant qui, comme on le verra en détails par la suite, motive pour une bonne part ces travaux est le cas de signaux de musique sur CD audio. Ce cas de figure représente en effet une configuration (quasi) extrême de séparation sous-déterminée car on ne dispose que de deux voies stéréo (qui sont de plus très redondantes) pour un grand nombre potentiel de sources dans le cas d'une formation musicale riche. Le tatouage des signaux tel qu'il est proposé dans ce procédé doit précisément permettre de remédier à cette

difficulté et aider à séparer les diverses sources d'une scène polyphonique, de signaux de parole ou de musique.

1.1.2 Une configuration originale

Le cas le plus général de la séparation de sources est la séparation de sources dite "aveugle", dans laquelle on ne possède pas d'informations *a priori* sur les sources, ni même sur la nature du mélange. Le cas sous-déterminé où l'on dispose de moins d'observations de mélanges des sources que de sources elles-mêmes en est le cas de figure le plus délicat à traiter. Lorsque l'on possède une information partielle sur les sources (qui peut être de différents types, par exemple la nature même du signal, sa stationnarité...) ou sur la nature du mélange, on parle de séparation semi-aveugle. Bien sûr, on cherche dans ce cas à s'appuyer sur cette information partielle pour augmenter les performances de séparation.

Au cours de la présente thèse, nous poussons plus loin les hypothèses faites sur les signaux sources, et nous plaçons dans une configuration inédite en séparation de sources, quasiment à l'opposé de la séparation aveugle, et surprenante au premier abord : nous supposons que les signaux sources originaux (non mixés) sont disponibles. Plus précisément, pour que la notion même de séparation ait un sens, nous supposons que ces sources sont disponibles en amont du mélange, au niveau d'un utilisateur "fournisseur" qui réalise ce mélange [1] à partir des signaux sources, de façon plus ou moins contrôlée [2]. Le problème de la séparation se pose alors pour un second utilisateur "client" qui n'a lui à sa disposition que le signal mélange, et qui désire retrouver les signaux sources à partir de ce mélange. Il s'agit donc de profiter du fait que l'utilisateur fournisseur connaît les signaux sources, pour en extraire une série d'informations capables à la fois d'aider à la séparation, et d'être transmises implicitement à l'utilisateur client. C'est cette transmission qui est réalisée ici par un procédé de tatouage. En effet, d'une part on suppose qu'il n'y a pas de canal additionnel spécifique pour la transmission de ces informations. Et d'autre part, il faut que ces informations soient insérées de façon imperceptible pour n'avoir aucune incidence négative sur la qualité audio des signaux.

On comprend donc que la méthode de séparation que nous développons porte le nom de *séparation de sources informée*. En effet, le travail présenté au cours de cette thèse, s'il correspond à un cas sous-déterminé, voire au cas sous-déterminé extrême où on n'a qu'un seul signal de mélange, n'est en revanche en aucun cas une séparation aveugle. Au contraire, le terme "informée" fait référence à la somme d'informations sur

1. Le terme mixage en traitement audio de type réalisation de CD est plus approprié.

2. Dans la suite de cette introduction, par commodité on parle d'un signal de mélange, mais il peut s'agir d'un mélange multi-dimensionnel, et donc de plusieurs signaux de mélange, deux dans le cas stéréo, et plus dans d'autres applications. Comme pour les méthodes de séparation aveugles, différentes méthodes de séparation informée peuvent être envisagées en fonction du degré de sous-détermination du problème de séparation traité. Nous traiterons ce problème plus en détails par la suite. Nous restons à un niveau général dans cette introduction, sachant que l'on peut effectivement aller ici jusqu'à n'avoir qu'un seul signal de mélange.

les signaux sources utilisées pour aider à la séparation. Comme nous le verrons plus en détails par la suite, ces informations résultent d'un ensemble de traitements appliqués aux signaux sources pour en extraire, soit des descripteurs qui les caractérisent individuellement et doivent permettre de les différencier les uns des autres au sein de leur mélange, soit une information sur la composition locale du mélange pour permettre par la suite d'inverser le mélange.

1.2 Objectif de la thèse : une boucle complète de traitement

Le travail présenté dans cette thèse aborde la boucle complète de traitement en appliquant les principes décrits ci-dessus : à la fois au niveau de l'utilisateur "fournisseur" et au niveau de l'utilisateur "client". Ainsi, le but de ces travaux est la réalisation d'un procédé permettant à un utilisateur client de séparer les différents signaux numériques sources composant un mélange audio à partir de ce seul mélange. Pour ce faire un marquage du signal est effectué, en amont, par un utilisateur fournisseur avant la fixation du mélange sur son support numérique. Ce marquage consiste en l'insertion sur le signal lui-même d'informations utiles à la séparation, et ceci de façon imperceptible sous forme d'un tatouage numérique. Ces informations doivent être récupérées par l'utilisateur client qui, lui, ne dispose pas des signaux sources, mais seulement du signal de mélange. Elles doivent alors "guider" autant que possible la séparation des sources que l'on qualifie pour cette raison d'informée.

On peut noter que le tatouage des informations utiles à la séparation peut être effectué soit directement sur le signal de mélange audio, soit "indirectement" sur les signaux sources utilisés pour réaliser le mélange, ces signaux sources étant dans tous les cas disponibles à l'utilisateur fournisseur [3]. Ce principe est plus largement décrit au chapitre 4. Nous verrons par la suite qu'à ces deux approches fondamentales correspondent deux procédés relativement différents dans leurs fondements théoriques et techniques. Nous nous sommes concentrés durant ces travaux de thèse sur le tatouage du signal de mélange. Quelle que soit l'approche envisagée, deux systèmes composent ce procédé : un encodeur qui permet à l'utilisateur fournisseur de réaliser la phase de mélange et de marquage, et un décodeur qui permet à l'utilisateur client de contrôler la séparation (voir le chapitre 4). Cette thèse vise à réaliser l'élaboration de ces deux systèmes à partir de considérations théoriques, leur implémentation technique, et leur évaluation, notamment dans le cadre de l'application privilégiée de la conversion PCM 16 bits utilisée dans le format CD-audio ou le format non-compressé wav.

3. Le mélange et le marquage sont effectués dans cet ordre dans le cas du tatouage du mélange et dans l'ordre inverse dans le cas du tatouage des signaux sources.

Un exemple concret d'application

Reprenons l'exemple clé du CD audio évoqué en 1.1. De par son importance et de par le fait qu'il représente un cas d'école pour notre travail, cet exemple, et plus généralement la conversion PCM 16 bits, constituent un cadre d'étude et une application privilégiés dans la suite de cette thèse.

L'utilisateur client est ici l'utilisateur possédant un CD audio tatoué au moyen du procédé d'encodage proposé, et possédant le procédé de décodage correspondant. Ce décodeur peut être implémenté sous forme logicielle (pour des lecteurs multimédia) ou dans un stade plus avancé dans un lecteur dédié autonome, de type lecteur CD. Grâce à ce décodeur, l'utilisateur sera capable de manipuler séparément chaque source sonore de l'enregistrement. Cette manipulation pourra se faire au cours de la restitution d'un morceau (en jouant la piste) ou indépendamment de cette restitution (sans jouer cette piste). Il pourra ainsi, par exemple, modifier le volume d'une voix ou d'un des instruments du morceau et ainsi rehausser ou diminuer la contribution de cette voix ou de cet instrument dans la scène sonore, à sa convenance. Il pourra plus généralement leur appliquer un certain nombre d'effets audio (écho, changement de timbre, spatialisation...). Un musicien sera en mesure "d'éteindre" un instrument d'un morceau et de pouvoir jouer sur ce morceau à la place de l'instrument original (application de type karaoké généralisée aux instruments). Il sera également possible à un musicien "d'utiliser" l'instrument qu'il vient d'isoler comme échantillon pour la création de musique électronique ou à des fins pédagogiques [4]. Notons que le CD "tatoué" envisagé est compatible avec un lecteur de CD classique car d'une part le signal tatoué reste encodé au format CD-audio standard, et d'autre part le tatouage est inaudible et peut très bien ne pas être exploité.

Notons que l'utilisateur fournisseur est ici le producteur du CD (au sens aussi bien technique que commercial). La phase de tatouage compatible avec le procédé de séparation intervient juste avant ou juste après le mixage, selon la configuration de séparation envisagée (cf. sous-section précédente). On peut envisager la mise au point de logiciels de tatouage dont l'utilisation soit quasi-transparente pour l'ingénieur du son chargé du mixage, voire d'intégrer directement le procédé de tatouage dans un logiciel de mixage.

1.3 Plan de la thèse

Nous nous attachons au cours de ce manuscrit à répondre à quatre questions concernant la séparation de sources informée. Quand, c'est-à-dire dans quel cas de figure s'applique la séparation de sources informée ? Pourquoi, c'est-à-dire dans quel but cette méthode a-t-elle été mise au point ? Où, c'est-à-dire dans quel domaine de représentation s'applique-t-elle ? Et enfin comment, c'est-à-dire par l'utilisation de quelles méthodes

4. Le problème de copyright devra vraisemblablement être géré au niveau des différentes pistes utilisée dans le mixage.

et de quels outils la séparation informée est-elle mise en œuvre ? Nous avons déjà fournit des éléments de réponse aux deux premières questions : la séparation de source informée est utilisée lorsque l'on dispose des signaux sources avant mixage et éventuellement d'un seul exemplaire de leur mélange (ou deux canaux redondants dans le cas d'un mélange stéréo), afin de pouvoir manipuler les signaux sources estimés individuellement pour une écoute active de la musique (modification du volume, du timbre, de la spatialisation, ou tout autre type de remixage).

L'ensemble du manuscrit s'articule de la manière suivante.
Nous dressons dans un premier temps un état-de-l'art des techniques existantes de séparation de sources (au Chapitre 2) et de tatouage numérique (au Chapitre 3) se rapprochant le plus des méthodes utilisées au cours de ces travaux de thèse.
Nous étudions ensuite une première configuration de séparation de sources par codage des signaux sources, à partir d'un mélange linéaire instantané monophonique, puis stéréophonique. Une étude théorique est tout d'abord proposée (au Chapitre 4), suivie des détails d'implémentation en tant que telle (au Chapitre 5).
Nous abordons ensuite (au Chapitre 6) une technique de séparation de sources plus *classique*, ne recourant pas au codage des signaux sources, mais plutôt basée sur leur parcimonie, et l'exploitation de la multi-dimensionnalité du mélange (le mélange stéréophonique sera traité plus spécifiquement).
Une troisième approche combinant les atouts des deux précédents systèmes est ensuite détaillée au Chapitre 7.
Un bilan des contributions apportées et des résultats obtenus, ainsi qu'une présentation des perspectives de travail futur sont finalement présentés au dernier chapitre de ce manuscrit.

Chapitre 2

État de l'art de la séparation de sources

Comme nous l'avons déjà mentionné en introduction, la séparation de sources informée opère une jonction entre deux grands domaines du traitement du signal que sont la séparation de sources et le tatouage numérique des signaux. Ces deux domaines étant particulièrement vastes, nous restreignons volontairement l'étude bibliographique aux aspects de ces spécialités qui nous concernent directement. Cette étude bibliographique se scinde donc en deux grandes parties, l'une axée sur la séparation de sources, développée au cours de ce chapitre, et l'autre axée sur le tatouage numérique, présentée dans le Chapitre 3. Dans ce chapitre, nous présentons le problème de la séparation de sources de manière générale, les différents types de mélanges sonores, nous mentionnons les principes de l'analyse en composantes indépendantes (ACI) et de l'analyse de scène auditive computationnelle, et enfin nous considérons tout particulièrement le cas des méthodes de séparation basées sur la parcimonie, plus proches de notre domaine d'étude.

2.1 Présentation générale du problème de la séparation de sources

La séparation de sources est un domaine du traitement du signal qui a connu un essor majeur depuis une vingtaine d'années. Elle consiste à retrouver des signaux sources originaux, non observés, à partir de l'observation d'un ou plusieurs mélanges de ces signaux sources. L'exemple classique est celui de la "cocktail party" où plusieurs personnes parlent en même temps dans une même pièce et qu'une personne essaie de suivre dans ce flot d'informations le signal de parole prononcé par un locuteur particulier [Bronkhorst, 2000]. Le cerveau humain est capable de traiter avec une efficacité surprenante ce type de séparation de sources. Cependant, c'est un problème qui reste extrêmement complexe en traitement numérique du signal. L'estimation de signaux superposés est réalisée en tenant compte de la structure du mélange (voir Section 2.2) et en faisant des hypothèses sur les sources.

D'un point de vue mathématique, le problème de la séparation de sources se formalise de la façon suivante. Considérons un vecteur $\mathbf{x}[n]$ de J observations de mélanges $x_j[n], j \in [1, J]$ de I sources $s_i[n], i \in [1, I]$, $\mathbf{s}[n]$ étant le vecteur des signaux sources. $\mathbf{x}[n]$ est obtenu à partir d'une application inconnue $\mathcal{A}$ de $\mathbb{R}^I$ dans $\mathbb{R}^J$, telle que

$$\mathbf{x}[n] = \mathcal{A}(\mathbf{s}[n]) \tag{2.1}$$

où n désigne un indice de réalisation qui peut être un échantillon dans le cas discret. En fonction du nombre d'observations J par rapport au nombre de sources I, trois configurations de mélanges existent [Comon and Jutten, 2007b] :

- s'il y a plus d'observations que de sources ($J > I$), le mélange est dit *sur-déterminé*,
- s'il y a autant d'observations que de sources ($I = J$), le mélange est dit *déterminé*,
- s'il y a plus de sources que d'observations ($J < I$), le mélange est dit *sous-déterminé*.

Enfin, l'application $\mathcal{A}$ peut être linéaire ou non, à mémoire ou non, et stationnaire ou non (*i.e.* invariante dans le temps ou non) [Comon and Jutten, 2007b]. Dans cette thèse, nous ne considérons que le cas de mélanges linéaires stationnaires. Un mélange linéaire sans mémoire est aussi appelé mélange linéaire instantané, et un mélange linéaire à mémoire est appelé mélange linéaire convolutif. Nous décrivons plus en détails dans la section suivante les principaux types de mélanges utilisés dans les applications audio. Il existe plusieurs types d'approches dans la séparation de sources plus ou moins adaptées aux différents types de mélanges. Parmi elles, on peut citer la séparation aveugle, l'analyse de scènes auditives computationnelle, la séparation basée sur des modèles adaptés aux signaux ou non, la séparation basée sur la parcimonie des sources, ainsi que les multiples combinaisons possibles de ces méthodes [Comon and Jutten, 2007b] [Comon and Jutten, 2007a]. Nous décrirons les principales approches à partir de la Section 2.3.

2.2 Les différents types de mélanges audio

Nous introduisons dans cette section les trois grands types de mélanges audio linéaires, du plus simpliste au plus réaliste. Les différents types de mélanges se distinguent selon deux principaux critères : la prise en compte ou non du délai mis par le son pour parvenir d'une source sonore à un capteur, et les trajets empruntés par l'onde sonore, soit directs soit tenant compte de la réverbération, c'est-à-dire des multiples réflexions subies par les ondes sonores sur les différents éléments de l'environnement acoustique (parois, obstacles,...).

2.2.1 Le mélange linéaire instantané

La configuration de mélange la plus simple est le mélange linéaire instantané dans lequel l'hypothèse est faite que les signaux sources arrivent en même temps sur tous les

capteurs mais avec des intensités différentes, et ce quelles que soient les positions des sources par rapport aux capteurs. Les J observations du mélange s'expriment alors en fonction des I signaux sources de la façon suivante :

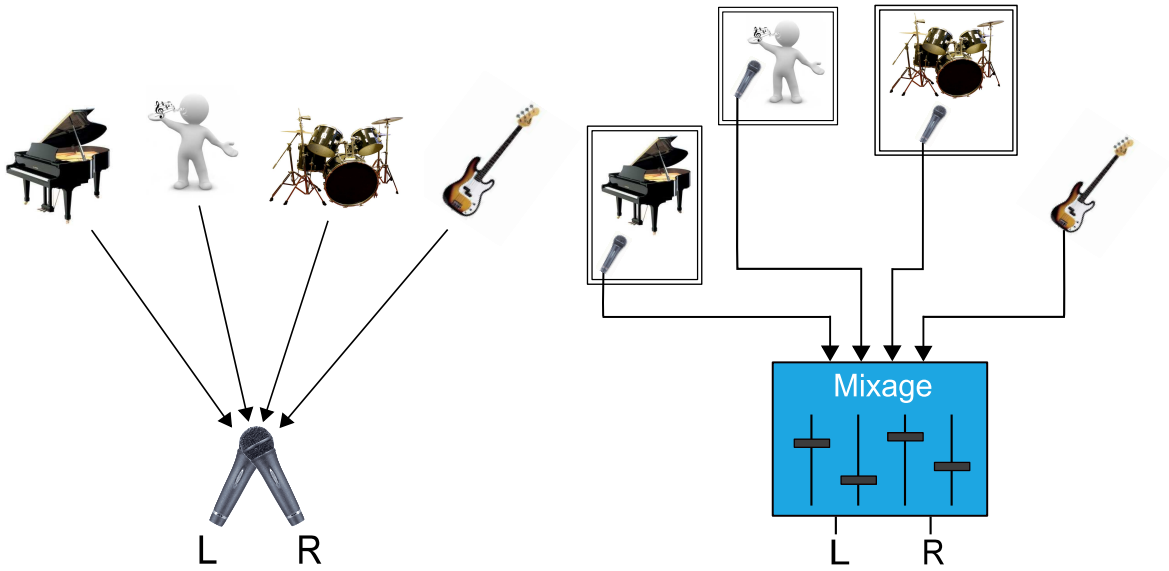
$$x_j(t) = \sum_{i=1}^{I} a_{ji} s_i(t), \quad j = 1, ..., J. \tag{2.2}$$

ou encore pour l'ensemble des observations et des sources, selon la formulation matricielle suivante :

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_J(t) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1I} \\ a_{21} & a_{22} & \cdots & a_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ a_{J1} & a_{J2} & \cdots & a_{JI} \end{pmatrix} \cdot \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_I(t) \end{pmatrix} \tag{2.3}$$

noté de façon plus compacte $\mathbf{x} = \mathbf{A}\mathbf{s}$. La matrice $\mathbf{A}$ est définie par $J \times I$ coefficients a_{ji} constants. La séparation de sources à partir de mélanges instantanés passe d'abord par l'identification, pour chaque source s_i, des facteurs d'atténuation a_{ji}.

Les Figures 2.1a et 2.1b donnent un exemple de réalisation pratique d'un mélange instantané de trois instruments et un chanteur, soit en prise son par deux microphones positionnés en Y, soit en injection directe dans la table de mixage.



(a) Configuration stéréo XY.

(b) Injection directe sur la table de mixage.

FIGURE 2.1 – Deux configurations de mélanges linéaires instantanés.

Notons que pour que les différents signaux parviennent en même temps sur les deux microphones, ils faut que ces derniers soient aussi proches que possible[1]. Dans

1. Les signaux atteignent les microphones avec un retard, mais ce retard est identique pour les deux microphones.

la réalité, la Figure 2.1a ne correspond pas à un mélange parfaitement instantané, en raison de la géométrie des deux microphones, au contraire de la configuration présentée Figure 2.1b. Cette dernière configuration offre de plus l'avantage de contrôler bien plus facilement les coefficients de mélange a_{ji}.

2.2.2 Le mélange anéchoïque

Une extension du mélange linéaire instantané est le mélange dit *anéchoïque* pour lequel les temps d'arrivée des signaux sur les différents capteurs sont retardés d'un délai qui dépend de la position de chaque source par rapport à chaque capteur. La prise en compte des délais d'une onde sonore pour parvenir aux différents capteurs apporte plus de réalisme au processus de mélange, le temps de trajet d'une onde sonore augmentant avec sa distance à un capteur. Pour chaque source, seul le chemin direct de l'onde sonore à un capteur est ici considéré, ceci revenant à considérer que le mélange est réalisé en milieu ouvert (sans réverbération du son sur les parois de la pièce où a lieu l'enregistrement). Le mélange de sources introduit à l'équation (2.1) se note alors :

$$x_j(t) = \sum_{i=1}^{I} a_{ji} s_i(t - \delta_{ji}), \quad j = 1, ..., J. \tag{2.4}$$

Les coefficients d'atténuation a_{ji} et de retard δ_{ji} sont fonction de la position relative entre les sources et les capteurs. La notation matricielle du mélange anéchoïque est la suivante :

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_J(t) \end{pmatrix} = \begin{pmatrix} a_{11}\delta(t-\delta_{11}) & a_{12}\delta(t-\delta_{12}) & \cdots & a_{1I}\delta(t-\delta_{1I}) \\ a_{21}\delta(t-\delta_{21}) & a_{22}\delta(t-\delta_{22}) & \cdots & a_{2I}\delta(t-\delta_{2I}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{J1}\delta(t-\delta_{J1}) & a_{J2}\delta(t-\delta_{J2}) & \cdots & a_{JI}\delta(t-\delta_{JI}) \end{pmatrix} * \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_I(t) \end{pmatrix} \tag{2.5}$$

La matrice $\mathbf{A}$ est ici définie par $2 \times J \times I$ constantes pour les atténuations et les retards, et $*$ dénote l'opérateur de convolution.

La Figure 2.2 donne un exemple de réalisation pratique d'une configuration stéréo anéchoïque : on utilise deux microphones espacés de quelques centimètres ou dizaines de centimètres.

2.2.3 Le mélange convolutif

La configuration de mélange linéaire la plus complexe que nous considérons dans cette section, et de fait, la plus proche de conditions d'enregistrements réelles en milieu fermé, est le mélange convolutif. Le mélange convolutif peut être vu comme une extension du mélange anéchoïque dans lequel on considère plusieurs chemins entre les sources et les récepteurs. Dans ce type de mélange, on ne considère pas seulement les délais

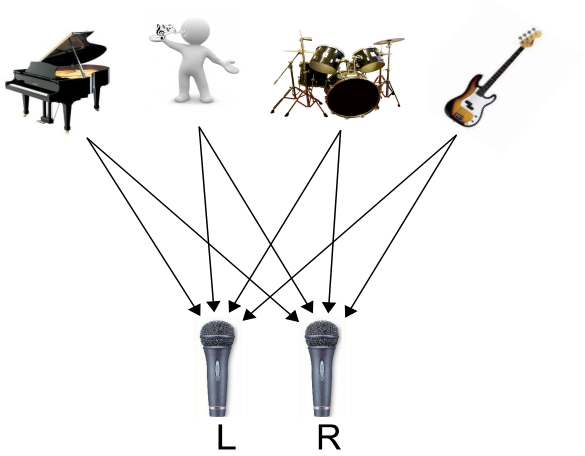


FIGURE 2.2 – Configuration de mélange anéchoïque.

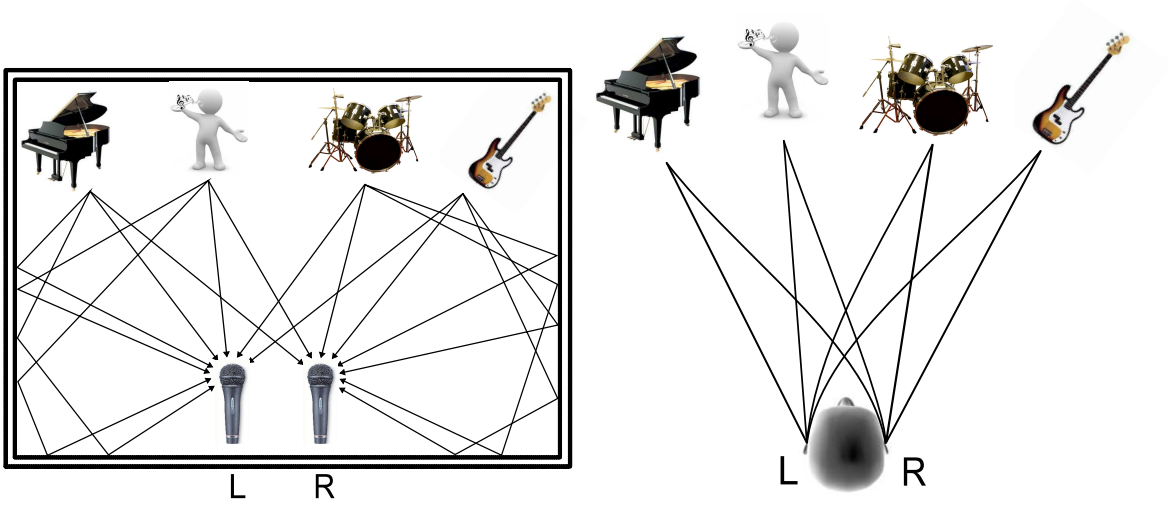
de transmission entre les sources et les capteurs, mais aussi les multiples réflexions des signaux sources sur les parois de la pièce d'enregistrement, phénomène appelé réverbération [2]. Les différents chemins empruntés par le signal entre le point d'émission de la source et les capteurs sont fonctions de la géométrie de la pièce où a lieu l'enregistrement. La matrice de mélange $\mathbf{A}$ peut être vue comme la matrice des filtres définis par l'environnement dans lequel sont enregistrés les mélanges. Les observations sont ici aussi générées par la convolution des sources avec la matrice $\mathbf{A}$, mais celle-ci est une matrice de filtres. En d'autres termes, chaque observation $x_j(t)$ est une combinaison linéaire des sources $s_i(t)$ filtrées par des filtres $h_{ji}(t)$ que nous considérons ici comme étant à réponse impulsionnelle finie :

$$x_j(t) = \sum_{i=1}^{I} h_{ji}(t) * s_i(t) = \sum_{i=1}^{I} \sum_{k=1}^{K_{ji}} a_{ji}^k s_i(t - \delta_{ji}^k), \quad j = 1, ..., J. \tag{2.6}$$

avec K_{ji} le nombre de chemins que peut prendre le signal entre la source s_i et le capteur j.

Les Figures 2.3a et 2.3b donnent un aperçu schématique de la réalisation pratique d'un mélange convolutif et d'un mélange binaural. Ce dernier est un cas particulier de mélange convolutif en milieu ouvert (pas de réverbération sur les parois de la pièce d'enregistrement) visant à simuler le système auditif humain. Le nombre de microphones utilisé est donc de deux (un pour chaque oreille), et ils sont placés de part et d'autre de la tête d'un mannequin qui simule l'influence de la tête d'un auditeur. Pour atteindre un bon mimétisme, les oreilles doivent être une réplique exacte du pavillon de l'oreille humaine, autant pour la forme que pour la consistance (voir Figure 2.4a). Les enregistrements binauraux sont destinés à être écoutés à l'aide d'un casque, voire d'écouteurs intra-auriculaires, de telle sorte que le signal audio binaural ne soit pas modifié une seconde fois par le pavillon de l'oreille de l'auditeur (il l'a déjà été par le procédé d'enregistrement binaural).

2. Réciproquement, on peut définir le mélange anéchoïque comme un mélange convolutif en milieu ouvert, dans lequel aucune réverbération n'est considérée.



(a) Mélange convolutif. (b) Mélange binaural.

FIGURE 2.3 – Deux configurations de mélanges linéaires convolutifs.

Dans un mélange binaural, une source $s(t)$ à l'azimut θ par rapport à un point fixé au centre des deux oreilles (voir Figure 2.5) est filtrée par un filtre binaural de réponse impulsionnelle $HRIR_L(\theta, t)$ sur le canal gauche (respectivement $HRIR_R(\theta, t)$ sur le canal droit), puis, sur chaque voie, les signaux sources filtrés sont sommés pour former le signal de mélange $x_L(t)$ (gauche) (respectivement $x_R(t)$ (droit)). Les filtres HRIR (*Head Related Impulse Response*) sont caractéristiques de la morphologie d'une tête humaine, différents pour chaque oreille, et fonction de la position de la source[3]. On a ainsi,

$$\mathbf{x}_L(t) = \mathbf{s}(t) * HRIR_L(\theta, t), \quad \text{et} \quad \mathbf{x}_R(t) = \mathbf{s}(t) * HRIR_R(\theta, t), \tag{2.7}$$

2.3 Séparation aveugle et Analyse en Composantes Indépendantes

La séparation *aveugle* est la forme la plus générale du problème de séparation de sources, dans laquelle aucune information sur les signaux sources ni sur les paramètres de mélanges n'est connue a priori, ce qui lui confère un intérêt certain dans des domaines aussi variés que le traitement de la parole ou l'analyse d'images médicales. En réalité la formulation *aveugle* est un peu abusive puisque sans hypothèse du tout, le problème de

3. Notons que dans un modèle encore plus complexe, les HRIR peuvent être fonctions de l'élévation des sources par rapport au plan horizontal des oreilles, ainsi que de l'éloignement des sources aux capteurs. Dans le cas où l'ensemble des sources sont supposées situées au même niveau que l'auditeur, et à une distance à la fois suffisamment faible pour négliger l'atténuation des ondes sonores dans l'air (fonction de la fréquence), et suffisamment importante pour que ces même ondes soient supposées planes lorsqu'elles atteignent les oreilles, une approximation courante consiste à dire que les HRIR ne sont fonction que de l'azimut.

(a) Tête seule. (b) Tête et torse.

FIGURE 2.4 – Modèle de mannequin utilisé pour l'enregistrement de mélanges binauraux.

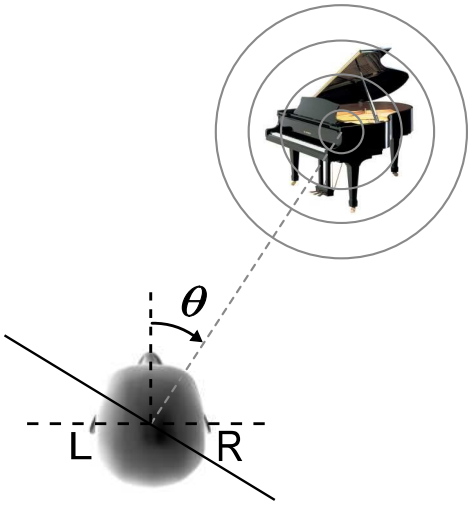


FIGURE 2.5 – Une source positionnée dans le plan horizontal à l'azimut θ propageant des ondes acoustiques vers la tête.

la séparation de sources décrit en 2.1 est insoluble. Darmois établit ainsi dans [Darmois, 1953] que le cas d'un mélange linéaire instantané ne peut être résolu si les sources sont gaussiennes et indépendantes identiquement distribuées (iid). On est donc amené à faire un certain nombre d'hypothèses sur les sources et/ou les mélanges. On peut par exemple supposer que les sources sont statistiquement indépendantes et que les mélanges sont linéaires, convolutifs ou instantanés. Dans le cas d'un mélange instantané, par exemple, le modèle est défini par l'équation : $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ où $\mathbf{x}$ est le vecteur des observations du mélange de dimension J, $\mathbf{s}$ est le vecteur des sources de dimension I et $\mathbf{A}$ est la matrice de mélange de dimension $J \times I$ à coefficients constants (voir Section 2.2.1). La séparation de sources consiste (si $J \geq I$) à estimer les paramètres de la matrice de séparation $\mathbf{B}$ de dimension $I \times J$ telle que ses sorties $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \simeq \mathbf{s}(t)$ soient des estimées des sources originales $\mathbf{s}$. Dans le cas où $I = J$, et $\mathbf{A}$ est inversible, l'estimation

optimale de $\mathbf{B}$ correspond à $\mathbf{A}^{-1}$. Si $J > I$, $\mathbf{B}$ est la matrice pseudo-inverse de $\mathbf{A}$ (au sens de Moore-Penrose). Si $I > J$, $\mathbf{A}$ est non-inversible, et le problème est traité spécifiquement, par exemple par des méthodes du type de celles abordées à la Section 2.5.

La première configuration envisagée en séparation de sources est celle de sources *iid* (et non-gaussiennes de manière à satisfaire le critère établi par Darmois) et a conduit aux méthodes d'analyses de données connues sous la dénomination d'analyse en composantes indépendantes (ACI, ou ICA pour Independent Component Analysis en anglais) [Comon, 1991] [Comon and Jutten, 2010] [Hyvärinen et al., 2001]. L'estimation de la matrice de séparation (cas $J \geq I$) se fait alors en maximisant un critère basé sur l'indépendance statistique des sorties du système de séparation. Une seconde catégorie de méthodes supposent cette fois les sources non-*iid* (et donc potentiellement gaussiennes). Alors que seule l'indépendance entre les signaux est originellement utilisée dans les méthodes de type ICA, les méthodes fondées sur l'hypothèse de sources non-*iid* s'appuient sur les propriétés temporelles des sources, comme l'hypothèse de sources colorées temporellement (identiquement distribuées mais non indépendantes temporellement) et potentiellement gaussiennes [Belouchrani and Abed-Meraim, 1993]. Cette méthode ainsi que les approches exploitant les propriétés de non-stationnarité des sources [Pham and Cardoso, 2001] exploitent des statistiques d'ordre deux ce qui les rend plus simples à mettre en oeuvre que les techniques dérivées de l'ACI, et sont de plus capables de traiter des sources gaussiennes. Plus généralement, lorsque des critères statistiques tels que la non gaussiannité des sources ne peuvent être appliqués, on peut avoir recours à d'autres hypothèses comme la non stationnarité des signaux sources, ou leur coloration spectrale, par exemple.

Les difficultés rencontrées en séparation de sources sont multiples, en fonction de la méthode de séparation utilisée, mais aussi du type de mélange, et du nombre de sources. Les techniques de séparation aveugles de sources de type ICA se heurtent, même dans le cas le moins complexe d'un mélange linéaire instantané, à des problèmes de séparabilité et d'indéterminations. Ainsi, même si $\mathbf{A}$ est (pseudo-) inversible ($I \leq J$), lors de l'identification de $\mathbf{B} = \mathbf{A}^{-1}$, les vecteurs colonnes de $\mathbf{A}^{-1}$ ne peuvent être retrouvés qu'à une permutation et un facteur d'échelle près (la maximisation de l'indépendance des signaux estimés ne dépend pas de leur ordre ou de leur gain) [Comon and Jutten, 2010]. Les estimées des signaux sources s'expriment alors sous la forme

$$y_i[n] = k_i s_{\sigma(i)}[n], \quad i = 1, ..., I. \tag{2.8}$$

où σ est une permutation de $\{1, ..., I\}$ et k_i un gain : les sources ne peuvent donc être estimées qu'à une permutation et un facteur d'échelle près. Des solutions à ce problème, qui reste cependant un problème délicat, ont été proposées dans la littérature (voir [O'Grady et al., 2005] pour un exemple de revue du problème).

Pour ce qui est des mélanges convolutifs, il est possible de grandement simplifier le problème de la séparation de sources en passant dans le domaine temps-fréquence. En

effet, résoudre le problème de la séparation de sources à partir de mélanges convolutifs dans le domaine temporel consiste soit à estimer les réponses impulsionnelles des filtres définis dans l'équation (2.6), soit à estimer directement les filtres de démélange correspondant, ce qui s'avère dans la plupart des cas extrêmement complexe. L'intérêt du domaine fréquentiel tient au fait qu'une convolution dans le domaine temporel devient, dans le domaine fréquentiel une simple multiplication scalaire pour chaque source. Si le processus de mélange est stationnaire (*i.e.* n'évolue pas au cours du temps), l'équation du mélange convolutif définie en (2.6), devient dans le domaine fréquentiel :

$$\mathbf{X}(f,t) = \mathbf{H}(f)\mathbf{S}(f,t) \tag{2.9}$$

où $\mathbf{X}(f,t)$ est la transformée de Fourier à court terme de $\mathbf{x}(t)$. La matrice de filtre $\mathbf{H}$ étant stationnaire, sa transformée de Fourier ne dépend que de la fréquence f. Ainsi, pour chaque fréquence f, l'équation (2.9) revient à un problème instantané. Le passage du domaine temporel au domaine fréquentiel permet de transformer un problème de séparation de sources à partir d'un mélange convolutif en N_f problèmes de séparation de sources à partir de mélanges instantanés, avec N_f le nombre de canaux fréquentiels de la transformée de Fourier. Cependant, si le problème est simplifié par la transformation temps-fréquence, le nombre d'indéterminations de permutations et de facteur d'échelle rencontré pour la séparation de mélange instantané est multiplié par N_f : les colonnes des matrices de démélanges à différentes fréquences peuvent très bien ne pas correspondre à la même source, ce qui peut entraîner une très forte distorsion des signaux estimés. Des méthodes ont été proposées pour résoudre ces indéterminations, comme par exemple celle proposée par Murata [Murata et al., 2001], qui exploite la non-stationnarité de signaux de parole. Les composantes séparées à chaque fréquence sont combinées aux composantes des autres fréquences qui présentent une enveloppe similaire durant une période de temps donnée. Sawada et al. [Sawada et al., 2004] y ajoutent un autre critère : une estimation de la direction d'arrivée (DOA en anglais) de chaque source qui permet d'ordonner les permutations par source.

De manière générale, les méthodes de séparation aveugle de sources de type ICA sont dédiées à la configuration (sur-)déterminée. On est donc ici dans un cadre assez différent de celui de notre procédé. En effet, la première caractéristique de nos travaux est qu'ils portent sur la séparation de sources dans le cas sous-déterminé avec seulement une ou deux observations du mélange au maximum (deux voies stéréo pour l'application CD audio), qui sont de plus significativement corrélées. On ne détaille donc pas plus ce pan de la bibliographie. Le lecteur intéressé par ce sujet peut se référer par exemple à l'ouvrage de synthèse sur la séparation de sources [Comon and Jutten, 2007b] et [Comon and Jutten, 2007a].

2.4 Séparation et Analyse de Scène Auditive Computationnelle

Une autre grande famille d'approches du problème de la séparation de sources est l'analyse de scènes auditives computationnelle, (CASA pour Computational Auditory Scene Analysis en anglais). Il s'agit d'une technique de séparation visant à simuler numériquement les mécanismes du système auditif humain pour séparer les sources de la même façon que le fait notre oreille, tout du moins de façon théorique. Un sujet est capable naturellement de décomposer une scène auditive complexe (un mélange), pour ensuite séparer les différents composants de cette scène, chaque composant correspondant à une source. Ce processus d'analyse de scène auditive se scinde en deux grandes étapes, la segmentation de la scène auditive en traits acoustiques élémentaires (onset, offset, harmoniques, modulation d'amplitude ou de fréquence), puis une deuxième étape consistant soit en le regroupement perceptif de ces traits en flux, chaque flux correspondant à une source, soit en la scission des traits entre plusieurs sources. On peut citer sur ce sujet les études de Ellis [Ellis, 1996] [Ellis, 1999]. Godsmark et Brown ajoutent la notion de timbre des sons étudiés [Godsmark and Brown, 1999] de même que Kinoshita [Kinoshita et al., 1999]. L'analyse de scènes auditives computationnelle se heurte à des difficultés notables quand les sources se situent dans des zones temps-fréquence voisines, ce qui est très souvent le cas en pratique pour de nombreux instruments. C'est pour cette raison qu'elle permet seulement d'expliquer un mélange du point de vue perceptif mais pas de séparer effectivement les sources. Nous verrons que nous serons confrontés à ce genre de difficultés dans nos travaux.

2.5 Séparation de source basée sur la parcimonie

Dans le cas où le mélange est *sous-déterminé*, $\mathbf{A}$ n'est pas inversible. Dans ce cas de figure, la séparation de sources se scinde généralement en deux problèmes distincts : l'identification du mélange $\mathbf{A}$, et l'estimation des sources à proprement parler. Dans la configuration sous-déterminée, les méthodes classiques de séparation aveugle ne parviennent pas à fournir de solution satisfaisante. C'est pourquoi un intérêt croissant est porté à des approches de séparation de sources s'appuyant sur une hypothèse de parcimonie des signaux sources [Gribonval and Lesage, 2006]. Les méthodes de séparation de sources basées sur la parcimonie offrent en effet pour de nombreux signaux *naturels*, en particulier pour les signaux audio, l'avantage de concentrer l'énergie des signaux en un faible nombre de coefficients. Il est alors possible de ramener localement une configuration sous-déterminée dans le domaine initial des signaux à une configuration déterminée, voire sur-déterminée, dans le domaine transformé en limitant le risque de superposition des différentes sources.

2.5.1 Principe de la parcimonie des sources audio

2.5.1.1 Définition de la parcimonie

Un signal est dit parcimonieux, dans un certain domaine de représentation (qui est généralement différent du domaine initial de représentation du signal), si la majorité de l'énergie de ce signal est concentrée sur un faible nombre de ses coefficients de décomposition, ou, autrement dit, si dans ce domaine de représentation la plupart de ses coefficients sont proches de zéro. Un signal parcimonieux $s(t)$ peut donc se décomposer sur une certaine famille de signaux élémentaires φ_k, appelés *atomes*, selon l'équation

$$s(t) = \sum_{k=1}^{K} c(k)\varphi_k(t) \tag{2.10}$$

où la plupart des coefficients $c(k)$ sont négligeables. La famille des atomes φ_k, appelée *dictionnaire*, peut être de différents types, et plus ou moins adaptée au signal. Ainsi pour le cas du traitement de signaux audio, la transformée de Fourier à court terme (TFCT) est souvent utilisée. Chaque coefficient $c(k)$ correspond alors à la TFCT du signal s à une position temps-fréquence $k = (f, t)$ Une autre façon de définir un signal parcimonieux revient à dire qu'un faible nombre de coefficients (au regard du nombre total de coefficients de la décomposition de ce signal) suffit à donner une (très) bonne approximation du signal.

D'un point de vue statistique, la densité de probabilité d'un signal parcimonieux présente donc un pic très significatif autour de zéro avec des queues de distribution décroissant très rapidement. Une distribution parcimonieuse utilisée dans la littérature [Zibulevsky et al., 2001] est une distribution de type Laplacienne, plus parcimonieuse qu'une distribution gaussienne comme le montre la Figure 2.6. C'est pour cette raison que l'hypothèse de parcimonie des signaux est parfois formulée comme une hypothèse de distribution Laplacienne (*Laplacian prior*).

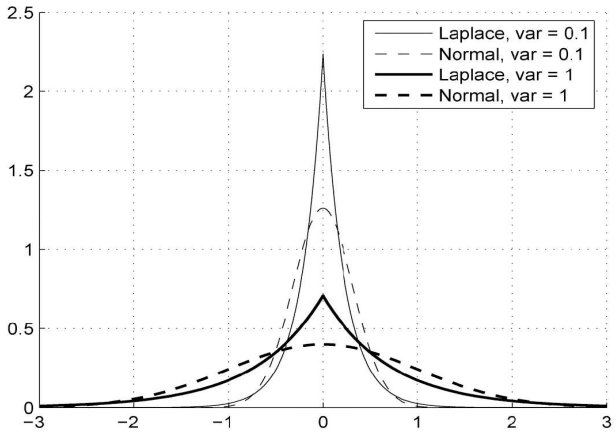


FIGURE 2.6 – Comparaison des distributions de Laplace et de Gauss pour des variances équivalentes.

Un des avantages évident d'une représentation parcimonieuse des signaux est que la probabilité que deux sources ou plus soient simultanément "actives" est faible. En effet, si les signaux sources ont, dans un domaine donné, une représentation parcimonieuse, la plupart de l'énergie de chaque signal dans le domaine de représentation en question est concentrée dans un faible nombre de coefficients, réduisant ainsi le risque de superposition de sources. Ainsi, contrairement aux hypothèses classiques d'indépendance des sources utilisées en séparation de sources aveugle, l'idée clé de la séparation de sources basée sur la parcimonie est essentiellement d'ordre géométrique. Prenons l'exemple d'un mélange linéaire instantané défini (cf Section 2.2.1) par

$$\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n] \tag{2.11}$$

L'hypothèse de parcimonie revient ici à considérer que pour chaque réalisation n, une source est significativement plus active que les autres. Notons Σ_i l'ensemble des réalisations où la source s_i est plus active que les autres sources :
$\Sigma_i = \{n \mid |s_i| \gg |s_k|, \forall k \in \{1, I\} \setminus \{i\}\}$. Donc, pour tout $n \in \Sigma_i$, (2.1) se réduit à

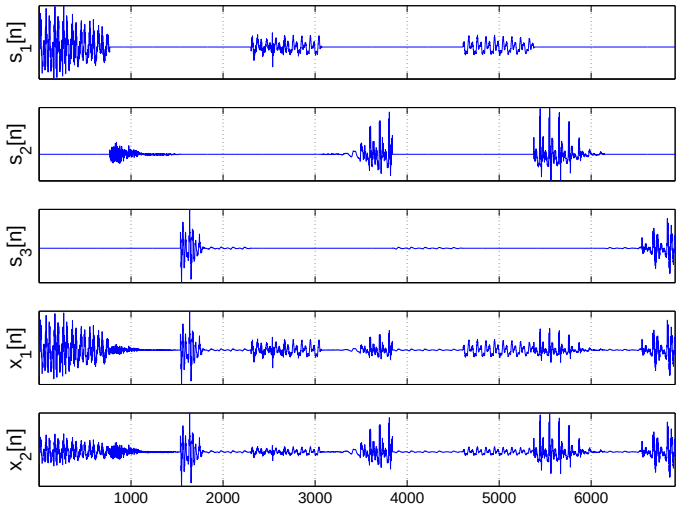
$$\mathbf{x}[n] \approx \mathbf{A}_i s_i[n], \quad n \in \Sigma_i \tag{2.12}$$

où $\mathbf{A}_i$ est la i-ième colonne de la matrice de mélange $\mathbf{A}$. De fait, l'ensemble des points $\{\mathbf{x}[n] \mid n \in \Sigma_i\}$ est aligné sur la droite coupant l'origine et de vecteur directeur $\mathbf{A}_i$. Cette propriété est facilement visualisable grâce à un *scatter plot* (ou diagramme de dispersion) du signal de mélange. En dimension $J = 2$, c'est à dire pour un mélange stéréophonique, cela revient à représenter l'ensemble des couples de points $(x_1[n], x_2[n])$. Nous donnons dans les sous-sections suivantes des exemples d'une telle parcimonie dans les domaines temporel puis fréquentiel.

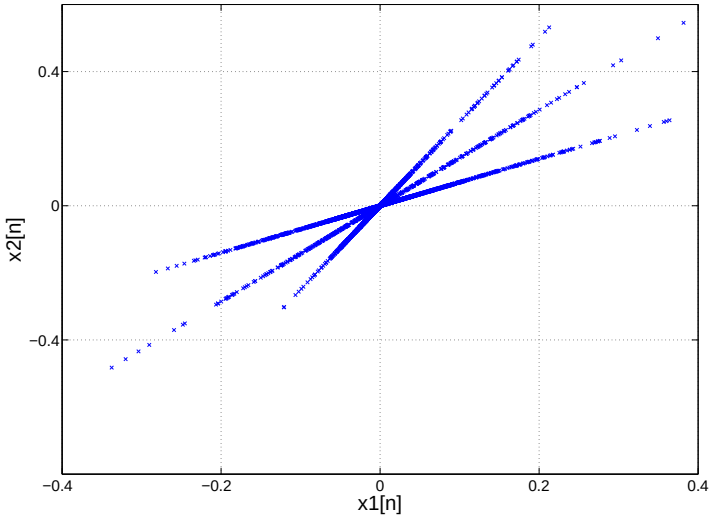
2.5.1.2 La parcimonie temporelle

La figure 2.7a présente un exemple de mélange instantané stéréophonique de trois sources de parole à supports temporels disjoints. Le diagramme de dispersion obtenu à partir des deux voies du mélange est donné figure 2.7b. Les trois droites correspondant aux trois colonnes de la matrice de mélange $\mathbf{A}$ apparaissent clairement. Géométriquement, l'identification des colonnes de la matrice de mélange revient à isoler chacune de ces droites. Une fois la matrice de mélange (inversible) estimée, l'estimation des signaux sources en elle-même est immédiate. L'estimation de la matrice de mélange peut être réalisée, à partir du diagramme de dispersion, à l'aide d'un algorithme de catégorisation (*clustering*), avec un nombre de clusters de points qui soit égal au nombre de sources I [Zibulevsky et al., 2002]. Dans [Hulle, 1999] Van Hulle utilise un diagramme de dispersion en coordonnées polaires construit à partir des deux observations d'un mélange stéréo. Des amas de points dans les directions des différents vecteurs colonnes de la matrice de mélange sont alors mis en évidence. Cette méthode est directement utilisée dans le domaine temporel pour la séparation de mélanges instantanés de signaux de parole. On peut remarquer que plus les amas de points sont distincts, plus la séparation est aisée. Moins les signaux sources sont à supports temporels disjoints, moins les

directions des colonnes de $\mathbf{A}$ sont clairement définies, et moins la détermination de la matrice de mélange est aisée. Or, dans la réalité, les signaux (en particulier les signaux audio) ne sont que très rarement à supports temporels disjoints, et l'estimation de la matrice de mélange doit être effectuée dans un autre domaine de représentation.



(a) Signaux sources temporels et mélange stéréo correspondant.



(b) Diagrammes de dispersion (x_1, x_2) en temps.

FIGURE 2.7 – Illustration de la séparation de sources basée sur la parcimonie. Exemple de signaux à supports temporels disjoints.

2.5.1.3 La parcimonie dans un domaine transformé

Le domaine temps-fréquence fournit une décomposition de la majorité des signaux, notamment les signaux audio, nettement plus parcimonieuse que le domaine temporel [Rickard and Yilmaz, 2002]. Il en résulte donc que les colonnes de la matrice de mé-

lange sont généralement beaucoup plus facilement et précisément identifiables dans le domaine temps-fréquence que dans le domaine temporel, comme l'attestent les figures 2.8a et 2.8b. Un diagramme de dispersion dans le domaine temporel fournit ici un amas de points dans lequel il est impossible d'isoler les différentes directions correspondant aux vecteurs colonnes de la matrice de mélange, alors que ces trois directions sont clairement différenciées dans le domaine de projection temps-fréquence obtenu après transformation des signaux de mélange par MDCT (Modified Discrete Cosine Transform dont nous reparlerons en détails dans la suite de ce manuscrit). Un des avantages essentiels de la séparation de sources basée sur la parcimonie est qu'il est possible, dans un domaine donné, de séparer un grand nombre de sources. Il est en particulier possible de solutionner le cas de mélanges fortement sous-déterminés là où des méthodes de séparation aveugle classiques ne fournissent pas de résultats satisfaisants.

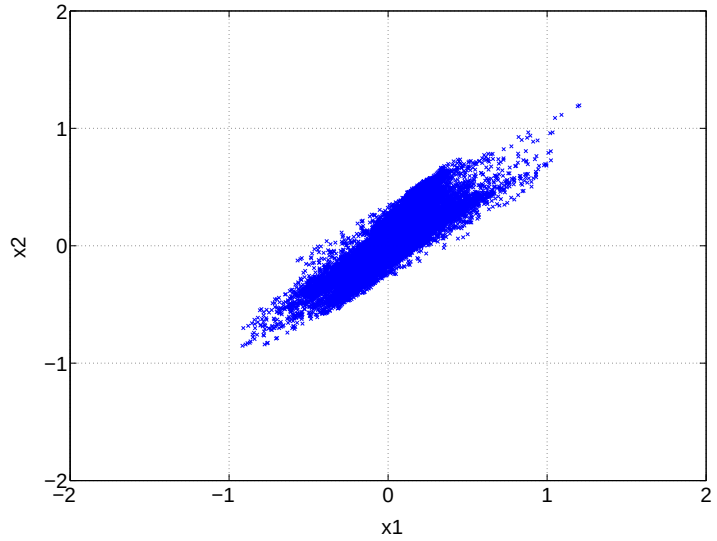
2.5.2 Exploitation de la parcimonie des sources dans la séparation

Les techniques de séparation de sources que nous développons s'appuient à divers degrés sur la parcimonie des signaux sources. Aussi nous détaillons dans la section suivante quelques méthodes de séparation de sources basées sur la parcimonie.

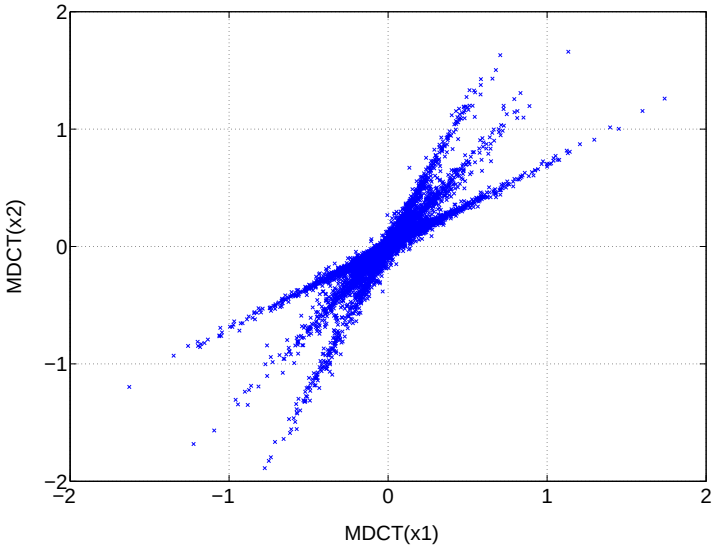
La première étape d'une technique de séparation de sources basée sur la parcimonie consiste en la décomposition des signaux de mélanges dans un domaine de représentation approprié, accroissant la parcimonie des signaux par rapport à leur représentation dans leur domaine d'origine. La Figure 2.9 présente brièvement le processus de séparation exploitant la parcimonie des sources. L'estimation des sources est faite, bien entendu, dans le domaine transformé où les signaux sont plus parcimonieux, et donc se superposent moins.

2.5.2.1 À partir d'un mélange monophonique

Nous avons présenté dans la section précédente une exploitation de la parcimonie dans le cas de mélange stéréophonique (utilisation du diagramme de dispersion pour identifier la matrice de mélange), mais la parcimonie des sources peut également être exploitée dans le cas de mélange monophonique. Dans le cas d'une décomposition parcimonieuse des signaux dans le domaine fréquentiel, divers algorithmes permettent de choisir le type de dictionnaire et la décomposition correspondante la plus vraisemblable. Pour le domaine temporel, on peut citer les travaux de Benaroya qui utilise l'algorithme de Basis Pursuit [Benaroya et al., 2001] ou ceux de Wolfe [Wolfe and Godsill, 2003]. Pour les approches de décomposition parcimonieuse en fréquence, on peut citer les études de Casey et Westner [Casey and Westner, 2000] qui ont introduit l'analyse en sous-espaces indépendants (ISA). Cette méthode consiste à décomposer le spectre d'amplitude à court terme du signal mélange (calculé par transformée de Fourier à court terme (TFCT)) sur une base d'atomes, et ensuite à regrouper les atomes en sous-espaces indépendants, chaque sous-espace étant propre à une source, pour ensuite



(a) Échantillons temporels.



(b) Coefficients temps-frequence après décomposition par MDCT.

FIGURE 2.8 – Diagrammes de dispersion (x1,x2) en temps (a) et en temps-fréquence (b) pour un mélange stéréo instantané de 3 sources de parole à supports temporels non disjoints.

resynthétiser les sources séparément.

Parmi les recherches les plus récentes et les plus abouties sur la séparation de sources dans un domaine parcimonieux à partir d'une seule piste audio (qui représente la première approche développée et détaillée dans le Chapitre 5), on peut également citer celles de Benaroya, Bimbot et Gribonval [Benaroya et al., 2006] qui utilisent des modèles statistiques de la densité spectrale des différentes sources. Cette technique s'apparente à une méthode de séparation supervisée de sources. Il faut noter que les paramètres de ces modèles sont réglés avant la phase de séparation à partir d'exemples de pistes audio des différents instruments à séparer. Ces modèles sont ensuite exploités pour la

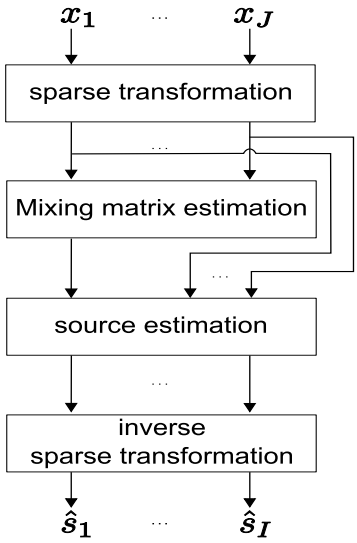


FIGURE 2.9 – Séparation de source après projection sur une base de décomposition parcimonieuse.

séparation proprement dite dans un processus de filtrage de Wiener non-stationnaire. Molla et Hirose [Molla and Hirose, 2003] ont quant à eux travaillé sur une séparation de sources par une décomposition du spectre de Hilbert du mélange en sous-espaces indépendants, la transformée de Hilbert fournissant de meilleurs résultats de discrimination des différentes sources que la transformée de Fourier. La combinaison entre la transformée de Hilbert et l'EMD (empirical mode decomposition) est apparue comme particulièrement adaptée à l'étude de signaux non-stationnaires dont la musique et la parole sont un bon exemple. Enfin, Cho, Shiu et Kuo [Cho et al., 2007] proposent une séparation par décomposition du mélange sur une base d'atomes de Gabor appris pour un instrument particulier, et pour les différentes notes de cet instrument. Par technique de matching pursuit, certains de ces atomes sont retenus puis rassemblés en un sous-espace adapté à la note extraite. Cependant, notons que, pour toutes ces études, les tests sont effectués sur des mélanges artificiels peu réalistes et en conditions très contrôlées par rapport au cas de plusieurs instruments jouant le même morceau de musique.

2.5.2.2 Exploitation de la parcimonie des sources dans la séparation à partir d'un mélange stéréophonique

Cas du mélange instantané

Les méthodes de séparation de sources stéréophoniques basées sur la parcimonie se décomposent en deux phases distinctes, l'estimation de la matrice de mélange, puis l'estimation des signaux sources à proprement parler.

En ce qui concerne l'identification des directions de la matrice de mélange, O'Grady et Pearlmutter utilisent un algorithme dérivé de l'algorithme de catégorisation *k-Means* dans [O'Grady and Pearlmutter, 2004]. Cet algorithme est similaire au *k-Means* original dans lequel les directions des colonnes de la matrice de mélange remplacent les

barycentres, et la distance d'un point à la droite suivant cette direction remplace la distance d'un point à un barycentre.
Des techniques issues du traitement d'images numérique sont également utilisées pour localiser les directions des droites dans les diagrammes de dispersion. Lin et al. [Lin et al., 1997] présentent un algorithme basé sur la transformée de Hough pour identifier les directions des droites. Le diagramme de dispersion est alors considéré comme une image qui est convoluée avec un opérateur de détection de contours, normalisée et convertie en image binaire. Une transformée de Hough est ensuite appliquée sur cette image. Les directions des vecteurs colonnes de la matrice de mélange sont identifiées grâce aux pics apparaissant dans l'espace de la transformée de Hough : à chaque pic correspond une direction propre à chaque source. Bofill et Zibulevski présentent quant à eux une méthode dédiée aux mélanges stéréophoniques [Bofill and Zibulevski, 2001]. Pour déterminer la matrice de mélange, le module et la phase (dans le domaine transformé) sont calculés à chaque point du diagramme de dispersion de coordonnées $(x_1[n], x_2[n])$ par $l_n = \sqrt{x_1^2[n] + x_2^2[n]}$ et $\theta_n = \tan^{-1}(x_2[n]/x_1[n])$. Notons α la différence de phase entre θ_n et une position arbitraire. Une fonction potentielle ϕ de α est ensuite définie par

$$\begin{aligned} \phi(\alpha) &= 1 - \tfrac{\alpha}{\pi/4}, \quad \text{si} \quad |\alpha| < \pi/4 \\ &= 0 \quad \text{sinon} \end{aligned} \tag{2.13}$$

et une fonction potentielle est définie comme

$$\Phi(\theta_k) = \sum_n l_n \phi(\theta_k - \theta_n) \tag{2.14}$$

avec $\theta_k = \pi/2K + k\pi/2K, \quad k = 1 \cdots K$ un des K éléments de la grille discrète d'azimuts équi-répartis entre 0 et π. Les maxima locaux de la fonction Φ correspondent aux directions des vecteurs colonnes de la matrice de mélange $\mathbf{A}$. Une fois la matrice de mélange estimée, les signaux sources sont estimés par une méthode géométrique relativement simple. Estimer les sources revient alors, en chaque point du plan TF, à une minimisation de norme l_1 : c'est à dire trouver la représentation optimale du point $\mathbf{x} = \sum_i \mathbf{a}_i \mathbf{s}_i$ tel que $\sum_i |s_i|$ soit minimum. En dimension 2, cela revient à trouver le couple de vecteurs $(\mathbf{a}_m, \mathbf{a}_n)$ qui offrent le plus court chemin vers le point $\mathbf{x}$, c'est à dire, le couple de vecteurs $(\mathbf{a}_m, \mathbf{a}_n)$ qui encadrent $\mathbf{x}$ comme le montre la figure 2.10 extraite de [Bofill and Zibulevski, 2001]. L'extraction des sources s_m et s_n est alors obtenue par l'opération $\hat{\mathbf{s}} = [\mathbf{a}_m \ \mathbf{a}_n]^{-1} \cdot \mathbf{x}$, et les autres sources, considérées comme non actives, sont mises à zéro. Sur cet exemple, nous voyons clairement que le couple de vecteurs minimisant le chemin de O à $\mathbf{x}$ est $(\mathbf{a}_1, \mathbf{a}_2)$.

Une autre étude de séparation aveugle de sources dans la configuration sous-déterminée est proposée par Aïssa-El-Bey et al. [Aïssa-El-Bey et al., 2007] [Aïssa-El-Bey et al., 2007], qui relaxe également l'hypothèse d'une seule source active à chaque bin du plan TF, mais supposant seulement le nombre de sources actives strictement inférieur au nombre d'observations. Un seuillage du pavage TF du signal de mélange est d'abord réalisé pour sélectionner les bins dits "autosources". Les directions de la matrice de

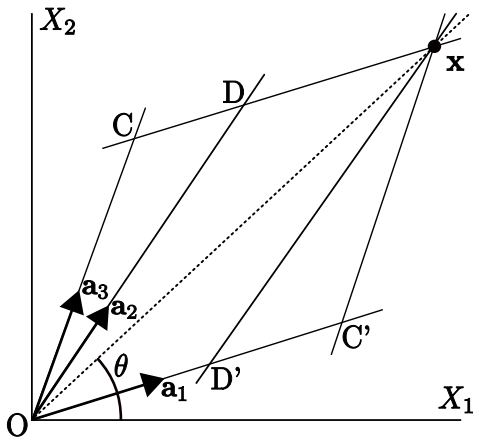


FIGURE 2.10 – Plus court chemin de l'origine au point **x**.

mélange sont obtenues par apprentissage non-supervisé sur les coefficients TF du mélange grâce à un algorithme de type k-means. À chaque point autosource, les index des sources prédominantes sont estimés par projection du mélange sur le sous-espace de la sous-matrice de mélange des sources prédominantes.

Cas du mélange anéchoïque

Jourjine et al. proposent une approche de séparation aveugle de signaux de parole pour des mélanges stéréo anéchoïques dans [Jourjine et al., 2000]. Cette méthode est ensuite développée par Yilmaz et Rickard [Yilmaz and Rickard, 2004]. Les deux observations du mélange sont transformées dans le domaine temps-fréquence par transformée de Fourier à court terme, puis les différences de phase (en anglais *Interaural Phase Difference (IPD)*) et d'amplitude (*Interaural Level Difference (ILD)*) entre les deux voies sont calculées, à chaque position (t, f) du plan temps-fréquence, à partir du ratio des coefficients temps-fréquence des deux voies. Cette méthode présuppose qu'une seule source est prédominante à chaque temps et à chaque fréquence. Le but de cette méthode est précisément de retrouver, en chaque point temps-fréquence, quelle est la source prépondérante. Pour ce faire, les paramètres (IPD,ILD) entre les deux voies sont représentés dans un histogramme 2D pondéré par la densité spectrale de puissance du signal. Du fait de la parcimonie, chaque source crée dans l'histogramme un pic résultant du cumul des valeurs des paramètres obtenus lorsque la source est prépondérante. Ce principe est illustré par la figure 2.11, obtenue pour le mélange de quatre signaux de parole répartis selon les quatre azimuts -30°, +10°, +25°, -15°, à environ un mètre de distance des microphones. Pour procéder à la séparation, un algorithme de catégorisation permet ensuite d'associer chaque point du plan temps-fréquence à une source distincte, en l'associant au pic de l'histogramme le plus proche. Un masque binaire est ensuite généré à partir de cette catégorisation, puis appliqué sur la décomposition temps-fréquence du signal de mélange pour isoler la source correspondant au masque. Finalement, chaque signal source est estimé par transformée temps-fréquence inverse de la décomposition temps-fréquence masquée. Woodruff et al.

proposent dans [Woodruff and Prado, 2007] une méthode de séparation de sources plus dédiée aux signaux de musique, pour des mélanges anéchoïques stéréophoniques. Les paramètres de mélanges sont estimés par une version améliorée de l'approche DUET, introduite originellement pour des mélanges de parole. À chaque trame, le nombre et l'identité des sources présentes sont ensuite estimés grâce à la fréquence fondamentale de chaque source, préalablement estimée, puis l'énergie du mélange est redistribuée entre les sources.

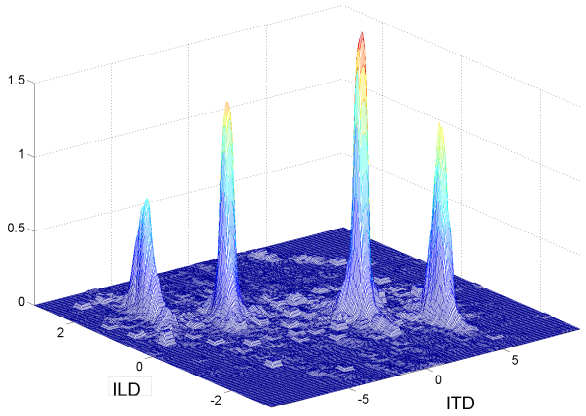


FIGURE 2.11 – Histogramme 2D pondéré en puissance. Chaque pic correspond aux paramètres de mélange d'un signal source.

Les diagrammes de dispersion peuvent également être utilisés dans le cas de mélanges anéchoïques [Bofill, 2002]. Le processus de séparation de sources comporte trois étapes. Dans la première étape, la matrice des atténuations d'amplitudes (ILD) est obtenue par un raisonnement similaire à celui effectué dans le cas de mélanges instantanés dans [Bofill and Zibulevski, 2001]. Les signaux de mélanges sont dans un premier temps décomposés dans le domaine temps-fréquence par transformée de Fourier court terme. Un diagramme de dispersion est construit à partir du module des observations, puis les directions de la matrice de mélange sont obtenues par la même méthode d'estimation de densité de probabilité (méthode d'estimation par noyau) que celle présentée au début de la section 2.5.2.2. Une première approximation de *clusters* des sources est alors obtenue. Chaque point temps-fréquence dont la direction dans le diagramme de dispersion est proche de celle définie par un vecteur colonne (par exemple le $k^{\text{ième}}$) de la matrice des atténuations des amplitudes obtenue à la première étape est considéré comme appartenant à la source correspondante (ici la $k^{\text{ième}}$ source s_k). Dans une deuxième étape, la matrice des atténuations de délais est obtenue en prenant les parties réelle et imaginaire des coefficients assignés à la source s_k et en corrigeant itérativement le paramètre de délai jusqu'à ce qu'il maximise la fonction potentielle définie en (2.14). Enfin une troisième étape consiste à estimer les signaux sources en supposant que leurs décompositions spectrales possèdent une distribution Laplacienne (hypothèse de parcimonie des sources), ce qui revient à minimiser la somme des amplitudes des sources.

Les adaptations de DUET sont nombreuses, dans le cas de mélanges anéchoïques bien sûr [Balan et al., 2003a] [Balan et al., 2003b], mais aussi convolutifs [Blin et al., 2004] [Melia and Rickard, 2005] [Melia and Rickard, 2007].

2.6 Spatial Audio Coding

Faisons maintenant une brève remarque sur la distinction entre l'approche de séparation de sources informée, et l'approche développée par Baumgarte, Faller et Herre dans une multitude de publications dont les principales sont [Baumgarte and Faller, 2003] [Faller and Baumgarte, 2003] [Faller, 2004] [Herre et al., 2004] [Herre et al., 2005] [Faller, 2006]. Par certains aspects, l'approche informée que nous développons dans cette thèse peut paraître proche, dans l'esprit, de celui du système de Spatial Audio Coding (SAC) MPEG développé par Faller et al., mais gardons à l'esprit que le but de la séparation de sources informée est de **complètement** séparer les signaux sources à partir des signaux de mélange non compressés. Au contraire, l'objectif de la méthode de MPEG-SAC est seulement de resynthétiser/respatialiser la scène audio grâce à une version compressée du mélange dont la dimension a été réduite (*downmix* en anglais). Par conséquent, la nature de l'information transmise, la façon dont elle est transmise, et la façon dont elle est exploitée (pour la séparation et non la spatialisation) sont complètement différentes de celles de l'approche SAC.

2.7 Conclusion

Les méthodes de séparation de sources basées sur la parcimonie exploitent le fait que, dans un domaine transformé, les signaux sources composant un mélange se superposent (beaucoup) moins que dans leur domaine initial (le plus souvent le domaine temporel). De telles méthodes permettent, en réduisant localement la dimension du mélange, de fournir des solutions efficaces à des problèmes de séparations de sources sous-déterminés pour lesquelles les méthodes de type ACI sont inadaptées. Cependant, les hypothèses fortes faites sur les sources (une à deux sources localement actives dans [Yilmaz and Rickard, 2004], respectivement dans [Bofill and Zibulevski, 2001]) sont encore insuffisantes pour traiter efficacement des mélanges musicaux réalistes d'un grand nombre de sources. D'une configuration initialement aveugle où il s'agit de *séparer pour connaître*, les hypothèses faites sur le processus de mélange ou bien les signaux sources ont conduit à des techniques de séparation de sources semi-aveugle. Nous nous proposons au cours de ce travail de thèse d'utiliser les grands principes de séparation de sources basée sur la parcimonie en enrichissant un certain nombre de ces hypothèses par une approche informée : il s'agit donc dans notre cas de *connaître pour séparer*. Dans les chapitres 4, 5 et 6, nous détaillons le principe de la séparation de sources informée qui possède les caratéristiques suivantes :

– parfaite connaissance des signaux sources avant mélange (hypothèse *a priori* poussée à l'extrême),

– séparation d'un grand nombre de sources à partir d'un mélange monophonique ou stéréophonique (cas extrêmes de mélange sous-déterminé),
– séparation des signaux audio avec la plus grande qualité d'écoute possible.

Chapitre 3

Bref état de l'art du tatouage audio-numérique

Abordons maintenant les techniques existantes de tatouage audio. Nous donnons dans cette section un bref aperçu des techniques de tatouage classiques et leur utilisation dans un but sécuritaire. Nous nous intéressons ensuite au cas de figure qui nous concerne directement, totalement différent d'un tatouage de type sécuritaire, mais qui s'apparente plus au transport de métadonnées, avant de présenter deux grandes techniques de tatouage numériques adaptées à cet objectif.

3.1 Principe général et application du tatouage de type sécuritaire

Le tatouage d'un signal, aussi appelé *watermarking* dans la littérature anglo-saxonne, exploite les caractéristiques du système perceptif humain pour insérer dans un média, en l'occurrence un signal sonore, une information qui soit imperceptible. La règle première de tout watermark audio est donc qu'elle doit être inaudible. Dans la grande majorité des cas, le tatouage audio est utilisé dans le cadre de la protection et du contrôle des droits d'auteur (en anglais on parle de Digital Right Management ou DRM) [Boney et al., 1996]. Le développement d'internet, la facilitation de l'échange de données ont fait du respect de la propriété intellectuelle un problème crucial dans le domaine de la création d'oeuvres à support numérique. On peut ainsi tatouer sur une chanson des informations permettant d'identifier l'auteur ou le propriétaire du fichier. Typiquement, les techniques employées sont de type étalement spectral, comme celles développées dans [Garcia, 1999] ou [Cox et al., 1997]. Dans ce cas, l'objectif est d'insérer l'information de tatouage de façon très robuste, c'est-à-dire résistante à de possibles manipulations plus ou moins licites du signal. Cette information, de taille relativement faible, est étalée dans une large plage temps-fréquence du signal puis ajoutée à celui-ci, de sorte qu'il est très difficile de pouvoir l'isoler pour la supprimer. Ce cadre d'application est donc relativement éloigné de notre objectif et on ne détaillera pas plus ce pan de la bibliographie.

3.2 Un tatouage informant

Dans notre cas de figure, nous proposons une utilisation tout à fait originale du tatouage, puisque nous proposons d'insérer une information permettant la séparation de sources à partir d'un mélange. L'information insérée porte ici sur les sources elles-mêmes et/ou le processus de mélange. Il ne s'agit pas d'une information sécuritaire de type copyright, mais de descripteurs caractéristiques des signaux sources et de leur contribution au mélange, au sens du traitement du signal, ces descripteurs devant permettre d'aider à la séparation des signaux. Il peut s'agir par exemple, comme on le détaillera par la suite, de descripteurs du contenu spectral des sources à séparer permettant d'estimer la participation respective de chaque source dans le mélange, notamment dans les zones de superposition du plan temps-fréquence. Il s'agit donc ici d'une information à la fois relativement volumineuse et répartie de façon bien localisée et bien contrôlée dans le plan temps-fréquence (le raffinement de la watermark sera fonction de la place dont nous disposerons pour l'insérer sans pour autant que cette modification du signal initial ne soit audible). En contrepartie, nous n'avons pas de contrainte sur la robustesse du tatouage. En effet nous nous plaçons dans un cadre applicatif tel que l'on suppose que le support ne subit pas d'attaques intentionnelles.

Les techniques qui se rapprochent le plus de notre cas d'étude sont celles où le watermarking devient un moyen de rajouter une "meta-information" de type non-sécuritaire à un média. Le média intégrant le tatouage devient alors canal pour la transmission de données. Par exemple, le tatouage pour la transmission de données est actuellement utilisé pour l'annotation de documents en vue d'une indexation dans une base de données [Tachibana, 2003], ou pour l'identification de documents dans le but d'établir des statistiques sur la diffusion de ce document [Nakamura et al., 2002]. Iliev et al. [Iliev and Scordilis, 2004] proposent une méthode de tatouage haute capacité utilisant la différence de phase entre les deux voies d'un mélange stéréophonique pour déterminer les portions du signal où insérer l'information de tatouage. L'utilisation d'un modèle psychoacoustique permet de contrôler la distorsion de la phase provoquée et d'assurer son inaudibilité. Des taux d'insertion supérieurs à 105kbits/s ont pu être atteints pour des mélanges audio de qualité CD audio. Un système de tatouage se compose alors de deux parties à l'image d'une chaîne de communication : un encodeur-émetteur où le tatouage est construit et inséré dans le signal porteur, et un récepteur-décodeur où la watermark est extraite du signal. Entre les deux, un canal de transmission peut engendrer des perturbations. Cox *et al.* [Cox et al., 1999] soulignent le fait que le tatouage peut être vu comme une transmission de données avec informations supplémentaires (*communication with side information*). Ils mettent en évidence le fait que l'on puisse mettre à profit les propriétés du signal hôte, toujours connu à l'émetteur (là où est formé le tatouage), pour améliorer les performances de détection du tatouage au décodeur. Le but est alors de choisir un tatouage optimal adapté au signal sur lequel il est inséré, et les contraintes à satisfaire sont d'obtenir un débit de transmission le plus élevé possible sans pour autant que la watermark soit audible, et également d'assurer une fiabilité de transmission la meilleure possible (peu d'erreurs faites au cours de la

transmission). On peut parler de tatouage informé par le signal hôte. Notons que nous pouvons donc bien considérer notre procédé de séparation de source comme une "séparation de source informée", avec un sens un peu différent de celui ci-dessus : dans notre cas c'est la séparation qui est informée par le tatouage sur les caractéristiques des signaux sources, et pas seulement le procédé de tatouage lui-même. En d'autres termes, le tatouage n'est pas seulement informé, ici il devient aussi informant.

3.3 Techniques de tatouage LSB et QIM

Au niveau de la technique de tatouage proprement dite, dans le cadre du tatouage pour la transmission de données, on peut citer la technique de watermarking substitutif (les caractéristiques du signal hôte sont remplacées par celles du tatouage) par quantification de Chen [Chen and Sundberg, 2000], ou celle de substitutions de sous-bandes fréquentielles du signal audio présentées par Bourcet [Bourcet et al., 1995]. Compte tenu des contraintes et des spécificités qui sont les nôtres en séparation de source informée, deux techniques de tatouage sont particulièrement intéressantes : le tatouage par LSB et le tatouage par QIM.

Le tatouage par *Least Significant Bit* (LSB) consiste à insérer une information sur les bits de poids faible du signal à tatouer. Les derniers bits (les plus faibles) d'une valeur quantifiée sont remplacés par l'information à tatouer, ou plus précisément le code de l'information à tatouer, l'information initialement portée par ces derniers bits étant définitivement perdue. Les principaux avantages de cette méthode de tatouage sont sa facilité de mise en oeuvre, sa rapidité tant au codage qu'au décodage, ainsi qu'un calcul simple du débit de tatouage (fixé par le nombre de bits modifiés). Des exemples d'utilisation du tatouage LSB sont donnés dans [Gil-Je et al., 2008] et [Cvejic and Seppanen, 2004]. Ces avantages sont contrebalancés par un inconvénient majeur, la robustesse limitée de ce type de tatouage, qui nécessite souvent l'utilisation de codes correcteurs d'erreurs en post traitement pour obtenir une récupération correcte de la watermark. Cvejic et al. proposent dans [Cvejic and Seppanen, 2002a] une technique de tatouage LSB plus robuste qui permet d'insérer le tatouage jusqu'au sixième bit de poids faible (dans un signal audio de qualité CD, *i.e.* encodé sur 16 bits. Autrement dit les 6 derniers bits sont modifiés par tatouage.), tout en assurant l'inaudibilité du tatouage en adaptant le bruit causé par la modification des bits de poids faible. Dans [Cvejic and Seppanen, 2002b], les mêmes auteurs introduisent une autre technique de tatouage LSB, cette fois dans le domaine des transformées en ondelettes, leur permettant d'atteindre des débits de l'ordre de 150 à 200kbits/s. La quantification par LSB est réalisée à l'aide de quantificateurs. Techniquement, le tatouage par LSB revient à utiliser deux quantificateurs, l'un grossier et l'autre plus fin. Le quantificateur grossier est utilisé pour fixer un niveau de référence sur lequel est d'abord quantifié le signal. Puis une seconde quantification est réalisée grâce au quantificateur fin, véritable porteur de l'information de tatouage. Le tatouage est alors la différence entre la valeur du signal tatoué sur le quantificateur fin et sa valeur sur le quantificateur grossier.

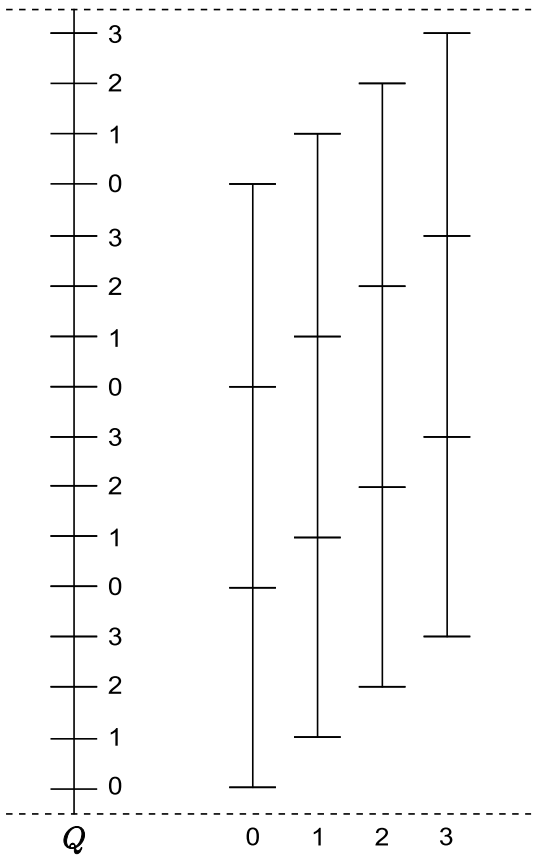


Figure 3.1 – Exemple d'un jeu de 4 quantificateurs pour la QIM. À droite les quantificateurs individuels, à gauche, le quantificateur résultant de leur réunion.

Chen et Wornell [Chen and Wornell, 2001] sont les premiers à introduire le tatouage par QIM pour *Quantization Index Modulation*. Il s'agit aussi d'un tatouage par quantification dans lequel le tatouage est porté par une modification des niveaux de quantification des descripteurs du signal hôte. En ce sens, cette technique peut être vue comme une généralisation de la technique de quantification LSB. Le signal tatoué x^W est vu comme une fonction du signal original x indexée par l'information à tatouer m : $x^W = f(x, m)$ avec la contrainte que $f(x, m) \approx x, \quad \forall m$. Dans la technique de tatouage par QIM, il y a autant de quantificateurs que de messages différents à insérer : chaque message m est quantifié sur un quantificateur dédié $f(.\,, m)$. Le schéma présenté Figure 3.1 donne un aperçu d'un jeu de quatre quantificateurs utilisés pour le tatouage respectif des quatre messages 0, 1, 2 et 3. À gauche, est présenté le quantificateur Q obtenu par la réunion des quatre quantificateurs individuels. Q correspond ici au quantificateur fin utilisé dans le tatouage par LSB, dans le cas de figure où les quantificateurs utilisés sont uniformes et où l'entrelacement des quantificateurs est régulier. Pour former le signal tatoué à partir du signal hôte et du tatouage m, le signal hôte est quantifié directement sur Q par le niveau le plus proche codant le message m. Pour décoder la watermark, il suffit de quantifier à nouveau le signal tatoué sur le quantificateur Q : le symbole décodé est celui correspondant au quantificateur individuel composant Q sur lequel est quantifié le message tatoué.

Les performances théoriques de cette technique s'approchent du modèle de Costa

[Costa, 1983] qui fixe la limite théorique de la capacité de transmission d'une chaîne de transmission si l'on connaît a priori le signal à l'émetteur. Comme nous le verrons plus en détails à la section 4.4, cette technique permet de tatouer un débit d'information relativement important, et nous la considérerons tout particulièrement dans notre application de séparation de sources informée.

Pour conclure ce bref état de l'art, rappelons que notre procédé est donc une application originale du tatouage pour la séparation de source. A notre connaissance, il n'existe qu'une seule étude, très récente, utilisant le watermarking dans un but de différencier des signaux d'un mélange. Il s'agit de l'étude proposée par Yi-Wen Liu [Liu, 2007]. Cependant cet article ne traite pas de séparation de sources à proprement parlé mais d'une simple ségrégation des signaux, selon les propres mots de l'auteur : la nuance est justifiée par la mauvaise qualité des signaux reconstruits (cette étude ne propose d'ailleurs pas de résultats qualitatifs sur la séparation). L'intérêt de la méthode en l'état semble se limiter à une simple estimation de la mélodie (évolution de la fréquence fondamentale du signal). De plus, dans nos travaux, le tatouage se fait sur le mélange des signaux et non sur les signaux sources à séparer comme c'est le cas dans l'étude citée.

Chapitre 4

Principes généraux pour la séparation de sources informée par tatouage

Dans ce chapitre, nous décrivons plus en détails les principes de la séparation de sources informée donnés en introduction. Nous précisons les conséquences de cette approche spécifique, en terme de structure de réalisation du procédé, et en terme d'information de tatouage. Plusieurs structures se différencient nettement, en fonction de la configuration du mélange et de la nature de l'information à tatouer. Nous ne donnons ici que les grandes lignes des différentes phases qui composent ces structures, de façon à en avoir rapidement une bonne vision globale. Le détail des implémentations et expérimentations associées sera donné dans les chapitres suivants.

4.1 Un tatouage porteur d'informations sur le signal lui-même

Contrairement à la séparation aveugle de sources où aucune information *a priori* n'est connue sur les signaux sources, la séparation de sources informée se place dans le cadre où une quantité possiblement importante d'informations sur les signaux sources est insérée lors du processus de mixage, dans le but d'aider à la séparation. Le tatouage porte, selon les cas, une image de la structure des signaux que l'on cherche à séparer, de leurs caractéristiques, ou encore de leur contribution au mélange. Il contient dans tous les cas des informations relativement précises sur les signaux sources : des informations qui doivent être suffisamment représentatives de chaque signal source pour permettre de le différencier des autres signaux auxquels il est mélangé. Ainsi, dans le Chapitre 5, nous proposons d'utiliser comme watermark une batterie de descripteurs de la structure temps-fréquence locale des signaux sources. Dans le Chapitre 6, nous utilisons comme watermark des index qui identifient la ou les sources prédominantes dans le mélange. Au décodage, ces informations pilotent la phase de séparation.

La distinction entre ce type de watermark et un tatouage de type DRM est fondamentale. Dans le cadre du watermarking sécuritaire, la watermark n'a pas de lien direct avec la structure ou les caractéristiques de son support en tant que signal, dans

la mesure où elle ne cherche pas à décrire ce support. Les informations portées sont simplement relatives à l'identification du signal (au sens de la propriété intellectuelle ou industrielle) et sont de fait peu volumineuses en terme de données utiles. À l'inverse, la watermark insérée en séparation de sources informée est profondément liée à la structure de chaque signal sur lequel elle est fixée.

Le signal de tatouage considéré ici est généralement beaucoup plus volumineux qu'un simple tatouage de type copyright, car beaucoup plus riche en terme d'informations utiles. Ce facteur de taille de la watermark rend très difficile, voire impossible toute addition supplémentaire de redondance au message lors de l'insertion (comme c'est le cas pour améliorer la robustesse du tatouage dans certaines applications de type sécuritaire). Ceci implique d'utiliser des techniques d'insertion du tatouage différentes de celles du watermarking sécuritaire, où tout du moins à les utiliser différemment : les techniques considérées ici seront moins axées sur la robustesse de la watermark que sur un débit élevé d'informations utiles à transmettre par tatouage. En résumé on peut dire que dans notre étude l'information portée par la watermark constitue un signal à part entière, relativement volumineux, d'une nature intrinsèquement liée à celle de son support, implanté localement sur un signal hôte dont il décrit la structure ou celle des signaux qui le composent.

4.2 Plusieurs structures possibles pour un système de séparation de sources informée

La technique de séparation de sources informée introduite dans ce manuscrit étant à la croisée de domaines de traitement du signal particulièrement riches, elle possède de fait de multiples variantes. Nous présentons ici les principales configurations que nous avons imaginées et détaillons celles que nous avons plus particulièrement étudiées. Cette technique de séparation de sources reste cependant relativement ouverte, et possède vraissemblablement d'autres variantes non traitées dans ce manuscrit. Les différentes configurations que nous présentons ici se différencient avant tout par la nature de l'information de description des sources à insérer, ainsi que le signal support de cette information. Trois structures émergent :

1. système de séparation de sources informée pour un signal de mélange monophonique,
2. système de séparation de sources informée pour un signal de mélange stéréophonique,
3. système de séparation de sources informée avec tatouage des signaux sources.

Les figures 4.1, 4.2 et 4.3 décrivent ces trois cas de figure. Dans un but de lisibilité, les schémas sont présentés dans le cas de deux signaux sources, mais le procédé proposé se généralise au cas de I sources, $I \geq 2$. Bien que cela ne soit pas apparent sur les schémas, la nature des descripteurs utilisés et les processus de séparation sont spécifiques pour chacune des trois structures (ils seront détaillés par la suite).

Les deux premières approches reposent sur une démarche commune. Tout d'abord, on effectue la comparaison du signal de mélange avec chacun des signaux sources afin d'en tirer une information quantitative sur la contribution de chacun de ces signaux sources dans le mélange. Cette comparaison est effectuée à l'encodeur, là où les signaux sources sont disponibles et où le mélange est généré. Le tatouage inséré sur le mélange encode alors le résultat de cette comparaison entre le contenu de chaque signal source et le mélange. Au décodeur, l'information de contribution relative de chaque source dans le mélange est extraite du signal de mélange, puis utilisée pour reconstruire chaque signal source à partir du mélange.

Dans les deux cas, le support de l'information de tatouage est le signal de mélange, et les deux approches diffèrent par la richesse de l'information tatouée. Dans le cas 1 d'un mélange monophonique (une seule observation du mélange), illustré Figure 4.1, une watermark particulièrement riche en informations sur les caractéristiques des sources est insérée. Le tatouage peut alors être assimilé à une technique de codage des signaux sources fonctionnant de façon relativement indépendante du comportement local du mélange. C'est le premier cas que nous traiterons, dans le chapitre 5, et auquel nous ferons désormais référence sous l'appellation de *SSI par codage des signaux sources*. Dans le cas 2, illustré Figure 4.2, le mélange considéré est stéréophonique (deux observations [1]). La nature de l'information insérée est alors relative à l'identification des sources prédominantes au sein du mélange, *i.e.* combien de sources sont actives localement, et quelles sont ces sources. On se réferrera dans la suite du manuscrit à cette configuration sous l'appellation de *SSI par indexation des sources* ou IISS pour *Index Informed Source Separation*. La watermark, particulièrement concise ici, nécessite des capacités d'insertions plus faibles que celles du cas 1. Du point de vue du contenu, cette information moindre est compensée par l'exploitation de l'information spatiale entre les deux canaux. Elle peut, quand cela s'avère nécessaire être couplée à un tatouage de type *codage des signaux sources* formant ainsi un système hybride, comme nous le verrons au Chapitre 7.

Enfin, dans le troisième cas de figure, les signaux sources sont tatoués avant d'effectuer le mélange. L'interaction entre les techniques de tatouage et de séparation suit alors une démarche générale différente de celle des deux cas précédents. Par exemple, comme cela est illustré Figure 4.3, plutôt que d'extraire le tatouage pour guider ensuite la séparation, on peut envisager de séparer d'abord les sources "grossièrement" avec une technique quasi-aveugle, puis d'extraire le tatouage pour régulariser le résultat de la séparation. L'information issue du tatouage pourrait permettre par exemple de lever les indéterminations de permutation et de facteur d'échelle propres à la séparation de sources aveugles (voir la Section 2.3). Le tatouage devrait être relativement compact et doit être robuste au processus de mélange ainsi qu'au processus de séparation. On peut donc penser à des techniques de watermarking inspirées du tatouage sécuritaire (étalement spectral, insertions multiples temporelle ou fréquentielle...). Les schémas de

1. En pratique, nous considérerons toujours un nombre de sources au moins égal à trois dans nos expérimentations, pour se placer ici en configuration sous-déterminée.

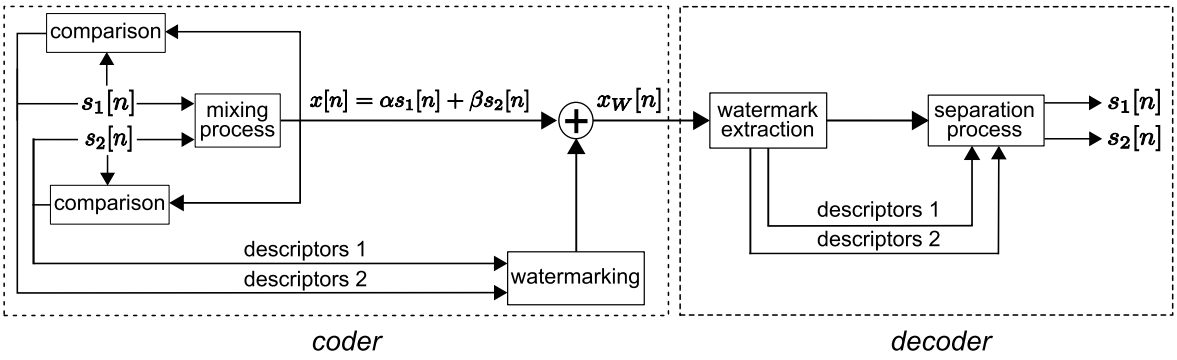


Figure 4.1 – Schéma simplifié du système codeur/décodeur dans le cas d'un mélange monophonique.

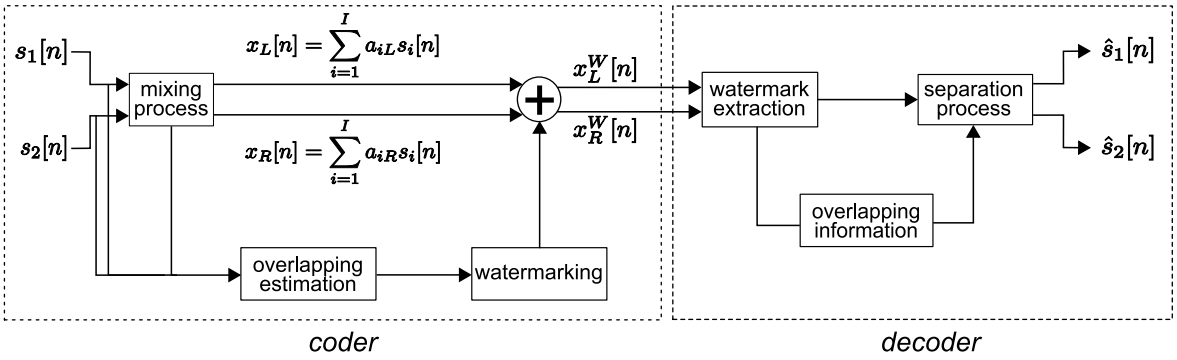


Figure 4.2 – Schéma simplifié du système codeur/décodeur dans le cas d'un mélange stéréophonique.

détection doivent bien évidemment être adaptés à ce type de techniques. Cette troisième configuration de séparation de sources informée n'a pas été étudiée en détails dans cette thèse et ne sera donc pas présentée plus amplement dans ce manuscrit. Elle est donnée à titre d'exemple de ce qu'il est possible d'envisager en séparation de sources informée, pour donner un aperçu de l'étendue du domaine.

Notons que dans toutes les configurations proposées, le signal hôte peut être considéré comme un canal de transmission de l'information de tatouage permettant, au décodeur, de guider la séparation des sources en tant que telle par l'extraction de la watermark. La technique de tatouage qui sera utilisée dans les deux configurations de tatouage du mélange étudiées (mélanges monophonique et stéréophonique) est inspirée de la technique de Quantization Index Modulation (QIM) de Chen et Wornell [Chen and Sundberg, 2000]. Cette technique de tatouage peut être vue comme un moyen d'insérer sur un signal hôte un autre signal d'informations complémentaires et de volume important, ce qui réaffirme l'originalité de l'utilisation du tatouage en SSI. Nous détaillons dans les sections suivantes les principes généraux de l'obtention des descripteurs des signaux, ainsi que de la technique de tatouage. Les détails d'implémentations seront eux donnés Chapitre 5.

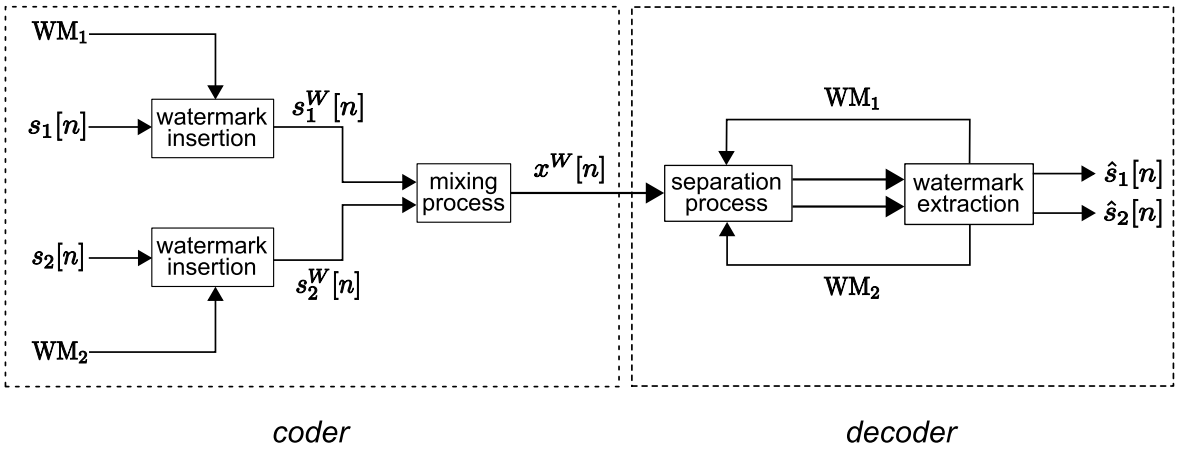


FIGURE 4.3 – Schéma simplifié du système codeur/décodeur dans le cas du tatouage des signaux sources.

4.3 Une représentation des signaux adaptée : un traitement dans le domaine en temps-fréquence

4.3.1 Principe général

Les signaux de parole et de musique présentent une large variabilité spectrale, et une nature (potentiellement fortement) non-stationnaire. Par exemple, on sait que les signaux audio et les signaux de parole ont une énergie beaucoup plus importante en basse fréquence qu'en haute fréquence. En musique, des passages "forts" avec une grande dynamique et une grande richesse spectrale peuvent alterner avec des passages plus doux présentant une plus grande parcimonie spectrale. C'est pourquoi il est nécessaire, en SSI, comme en séparation de sources classique, de tenir compte à la fois de la diversité spectrale des signaux sources (une même source peut contribuer fortement au mélange dans une région spectrale, et contribuer très faiblement au mélange dans une autre région spectrale) et de leur diversité temporelle, c'est-à-dire de leur non-stationnarité (si les composantes des signaux de parole et de musique évoluent continûment au cours du temps, il en est de même de leur contribution dans le signal de mélange). En d'autres termes, on ne peut pas se contenter de comparer les signaux dans le domaine des échantillons temporels, même à une échelle temporelle locale (car toutes les composantes spectrales des signaux sont "superposées" dans les échantillons du mélange), ni se contenter d'une étude spectrale globale sur l'ensemble du signal (qui ne permettrait pas de séparer avec précision les signaux en temps). Il faut au contraire être en mesure de collecter des caractéristiques les plus précises possibles sur chacun des signaux dans toute la bande de fréquence considérée, et de façon évolutive au cours de toute la durée du signal. C'est pour cette raison qu'il est judicieux de transformer les signaux sources aussi bien que celui du mélange, initialement fournis dans le domaine temporel, dans un domaine donnant conjointement leur évolution temporelle et fréquentielle : une représentation temps-fréquence (TF) des signaux semble particulièrement appropriée pour le problème.

Une deuxième justification de l'utilisation d'une représentation temps-fréquence des signaux étudiés tient aux propriétés de parcimonie des signaux audio. En effet, nous avons vu au Chapitre 2 que les signaux audio de parole ou de musique sont beaucoup plus parcimonieux dans le domaine temps-fréquence que dans le domaine temporel, d'où une concentration de l'énergie des signaux traités sur un faible nombre de coefficients. Ceci permet, dans notre étude, de pointer les zones les plus importantes au sens de la séparation qui devront porter une information de tatouage précise sur les sources pour permettre leur séparation au sein du mélange.

Enfin une troisième justification au traitement des signaux dans le domaine temps-fréquence est relative au tatouage. En effet, le tatouage proprement dit est réalisé par une modification judicieuse des coefficients de décomposition temps-fréquence du signal de mélange : ces coefficients sont quantifiés sur une échelle spécifique et les niveaux de quantification encodent l'information tatouée. Pour être plus précis, et comme ceci sera détaillé par la suite, ces coefficients subissent une sur-quantification par rapport à une échelle de quantification de référence. Cette modification des coefficients doit être sans conséquences sur la qualité audio du signal. Nous verrons que l'on peut exploiter la robustesse de coefficients temps-fréquence déjà mise en évidence dans le domaine du codage audio.

En résumé, la variabilité spectro-temporelle des signaux, leur parcimonie dans le domaine temps-fréquence, et la technique de tatouage par quantification choisie sont donc trois arguments majeurs justifiant le recours à une transformation des signaux traités dans le domaine temps-fréquence.

4.3.2 Définition de la décomposition temps-fréquence

La décomposition temps-fréquence d'un signal $s(t)$ consiste à réaliser une analyse spectrale du signal sur une fenêtre centrée autour d'un instant t_0 et à déplacer cette fenêtre sur l'ensemble de la durée du signal. On obtient donc une collection de spectres locaux dépendant de l'instant, noté b dans cette sous-section, autour duquel est centrée chaque fenêtre. Ces spectres peuvent être vus comme la décomposition du signal sur une base de fonctions $\varphi_{b,f}(t)$ dépendant des deux paramètres temps et fréquence. Ces fonctions sont les atomes de base du plan temps-fréquence. Les atomes forment un pavage régulier du plan temps-fréquence, lorsque les fenêtres d'analyse sont de taille constante. Les coefficients de décomposition du signal $s(t)$ sur la base des atomes $\varphi_{b,f}(t)$, notés $C_s(b,f)$, sont obtenus par la projection $\langle s, \varphi_{b,f} \rangle$, où $\langle ., . \rangle$ désigne le produit scalaire. Dans la suite, lorsque nous ferons référence au pavage temps-fréquence d'un signal s, il faudra comprendre l'ensemble des coefficients de décomposition $C_s(b,f)$ du signal s.

Notons que pour localiser de manière précise un signal en temps et en fréquence, on souhaite que les résolutions temporelle et spectrale tendent simultanément vers zéro. Cependant, une application des relations d'Heisenberg-Gabor aux représentations temps-fréquence précise qu'un signal ne peut avoir une localisation arbitrairement précise en temps et en fréquence : plus on veut se localiser précisément sur une portion

d'un signal, moins on peut spécifier les fréquences de ce signal précisément. En pratique, cela passe par un compromis entre la taille temporelle des segments d'analyse, et la précision du traitement (dans notre cas la séparation) en fréquence. On choisit une taille de fenêtre d'analyse telle qu'on puisse à la fois suivre correctement les non-stationnarités des signaux, et avoir une résolution fréquentielle satisfaisante pour la séparation.

4.3.3 Une approche à une échelle intermédiaire

Toutes les phases de traitement (comparaison des signaux, génération des descripteurs, mise en forme du watermark et insertion dans le mélange, extraction au décodeur, séparation des signaux) sont effectuées localement dans le domaine temps-fréquence, sur les coefficients de la décomposition choisie. Il faut donc dans un premier temps décomposer l'ensemble des signaux, sources et mélange, dans le plan temps-fréquence, et ceci avec la même transformation. Les pavages temps-fréquence résultant des sources et du mélange étant alors de mêmes dimensions, il est possible d'effectuer, dans un second temps, des comparaisons entre mélange et signaux sources sur l'ensemble de ce pavage temps-fréquence, en découpant le plan en zones élémentaires. La comparaison des coefficients de leur décomposition fournit le "comportement relatif" des sources par rapport au mélange dans chacune de ces zones élémentaires du plan temps-fréquence, en termes quantitatifs. Cette étude locale de chaque source et du mélange permet bien de tenir compte de la diversité spectrale et de la non-stationnarité des signaux dans la comparaison. Dans un troisième temps, les informations extraites de chaque zone de décomposition des signaux sources (les "descripteurs" de cette zone de signal) sont tatouées sur les coefficients de la zone de décomposition temps-fréquence correspondante du signal de mélange. Au décodeur, chaque source est estimée à partir de la décomposition temps-fréquence du signal de mélange dans chaque zone élémentaire et à partir de l'information de contribution de cette source fournie par le descripteur tatoué dans la même zone du signal de mélange. Le signal de mélange peut donc être vu à la fois comme un élément de comparaison des signaux sources par rapport à lui-même, un support pour le tatouage de l'information descriptive des sources, et un élément de reconstruction des sources.

Plutôt que de considérer le pavage temps-fréquence du mélange à l'échelle "microscopique", i.e. à l'échelle de chaque coefficient de la décomposition du mélange dans le plan temps fréquence, nous avons abordé le problème à une échelle plus grande que l'on peut qualifier de "macroscopique". Les coefficients ou *atomes* formant le pavage temps-fréquence des signaux ne seront pas considérés individuellement, mais groupés selon leur proximité temps-fréquence. Ainsi un ensemble d'atomes du pavage temps-fréquence localisés dans la même zone du plan temps-fréquence sont groupés en une *molécule*. Cette notion de groupement d'atomes dans le plan temps-fréquence a été abordée par Gribonval dans [Gribonval and Bacry, 2003] puis Daudet dans [Daudet, 2006] pour la décomposition de signaux (dans une optique générale d'application à l'analyse-synthèse des signaux de musique). Dans notre étude, outre le fait qu'une

molécule mette plus en évidence la structure du signal qu'un simple atome, cette organisation macroscopique est nécessaire en tant que structure élémentaire du tatouage : elle permet en effet de tatouer, sur la décomposition temps-fréquence du signal de mélange, une information nécessaire à une bonne séparation des sources d'une taille qu'il serait impossible d'insérer sur un seul coefficient temps-fréquence. Comme nous l'avons déjà mentionné en termes très généraux à la Section 4.1 et comme nous le verrons en détails par la suite, les descripteurs du signal contiennent en effet des informations potentiellement volumineuses qu'il est plus raisonnable de considérer à moyenne échelle dans le plan temps-fréquence. Ainsi, notons dès maintenant que les dimensions des molécules devront constituer un compromis entre la quantité d'information que l'on souhaite insérer (directement reliée à la qualité des descripteurs), et la précision de la séparation/reconstruction des signaux sources en terme de résolution TF. Si l'échelle microscopique (au niveau atomique) constitue un pavage trop fin pour l'insertion d'un tatouage conséquent, à l'inverse une molécule trop grande ne peut pas refléter de façon précise les caractéristiques d'une seule source dans les zones du plan temps-fréquence où plusieurs sources se superposent. En résumé, la taille des molécules doit permettre à la fois de séparer correctement les sources, et de porter une informations de tatouage suffisante pour cet objectif. Le choix de ces molécules et leurs dimensions feront l'objet de plus amples développements dans les chapitres consacrés à l'implémentation de nos systèmes de séparation.

4.3.4 Illustration du principe de séparation

La Figure 4.4 illustre schématiquement ce principe de séparation des signaux sources à partir du signal de mélange dans le domaine temps-fréquence. Notons, que s'il s'agit d'un mélange stéréophonique, le schéma représente la situation sur l'une des deux voies. On a représenté sur cette figure la séparation de deux sources comportant chacune deux zones temps-fréquence significatives. Ces zones représentent de façon très schématique deux cas de figure typiques et pourtant très différents : une des zones correspond à des composantes des deux sources isolées dans le plan temps-fréquence, et l'autre zone correspond à des composantes des deux sources largement superposées. Dans les zones où les composantes sont isolées, la séparation est particulièrement simple et immédiate : chacune des deux sources est seule à contribuer au mélange dans sa zone respective et les signaux sources peuvent donc être reconstruits dans chacune des zones en question directement à partir du mélange. Dans la zone où les sources sont superposées, la séparation n'est pas aussi immédiate. Sur cette figure, pour les deux signaux sources estimés, la forme du signal reconstruit dans cette zone est celle du signal de mélange, avec cependant une pondération énergétique à partir de l'information fournie par le descripteur tatoué. La forme de la zone de signal estimé est celle du mélange mais son énergie est donc celle du signal source : c'est ce principe qui est à la base de la séparation. Sur cet exemple, la forme du signal estimé diffère donc significativement de celle de chaque signal source dans cette zone du plan TF. Ceci traduit évidemment la difficulté de séparer des sources superposées dans le plan TF. Nous verrons plus

en détails dans le Chapitre 5 en quoi cela peut limiter la qualité de reconstruction des sources, et comment y remédier en "enrichissant" la nature des descripteurs des signaux sources. Nous verrons ensuite au Chapitre 6 comment exploiter l'information spatiale en séparation de sources informée à partir d'un mélange stéréophonique.

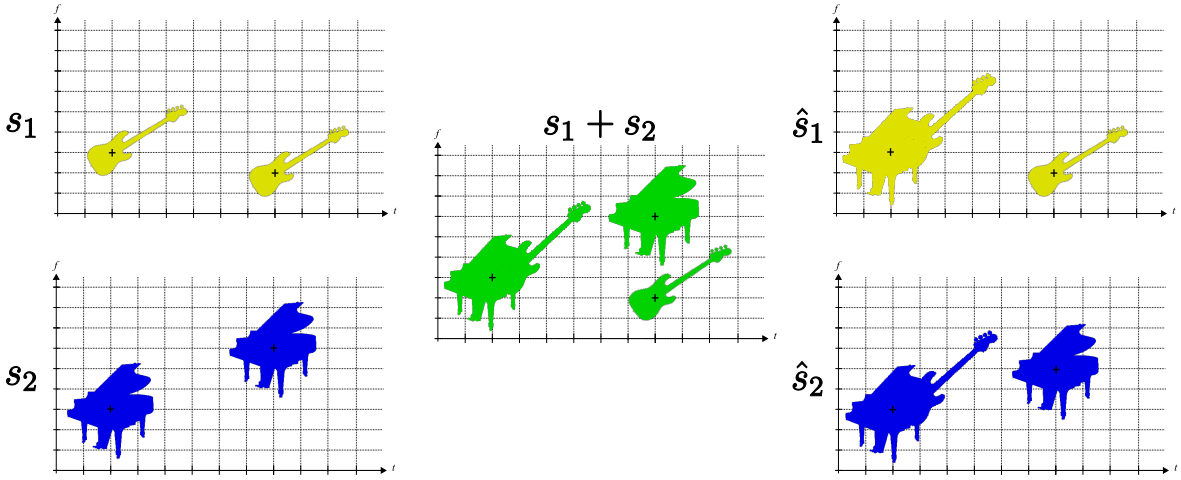


FIGURE 4.4 – Un principe simple de reconstruction de sources à partir du mélange, avec un descripteur énergétique des sources respectives.

4.3.5 Les descripteurs des signaux sources dans le domaine temps-fréquence

La décomposition de chaque signal dans le plan TF est unique. Les descripteurs des signaux sources utilisés pour guider la séparation sont donc construits dans le plan temps-fréquence à partir des pavages temps-fréquence des différentes sources, et du pavage temps-fréquence du mélange. Ces descripteurs offrent un moyen de comparaison quantitative des signaux sources entre eux ainsi qu'avec le mélange, et peut permettre, si la description est suffisamment précise, de les différencier les uns des autres. Ici aussi, l'étude proprement dite se fait à l'échelle d'une molécule : chaque molécule d'un signal source est comparée à la molécule correspondante du mélange (même localisation en temps et en fréquence). La localisation dans le plan TF constitue une première condition à la comparaison inter-molécules. Plusieurs types de descripteurs ont été envisagés, selon le type de mélange, mono ou stéréo, et pour un type de mélange donné. La multiplication des descripteurs permet de plus d'affiner la caractérisation des signaux et donc d'améliorer leur séparabilité. Des critères énergétiques sur une molécule entière (que le mélange soit à une ou deux voies) ou entre les différents coefficients d'une molécule (pour un mélange monocanal) semblent s'imposer, dans un premier temps, pour décrire à l'échelle moléculaire le pavage temps-fréquence des signaux à séparer. Les descripteurs temps-fréquence utilisés dans notre étude seront présentés en détails à la Section 5.1.4 pour la SSI à partir d'un mélange monophonique et à la Section 6.2.2 pour le cas de la SSI à partir d'un mélange stéréophonique.

4.4 La technique de tatouage

4.4.1 Principe général

Les deux configurations de séparation de sources informée que nous considérons (cf Section 4.2) sont toutes deux basées sur le tatouage du signal de mélange avec une contrainte de capacité significativement supérieure à celle des techniques de tatouage de type DRM. Même si la taille des informations à tatouer diffère entre ces deux configurations, la technique de tatouage sera la même dans les deux cas, avec des réglages différents, en raison de sa généralité et sa flexibilité. Nous introduisons cette technique dans cette section.

Comme la plupart du temps en matière de tatouage audio, la technique d'insertion de la watermark sur un signal utilise les limitations de perception de l'oreille humaine. La watermark que nous tatouons doit vérifier les trois principes de base suivants :

- être imperceptible pour l'oreille humaine
- être parfaitement détectable par le décodeur pour pouvoir utiliser l'information qu'elle transporte
- être résistante à un certain nombre de manipulations

En ce qui concerne l'inaudibilité de la watermark, on peut s'inspirer de ce qui existe dans le domaine de la compression audio par exemple. En effet l'utilisation de propriétés psychoacoustiques est très répandue dans les algorithmes de compression audio. Ainsi, la non-sensibilité de l'oreille humaine à une certaine quantification des coefficients temps-fréquence d'un signal audio a pu être utilisée dans certains standards MPEG [Brandenburg and Bosi, 1997]. La quantification du signal audio est variable en fonction de la pertinence perceptive des données. Les coefficients situés dans les zones où l'oreille est particulièrement sensible sont affectés d'un nombre de bits plus importants que ceux localisés dans des zones temps-fréquence moins perceptibles pour l'oreille. La technique de tatouage que nous utilisons pourra se baser sur le même principe, avec dans notre cas, non pas une quantification la plus parcimonieuse possible des coefficients de la décomposition temps-fréquence comme dans le cas de la compression, mais une quantification de ces coefficients permettant un ajout d'information. En d'autres termes, on ne supprime pas d'information inutile à l'oreille humaine, mais on modifie (par le tatouage) le signal dans les régions où cela est inaudible. En général, plus une portion du signal est énergétique, plus la capacité d'insertion de l'information est grande, ce qui permet de coder avec précision les descripteurs des signaux sources dans ces portions du signal. A contrario, la quantification des signaux dans ces zones est plus grossière (mais toujours inaudible). En conclusion, dans les zones les plus énergétiques du signal, la capacité d'insertion de l'information est élevée, d'où la possibilité de décrire avec précision ces portions du signal essentielles à la qualité audio des signaux.

Le tatouage que nous souhaitons insérer sur le mélange étant voué à transporter le plus d'information possible sur les sources, nous nous sommes intéressés à une technique

d'insertion par modification de niveaux de quantification, en nous basant sur les travaux de Chen et Wornell [Chen and Wornell, 2001] sur la technique Quantization Index Modulation (QIM), que nous avons déjà mentionnée au Chapitre 3. En séparation de sources informée, la QIM est directement appliquée sur les coefficients temps-fréquence de la décomposition du mélange. Le résultat de cette quantification doit être robuste à une conversion du signal dans le domaine temporel en fin de codeur, ainsi qu'à une nouvelle conversion dans le domaine temps-fréquence intervenant au décodeur pour permettre une bonne récupération de la watermark [2].

4.4.2 Allocation de bits

La richesse des descripteurs des signaux est à considérer à plusieurs niveaux : leur nature, leur nombre et leur précision de codage. De plus ceci est vrai pour chaque signal source, et dans le cas usuel où on l'on veut séparer plusieurs sources il faut tatouer les descripteurs de chacune de ces sources. En fonction de la place disponible pour insérer le tatouage, différents niveaux d'information (c'est à dire des descripteurs de richesse croissante) pourront être insérés sur le signal de mélange afin d'affiner la caractérisation des signaux sources. La répartition des bits de tatouage disponibles entre les différents descripteurs des différentes sources constitue l'étape d'allocation de bits (nommée ainsi par analogie avec la même étape de distribution de ressource binaire dans les algorithmes de compression). Plus la place disponible est élevée, plus on peut coder de sources et de descripteurs pour chacune d'entre elles, et ce avec une précision croissante. Cependant, si la place pour coder des descripteurs avec une précision optimale n'est pas disponible, on peut décider de tous les coder avec une précision moindre, ou de coder moins de descripteurs mais de façon plus précise. Un compromis est à trouver entre la nature des descripteurs utilisés, leur nombre, et la précision avec laquelle ils sont codés (pour un nombre de sources donné). Nous verrons des exemples de réalisations d'allocation dans la Section 5.1.4.4. Nous verrons aussi que cette allocation est réalisée de façon identique au codeur et au décodeur. La connaissance des critères d'allocation au décodeur permet de décoder les divers descripteurs avec la résolution appropriée et assure donc une bonne lecture de la watermark.

4.4.3 Conséquence sur le format du signal considéré en séparation de sources informée

L'insertion de la watermark dans le domaine temps-fréquence choisi consiste donc en une quantification spécifique des coefficients du pavage du mélange. Elle ne semble donc pas, au moins dans un premier temps, adaptée aux signaux compressés. En effet, du point de vue de la compression, une insertion d'information sur de tels signaux au-

2. S'il n'y a pas de perturbation du signal tatoué entre le codeur et le décodeur, le signal tatoué correspond exactement au point de reconstruction du quantificateur de tatouage ; s'il y a une perturbation, la quantification doit lui être robuste : voir la Section 5.1.3.4 pour le détail de ce problème dans notre application de la séparation de source informée au CD audio.

rait pour conséquence directe l'augmentation du débit de codage, ce qui va à l'encontre de la notion même de signaux compressés. A l'inverse, du point de vue du tatouage, exiger un débit de compression faible aurait pour conséquence de perdre l'information de tatouage et condamnerait la séparation de sources informée. Pour cette raison, le format PCM 16-bits, où les signaux ne sont pas compressés (signal échantillonné à 44100 kHz et quantification uniforme sur 16-bits des échantillons temporels) apparaît comme particulièrement adapté à la séparation de sources informée appliquée aux signaux de musique (pour les signaux de parole, on prendra également des signaux non compressés, avec une fréquence d'échantillonnage éventuellement plus faible). C'est pourquoi l'ensemble des signaux traités dans les chapitres suivants seront à ce format non compressé. Cependant, une piste de réflexion pour la poursuite des travaux développés dans cette thèse, est, comme nous le détaillerons par la suite, l'étude de la possibilité de tatouage pour des signaux traités par compression de type MPEG MP3 ou AAC.

Dans la suite nous serons amenés à faire référence à de nombreuses reprises à l'étape de conversion PCM 16-bits faisant référence soit à la fixation d'un signal sur le support CD soit à sa conversion au format wav, et à la quantification 16-bits des échantillons temporels qui accompagne chacune de ces deux conversions numériques. Cette quantification dans le domaine temporel est assimilable à l'ajout d'un bruit de quantification dans ce même domaine temporel. Bien que très faible, ce bruit a une répercution dans le domaine spectral et est donc susceptible d'avoir des conséquences sur le procédé de tatouage. Ce point sera plus amplement développé à la Section 5.1.3.4.

Rappelons ici qu'une autre caractéristique de la watermark insérée sur le signal audio du mélange dans le cadre de la séparation de sources informée est qu'elle n'est pas amenée à subir d'attaques, contrairement à un tatouage sécuritaire constitué d'un nombre de bits limité mais qui se doit d'être extrêmement robuste à tout type d'agressions. Il n'y a en effet aucun intérêt à vouloir supprimer la watermark fixée sur le mélange des signaux (elle est nécessaire à tout utilisateur pour que celui-ci puisse contrôler les différents signaux sources d'un mélange, et on exclut le cadre d'attaques mal intentionnées).

4.5 Conclusion sur ce Chapitre 4

Nous avons répondu au cours de cette section aux deux questions *où* et *comment* posées au cours de l'introduction. Où la séparation de sources informée est-elle réalisée ? Dans le plan temps-fréquence qui offre une représentation à la fois capable de capter les propriétés évolutives des signaux de musique et de parole, de fournir un traitement à la bonne échelle (moléculaire) et d'être un support approprié au tatouage (selon les critères de quantification détaillés plus tard). Comment est-elle mise en œuvre ? En insérant sur le mélange, par modification du niveau de quantification de ses coefficients de décomposition dans le domaine fréquentiel, des informations relatives à la contribution locale des signaux sources dans le mélange. Nous avons répondu à ces questions en termes de principes techniques très généraux. Nous allons maintenant décrire en

détails une première implémentation concrète de ces principes dans le cadre de la SSI par codage des signaux sources.

Chapitre 5

Une première implémentation : la séparation par codage des signaux sources pour un mélange linéaire instantané monophonique

Dans ce chapitre, nous présentons en détails une première réalisation d'un système de séparation de sources informée, dans le cas d'un mélange linéaire, instantané, stationnaire et monophonique (LISM), avec une approche basée sur un codage des signaux sources. Les grands principes de ce système ont été donnés au Chapitre 4 (cf Figure 4.1) et nous nous focalisons ici sur l'implémentation technique. Le mélange LISM traité au cours de ce chapitre est obtenu par combinaison linéaire des signaux sources avec des facteurs d'amplitude propres comme nous l'avons défini au Chapitre 2 et dont nous rappelons ici la forme. Nous considérons un mélange de I signaux sources, combinés linéairement tels que leur mélange x soit de la forme :

$$x[n] = \sum_{i=1}^{I} a_i s_i[n] \tag{5.1}$$

Il s'agit d'un cadre simple mais réaliste pour une première approche du mixage audio dans le cadre du format PCM 16-bits qui nous intéresse plus particulièrement. Les expérimentations et les résultats obtenus dans cette configuration pour cette implémentation sont présentés et commentés à la Section 5.2.

5.1 Implémentation

La description de cette implémentation se présente sous la forme suivante : on donne Figure 5.1 le schéma détaillé des différentes étapes de traitement des signaux mettant en oeuvre les principes développés au Chapitre 4, tant pour la partie encodeur où mixage et tatouage sont réalisés, que pour la partie décodeur où sont réalisées la détection du tatouage et son utilisation pour la séparation des sources. Nous décrivons

en détails les différents blocs de ce schéma dans les sections suivantes.

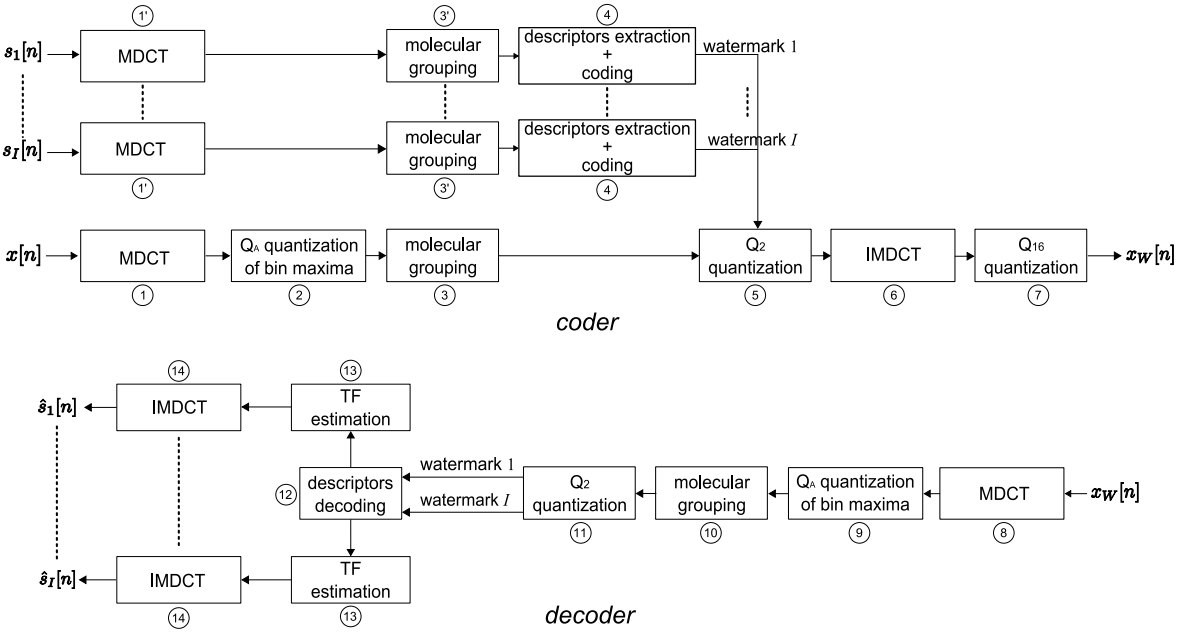


Figure 5.1 – Le bloc Codeur/Décodeur pour le système de séparation de sources informée par codage des signaux sources.

5.1.1 La transformée MDCT

Comme nous l'avons présenté à la Section 4.3, l'essentiel des traitements est réalisé dans le domaine temps-fréquence, et passe donc d'abord par une décomposition des signaux sources et du mélange dans ce domaine. Cette étape correspond au bloc 1 de la Figure 5.1 pour le codeur (le bloc 1 concerne le mélange ; le même traitement appliqué aux signaux sources est noté 1') et au bloc 8 pour le décodeur. Il existe un large éventail de transformées permettant une représentation des signaux temporels dans le plan temps-fréquence. Les transformées issues de la DCT pour *Discrete Cosine Transform* sont particulièrement adaptées aux représentations fréquentielles des signaux audio. La DCT est une transformation linéaire proche de la transformation de Fourier discrète (DFT). Le noyau de projection utilisé pour la DFT est représenté par une exponentielle complexe alors que le noyau de projection de la DCT est constitué par une base de cosinus. Les coefficients de la transformée en cosinus discrète ne sont donc pas complexes mais réels ce qui présente un avantage pour le codage et la quantification. Dans cette représentation, à l'inverse de la DFT, la phase des coefficients n'a pas besoin d'être codée explicitement en tant que telle (l'information de phase est inclue dans les coefficients DCT réels). L'information est, en général, essentiellement portée par les coefficients basses fréquences pour les signaux audio.

Dans la méthode de séparation que nous développons, nous avons fait le choix d'utiliser une transformée dérivée de la DCT : la transformée en cosinus discrète modifiée, plus connue sous l'appellation anglo-saxonne équivalente de Modified Discrete Cosine Transform (MDCT). La MDCT a été introduite par Princen et Bradley [Princen and

Bradley, 1986], puis développée sous sa forme actuelle par ces mêmes auteurs en 1987 [Princen et al., 1987]. C'est une transformation basée sur la Discrete Cosine Transform de type IV, mais appliquée sur des fenêtres temporelles à court terme (généralement quelques dizaines de millisecondes) qui se chevauchent. Le signal est ainsi découpé en trames temporelles consécutives de telle façon que la dernière moitié d'un bloc coïncide avec la première moitié du bloc suivant [1]. Notons que le recouvrement entre deux trames successives peut être inférieur à 50%, comme c'est par exemple le cas dans la transition "grande fenêtre"/"petite fenêtre" du codeur AAC [?]. Ce découpage en fenêtres ajouté aux propriétés de concentration de l'énergie des coefficients font de la MDCT une transformation relativement compacte particulièrement prisée pour la compression de signaux [Rulon et al., 1999] et [Sinha and Johnston, 1996] (voir aussi [Daudet and Sandler, 2004] pour une discussion plus approfondie sur le comportement des coefficients MDCT pour les signaux audio). Elle est par exemple utilisée dans les différentes normes MPEG.

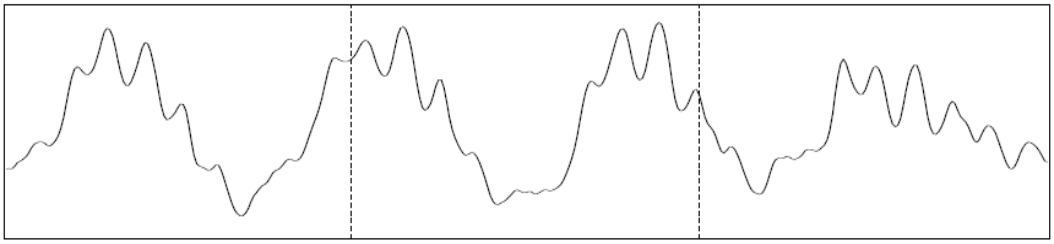
Au cours de la première étape du codeur et du décodeur de la figure 5.1, le signal temporel est donc découpé en fenêtres successives se recouvrant de moitié, puis une transformée est appliquée sur chaque fenêtre. Notons t l'indice entier de la fenêtre temporelle. Un bloc temporel de W échantillons d'un signal x commençant à l'instant $t \times W/2$, est transformé par MDCT en $W/2$ échantillons fréquentiels $m_t^x[f]$ selon la formule :

$$m_t^x[f] = \sum_{n=0}^{W-1} x[n + tW/2]\, w_a[n] \cos\left(\frac{2\pi}{W}(n + n_0)\left(f + \frac{1}{2}\right)\right) \tag{5.2}$$

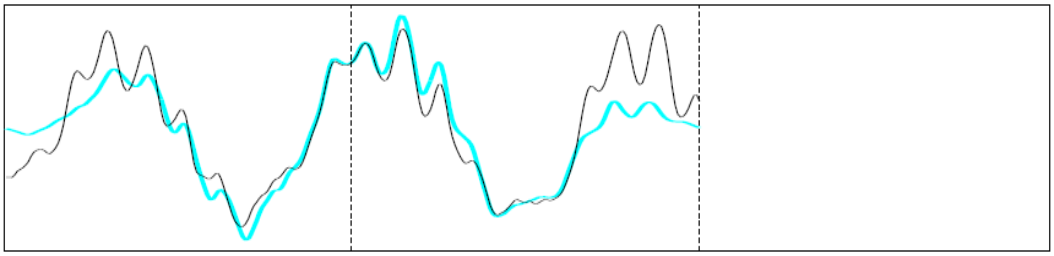
où $f \in [0, W/2-1]$, $n_0 = (W/2+1)/2$, et w_a est la fenêtre d'analyse temporelle de taille W. La taille de cette fenêtre conditionne la précision du pavage temps-fréquence. Ainsi pour un signal temporel de N échantillons et une fenêtre w_a de largeur W échantillons, avec un recouvrement de 50% on obtient un pavage de dimension $W/2$ en ordonnées, correspondant aux différents canaux fréquentiels, et de dimension $\frac{N}{W/2}+1$ en abscisses (valeur supposée entière [2]), correspondant aux nombres de fenêtres temporelles sur la totalité du signal. La décomposition MDCT fournit une matrice $\mathcal{M}_x = \{m_t^x[f]\}, f \in [0, W/2-1], t \in [0, 2N/W]$ (voir Figure 5.5). À la synthèse (qui a

1. La transformée MDCT n'est pas inversible à l'échelle d'une trame : l'inverse d'un bloc transformé ne fournit pas exactement le bloc de signal original comme le fait la transformée de Fourier discrète. Un phénomène de "Time-Domain Aliasing" apparaît. Il est analogue, dans le domaine temporel, au phénomène de repliement de spectre en fréquence. A la synthèse, une procédure de recouvrement et addition (overlap-add) de fenêtres temporelles successives permet d'annuler cet effet de repliement temporel et d'obtenir une parfaite reconstruction du signal comme le montre la figure 5.2 : le bloc de signal original est obtenu par un mélange du bloc de signal transformé inverse courant et des blocs suivant et précédent. On parle de transformée de type TDAC pour *Time Domain Aliasing Cancellation*.

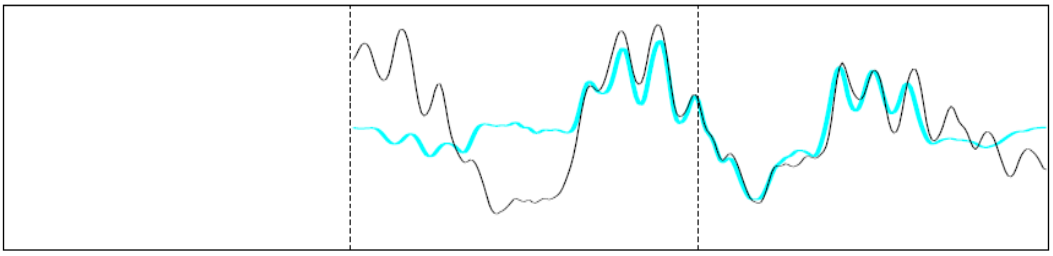
2. En pratique les signaux traités, de parole ou de musique sont de taille N grande devant les autres grandeurs, et il est toujours possible de rajouter des échantillons nuls en fin de signal de sorte à assurer cette condition.



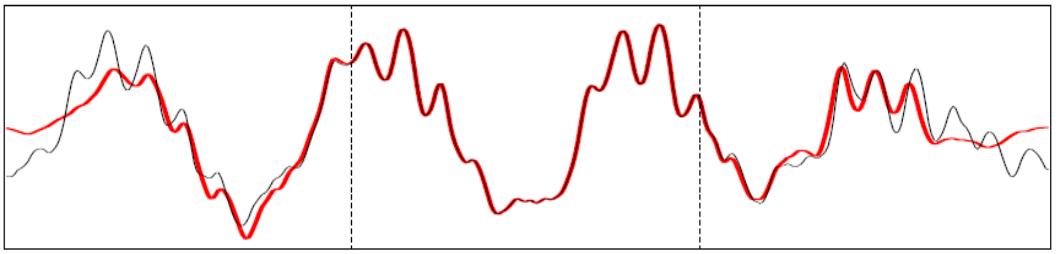
(a) Signal original.



(b) Fenêtre 1 : signal original en noir et après transformation MDCT puis IMDCT en bleu.



(c) Fenêtre 2 : signal original en noir et après transformation MDCT puis IMDCT en bleu.



(d) Signal original en noir et signal reconstruit par recouvrement des deux fenêtres 1 et 2 en rouge.

Figure 5.2 – Illustration du principe TDAC de la MDCT : la portion de signal comprise entre les pointillés obtenue après transformation MDCT puis MDCT inverse est identique à la portion correspondante de signal original.

lieu à l'étape 6 du codeur et à l'étape 14 du décodeur dans notre implémentation), les $W/2$ coefficients fréquentiels $m_t^x[f]$ sont retransformés par MDCT inverse (IMDCT) en un bloc de W échantillons temporels :

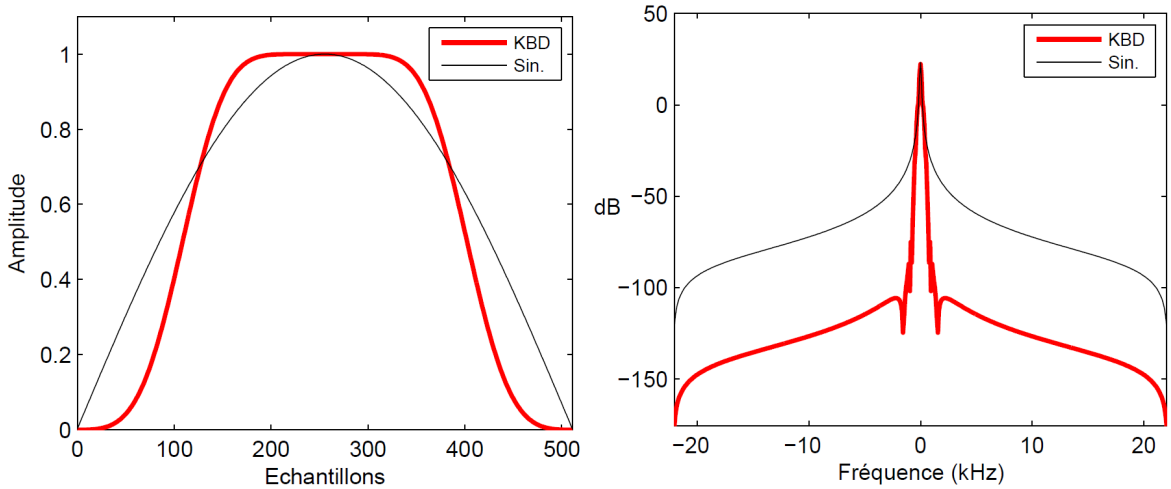
$$\tilde{x}[n+tW/2] = w_s[n]\frac{4}{W}\sum_{f=0}^{W/2-1} m_t^x[f]\cos\left(\frac{2\pi}{W}(n+n_0)\left(f+\frac{1}{2}\right)\right) \tag{5.3}$$

où w_s est la fenêtre de synthèse temporelle (identique à la fenêtre d'analyse dans notre

application). Ce bloc est superposé et additionné aux blocs précédent et suivant pour obtenir le bloc de synthèse final. Pour que la reconstruction du signal soit exacte par MDCT inverse, *i.e.* pour que $MDCT^{-1}(MDCT(x\,[n])) = x\,[n]$, il faut que les fenêtres d'analyse w_a et de synthèse w_s vérifient certaines conditions qui sont, dans le cas où $w_a = w_s$:

$$\begin{cases} w^2[n] + w^2[n + \frac{N}{2}] = & 1 \\ \\ w[n]\ w[\frac{N}{2} - 1 - n] = & w[n + \frac{N}{2}]w[N - 1 - n], \quad \forall n \in [0, \frac{N}{2} - 1] \end{cases} \tag{5.4}$$

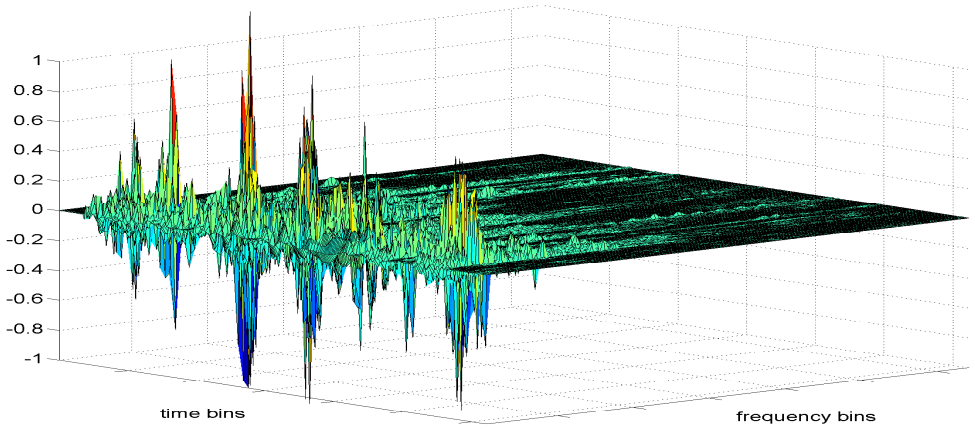
On prend en général des fenêtres symétriques temporellement (*i.e.* vérifiant $w[n] = w[N - 1 - n], \quad \forall n \in [0, N - 1]$). Les fenêtres les plus couramment utilisées sont la fenêtre KBD (Kaiser-Bessel dérivée) et la fenêtre sinusoïdale (arche supérieure de sinusoïde). La fenêtre de type KBD offre une meilleure atténuation spectrale que la fenêtre sinusoïdale comme le montre la Figure 5.3. Nous avons choisi pour w_a une fenêtre KBD de $W = 512$ échantillons correspondant à 32ms de signal pour une fréquence d'échantillonnage de $f_e = 16$kHz et environ 12ms si $f_e = 44.1$kHz (fréquemment utilisée pour la musique, c'est notamment la fréquence d'échantillonnage du format CD audio). Cette taille de fenêtre est choisie pour permettre de suivre la dynamique des signaux que nous étudions. A titre d'illustration, la Figure 5.4 présente une représentation dans le plan temps-fréquence des coefficients MDCT issus de la décomposition de signaux de parole avec les réglages mentionnés. La parcimonie de ces signaux est évidente : la majorité de l'énergie de chaque signal est concentrée dans les basses fréquences alors qu'en hautes fréquences, la plupart des coefficients ont une amplitude proche de zéro.



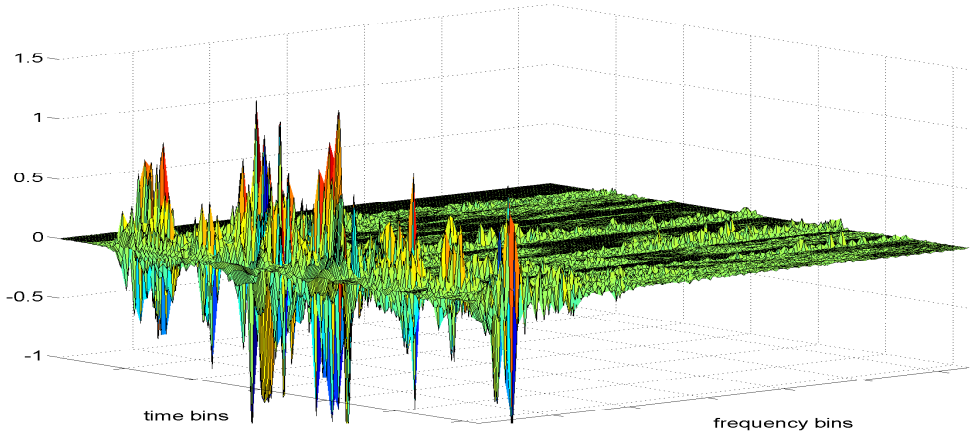
(a) Représentation temporelle.

(b) Représentation fréquentielle.

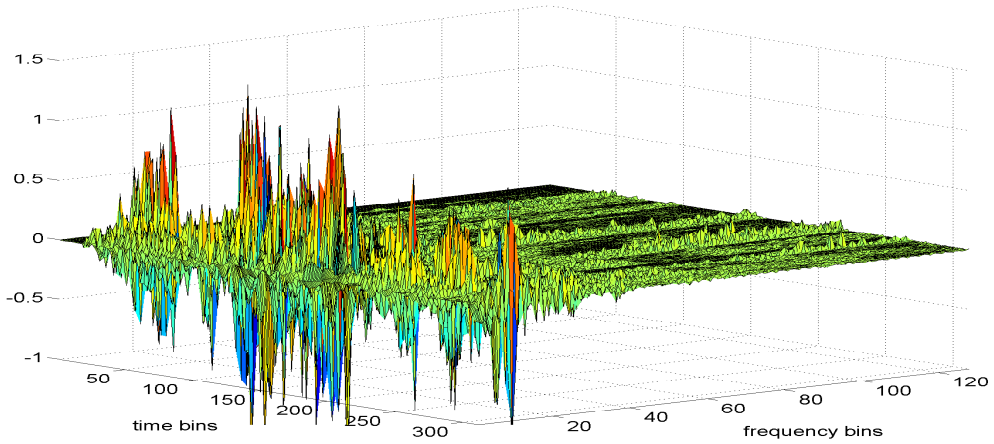
FIGURE 5.3 – Fenêtres KBD et sinusoïdale de 512 échantillons (la représentation fréquentielle correspond à $f_e = 44,1$kHz).



(a) Female



(b) Male



(c) Mix

Figure 5.4 – Représentation MDCT de 2 signaux de parole et de leur mélange.

5.1.2 Groupement moléculaire

Comme nous l'avons abordé à la Section 4.3.3, nous avons décidé de considérer le pavage temps-fréquence non pas à l'échelle de chaque coefficient MDCT, mais à l'échelle d'une *molécule* de coefficients MDCT voisins dans le plan temps-fréquence. Ce groupe de coefficients MDCT voisins est le support élémentaire de l'information de tatouage. Rappelons que ce groupement a pour but de fournir une capacité d'insertion de la watermark significativement plus importante qu'à l'échelle microscopique d'un seul coefficient MDCT, tout en représentant la structure globale du signal avec une granularité que l'on suppose suffisamment précise pour la séparation (nous verrons comment mettre l'accent sur les zones les plus "utiles" au niveau de la perception auditive du signal). L'organisation des coefficients MDCT en molécules est réalisée au niveau du bloc 3 (ou 3' pour les signaux sources) du codeur et au niveau du bloc 10 du décodeur de la Figure 5.1. Le regroupement moléculaire peut être effectué soit par un découpage régulier du plan TF, soit par l'utilisation d'un algorithme de décomposition de type Molecular Matching Pursuit. Cette dernière approche n'étant pas utilisée au cours de ces travaux, sera seulement présentée en Annexe, en prévision de raffinements futurs.

La façon la plus simple de procéder au regroupement moléculaire consiste en un découpage régulier du plan temps-fréquence en molécules rectangulaires adjacentes de taille fixe $F \times T$, où T est le nombre de coefficients MDCT d'une molécule en temps et F est le nombre de coefficients MDCT d'une molécule en fréquence. On parlera alors ici de *pavage moléculaire régulier*. Ceci se formalise de la façon suivante. Considérons un signal x de N échantillons temporels, de matrice MDCT $\mathcal{M}_x$ obtenue grâce à une fenêtre d'analyse de taille W, et qui a donc pour dimension $\frac{W}{2} \times \left(\frac{2N}{W} + 1\right)$ (voir Section 5.1.1). Notons p et q, les coordonnées de chaque molécule de coefficients MDCT dans le plan temps-fréquence : il s'agit donc de canaux fréquentiels et temporels *moléculaires*. Une molécule M^x_{pq} est définie de la façon suivante [3] :

$$M^x_{pq} = \left\{ m^x_t \left[f \right] \right\}_{\substack{f \in P = [(p-1)F, pF-1] \\ t \in Q = [(q-1)T, qT-1]}} \tag{5.5}$$

On obtient alors le pavage TF suivant pour un bloc temporel de signal :

$$\mathcal{M}_x = \begin{pmatrix} M^x_{11} & M^x_{12} & \cdots & M^x_{1L_t} \\ M^x_{21} & M^x_{22} & \cdots & M^x_{2L_t} \\ \vdots & \vdots & \ddots & \vdots \\ M^x_{L_f 1} & M^x_{L_f 2} & \cdots & M^x_{L_f L_t} \end{pmatrix} \tag{5.6}$$

où $L_t = \left(\frac{2N}{W} + 1\right)/T$ (supposé entier) et $L_f = W/2F$ (on choisit F tel que $W/2F$ soit entier).

3. On adopte désormais une notation homogène à celle utilisée sous *Matlab* : le 1^{er} indice d'une matrice adresse une ligne qui correspond pour nous à un canal fréquentiel (ordonnée du plan TF) et le 2^e indice adresse une colonne qui correspond à un canal temporel (abscisse du plan TF).

Un canal moléculaire constitue un découpage du plan temps-fréquence plus grossier que la notion de canal à l'échelle d'un coefficient MDCT. Une molécule possède un seul canal moléculaire qui est la réunion de l'ensemble des canaux de ses coefficients MDCT. La Figure 5.5 schématise un exemple de décomposition du plan temps-fréquence avec un tel pavage, où chaque molécule M_{pq}^{x} du mélange est de dimension 2×4 (2 canaux fréquentiels et 4 canaux temporels). La taille optimale des molécules en fonction de l'information de tatouage à insérer est un des problèmes clé de notre procédé de séparation de source informée, un compromis devant être trouvé entre une capacité suffisamment élevée pour tatouer une information utile au processus de séparation, et une taille moléculaire suffisamment faible pour réduire les risques de superposition des différentes sources. Ce problème a été traité empiriquement et des éléments de résultat seront plus amplement détaillés dans le paragraphe 5.2.

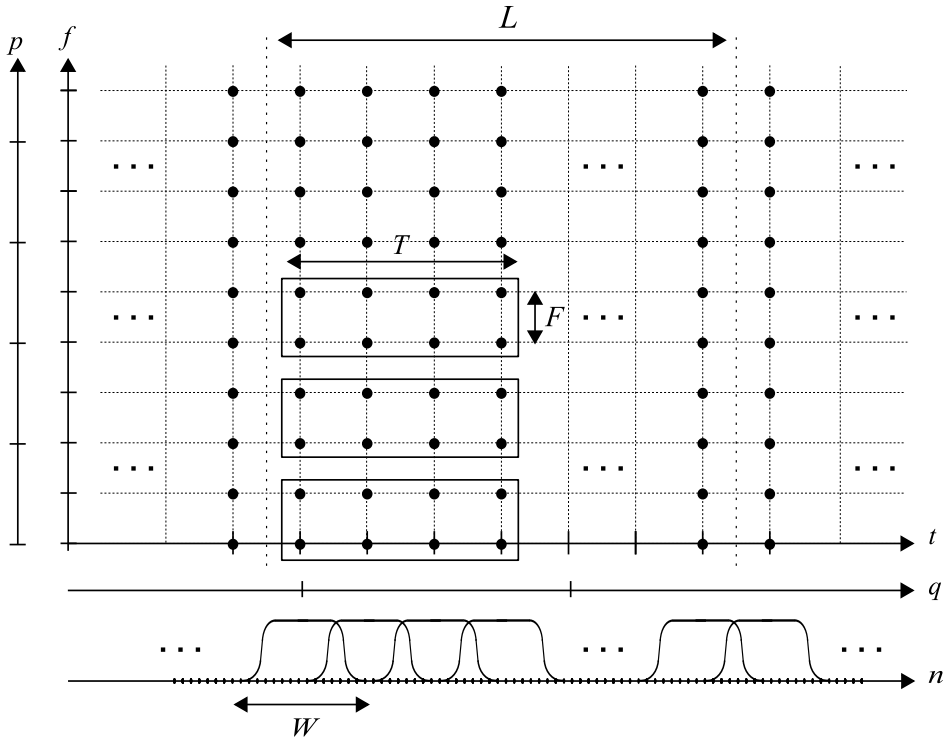


Figure 5.5 – Pavage régulier du plan temps-fréquence.

5.1.3 Tatouage par quantification des coefficients MDCT

5.1.3.1 Deux quantificateurs

Comme annoncé à la Section 4.4, l'insertion de l'information de watermark s'effectue par une modification de l'amplitude des coefficients MDCT du signal de mélange. Nous tirons avantage du fait que les coefficients MDCT du mélange peuvent généralement être quantifiés à une résolution relativement grossière sans pour autant que cela altère la qualité audio du signal de mélange (cette quantification est ajustée, comme nous le développons dans la suite, de manière à être imperceptible). Le tatouage, effectué par une technique de quantification de type QIM, est ici implémenté en deux

temps par l'utilisation de deux quantificateurs scalaires uniformes, notés Q_1 et Q_2. En réalité, seul le quantificateur Q_2 est utilisé en pratique au bloc 5 de la Figure 5.1. Le quantificateur Q_1 sert simplement de référence virtuelle, et permet de simplifier la présentation détaillée de la technique de tatouage.

Rappelons tout d'abord les caractéristiques d'un quantificateur scalaire uniforme (QSU). Deux paramètres définissent un QSU : son facteur d'échelle que nous notons A, relié à l'amplitude maximale du signal à quantifier, et sa résolution notée R, directement reliée au nombre de niveaux de quantification 2^R. L'écart entre deux niveaux de quantification successifs d'un QSU, appelée pas de quantification vaut $\Delta = 2A/2^R$. Enfin, la valeur correspondant au k-*ième* niveau de quantification est $-A+k\times\Delta, k \in \left[0, 2^R - 1\right]$.

Les QSU utilisés dans notre procédé de séparation, Q_1 et Q_2, ont pour résolutions respectives R_1 et R_2, avec $R_2 > R_1$, et possèdent un facteur d'échelle commun, A. Notons que le quantificateur Q_2 peut donc être vu comme un "sur-quantificateur" du quantificateur Q_1, dans la mesure où le pas de quantification Δ_1 est un multiple (de 2) du pas Δ_2. La Figure 5.6 présente le schéma général d'un tatouage par quantification de type QIM, dans lequel les deux QSU décalés de $\Delta_2/2$, l'un grossier et l'autre fin, sont représentés. Notons qu'il ne s'agit pas d'une véritable quantification à proprement parler : l'affectation d'un MDCT à un niveau de Q_2 se fait ici en fonction du résultat de la quantification Q_1 et du symbole à tatouer, et ne donne pas le même résultat qu'une vraie quantification de ce coefficient sur la grille Q_2 au sens du plus proche voisin. Partant du coefficient MDCT quantifié à Q_1, le tatouage de ce coefficient peut le rapprocher ou l'éloigner du coefficient original (dans la limite d'un demi pas de quantification de Q_1). Dans la suite du document, nous gardons cependant le terme "quantification" par simplicité, pour décrire une affectation sur la grille Q_2.

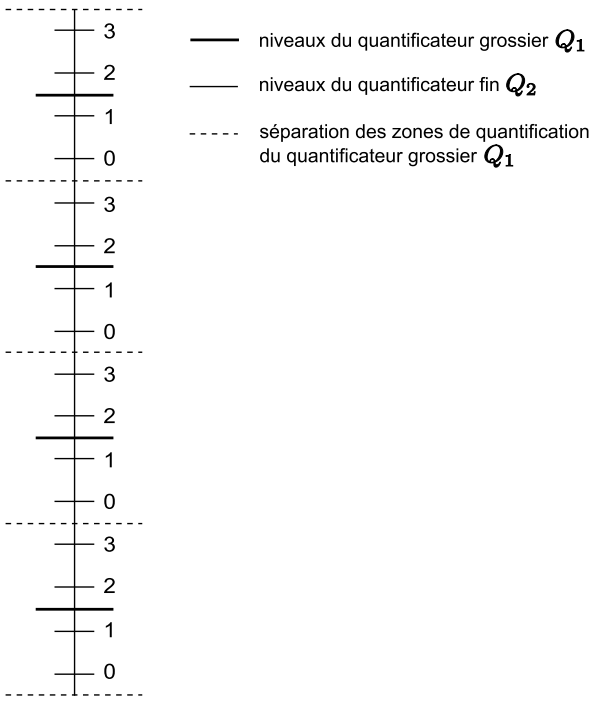


FIGURE 5.6 – Implémentation du tatouage par QIM avec deux QSU.

5.1.3.2 Dépendance en fréquence et en temps

Nous avons vu à la Section 4.3.1 que les coefficients temps-fréquence des signaux audio présentent une large variabilité spectro-temporelle. La gamme d'amplitude des coefficients varie donc en fonction de leur canal fréquentiel et en fonction du temps. Les caractéristiques des quantificateurs gagnent ainsi à être adaptées à cette dynamique, c'est pourquoi on choisit de définir les quantificateurs à chaque canal fréquentiel f. De plus, pour suivre la dynamique temporelle du signal, les quantificateurs utilisés sont mis à jour toutes les L trames temporelles (typiquement toutes les 1.5 secondes). Il s'agit d'un problème classique en quantification, où l'on parle de quantification adaptative. Dans la suite, tous les paramètres des quantificateurs introduits sont fonction à la fois du canal fréquentiel f et de l'indice l du bloc de L fenêtres temporelles. À chaque bloc d'indice l, on met donc à jour $W/2$ quantificateurs $Q_1(l, f)$ et $W/2$ quantificateurs $Q_2(l, f)$, c'est à dire qu'on met à jour les résolutions correspondantes $R_1(l, f)$ et $R_2(l, f)$, ainsi que le facteur d'échelle commun $A(l, f)$. Notons que la capacité d'insertion $C(t, f)$ de l'information de chaque coefficient MDCT, définie au canal fréquentiel f et au canal temporel t, est indépendante du temps sur le bloc l ; autrement dit, $C(t, f) = C(l, f)$ sur le bloc l. $C(l, f)$ est définie par

$$C(l, f) = R_2(l, f) - R_1(l, f) \tag{5.7}$$

On peut donc définir la capacité d'insertion d'une molécule d'un bloc l de la façon suivante :

$$C_M(l, p) = T \times \sum_{f \in P} C(l, f), \;\; P = [(p-1)F, pF - 1] \tag{5.8}$$

La figure 5.12 Section 5.2.3.3 donne un exemple du nombre moyen de bits disponibles pour le tatouage par coefficient MDCT en fonction du canal fréquentiel après une série de tests sur des signaux de musique. Notons la capacité confortable d'insertion d'information dans les plus basses fréquences avec 8 bits par coefficient dans le cas le plus favorable. Une telle capacité par coefficient laisse à penser qu'une quantité importante d'information sur les descripteurs de signaux sources pourra être insérée à l'échelle moléculaire sur le signal de mélange.

5.1.3.3 Détermination de $R_1(l, f)$

Dans le cadre de la SSI, et en particulier avec ce système reposant sur un codage des signaux sources, le volume d'information à insérer sur le signal de mélange est relativement important, d'où la nécessité de maximiser la capacité d'insertion tout en respectant les contraintes intrinsèques au tatouage. L'équation (5.7) montre clairement que pour maximiser $C(l, f)$, $R_1(l, f)$ doit être minimisée et $R_2(l, f)$ doit être maximisée. Comme on va le voir maintenant en détails, la nécessité d'inaudibilité de la watermark permet de définir une borne inférieure pour $R_1(l, f)$, et la nécessité de robustesse du tatouage à la conversion PCM 16-bits introduit une limite supérieure pour $R_2(l, f)$.

Observons d'abord le cas de la résolution $R_1(l, f)$. Il s'agit de minimiser cette résolution sous contrainte d'inaudibilité. Pour assurer cette contrainte, nous utilisons le fait que la précision sur la valeur des coefficients MDCT issus de la décomposition de signaux audio encodés au format PCM 16-bits est trop fine pour l'oreille humaine (avec ce format, les coefficients MDCT ne sont pas quantifiés, ce sont les échantillons temporels qui sont quantifiés sur 16 bits). Il existe une marge disponible pour quantifier ces coefficients avec une précision moindre sans que cela soit préjudiciable à la qualité audio du signal. C'est ce principe qui est largement exploité dans de nombreux algorithmes de compression audio, notamment les codeurs MP3 et AAC des normes MPEG [Brandenburg and Bosi, 1997]. Dans notre étude, c'est précisément dans cette marge que la watermark sera insérée (rappelons que c'est pour cette raison qu'on travaille, dans cette première approche, avec des signaux non compressés). Remarquons aussi que dans notre système complet de séparation, on suppose que si l'inaudibilité est assurée par quantification des MDCT à la résolution $R_1(l, f)$, elle est alors implicitement assurée pour la quantification par $Q_2(l, f)$ à la résolution $R_2(l, f)$, porteuse du watermark et supérieure à $R_1(l, f)$, même si on a vu qu'il ne s'agit pas à proprement parler d'une véritable quantification.

Pour assurer la qualité audio du signal de mélange après quantification sur la grille $Q_1(l, f)$, on assure d'abord une bonne dynamique du signal à l'intérieur de la grille de quantification. Pour cela, le facteur d'échelle $A(l, f)$ est déterminé pour chaque canal fréquentiel f et chaque bloc temporel l à partir de la valeur du coefficient MDCT maximum (en valeur absolue) sur ce bloc :

$$m^x_{max}(l, f) = \max_{t\in[(l-1)L,lL-1]}(|m^x_t\,[f]\,|)$$

Le facteur d'échelle du quantificateur $Q_1(l, f)$ étant adapté au signal de mélange, il est ensuite décidé, dans une première approximation, de fixer la résolution $R_1(l, f)$ à une constante. Ceci permet, dans un premier temps de simplifier la mise en place du processus complet de SSI par codage des signaux sources. Pour fixer cette résolution, des tests d'écoute ont été effectués sur une large base de données de signaux de parole et de musique de genres très variés. Une quantification des coefficients MDCT sur 8 bits est apparue comme n'ayant pas de conséquences audible sur la qualité audio des signaux temporels resynthétisés, avec une mise à jour de $A(l, f)$ effectuée toutes les 1.5 secondes. Des mesures de qualité ont permis de confirmer les tests d'écoute, en fournissant un rapport signal-sur-bruit entre les signaux initiaux et les signaux ayant subis une quantification sur 8 bits dans le domaine temps-fréquence toujours supérieur à 40dB. C'est pourquoi nous choisissons $R_1(l, f) = R_1 = 8$ bits pour chaque canal fréquentiel f et chaque bloc temporel l. Notons que cette résolution est relativement confortable au regard des standards utilisés dans les algorithmes de compression type MPEG. Ceci suggère la possibilité d'une amélioration significative de la capacité d'insertion dans une version plus raffinée, que nous introduirons à la Section 5.3. Une résolution $R_1(l, f)$ sera alors adaptée au contenu du signal hôte et déterminée pour chaque trame temporelle du signal grâce à un modèle psychoacoustique. Ce dernier

permet d'augmenter la capacité $C(l, f)$ en particulier en hautes fréquences, où l'oreille humaine est moins sensible et où, de fait, une résolution $R_1(l, f)$ inférieure à 8 bits peut permettre d'assurer la qualité audio du signal de mélange après tatouage.

5.1.3.4 Détermination de $R_2(l, f)$

En ce qui concerne la résolution $R_2(l, f)$, correspondant à la quantification de tatouage, elle doit être la plus élevée possible pour maximiser la capacité d'insertion d'information $C(l, f)$, sous contrainte d'assurer la robustesse à la conversion du signal tatoué au format CD-audio. Comme nous l'avons déjà mentionné à la Section 4.4.3, la conversion au format CD entraîne, dans le domaine spectral, un bruit susceptible de modifier le tatouage des coefficients MDCT. C'est pourquoi il faut veiller à ce que la quantification Q_2 au bloc 11 du décodeur fournisse le même résultat que celui obtenu après la quantification Q_2 au bloc 5 du codeur, *i.e.* avant la quantification 16 bits correspondant à la conversion PCM 16-bits.

Dans le domaine temporel, rappelons que la conversion au format PCM 16-bits correspond à une quantification du signal temporel grâce à un QSU de résolution 16 bits (bloc 7 de la Figure 5.1), ce qui se traduit par l'ajout sur les échantillons d'un bruit uniforme d'écart-type[4] $\sigma_{16} = \sqrt{\frac{\Delta^2}{12}} = \frac{2^{-15}}{\sqrt{12}} \approx 8,8.10^{-6}$. La MDCT étant une transformation linéaire orthonormée, ce bruit additif se traduit par un bruit additif dans le domaine des MDCT. Le théorème central limite généralisé assure le comportement gaussien de ce bruit dans le domaine fréquentiel [Feller, 1971], comportement que nous avons observé dans la pratique[5]. On peut montrer de plus que l'écart-type du bruit spectral est indépendant du canal fréquentiel f (ainsi que de t), et est le même que l'écart-type du bruit de quantification uniforme dans le domaine temporel, c'est à dire σ_{16} : en résumé, la MDCT normalisée transforme un bruit uniforme temporel en un bruit TF gaussien, blanc, et de même variance [Pinel et al., 2010a]. Pour que la watermark ne soit pas affectée par la conversion du signal tatoué au format CD-audio, le pas de quantification de $Q_2(l, f))$, $\Delta_2(l, f)$, doit respecter la contrainte :

$$|\Delta_{16}| < \frac{\Delta_2(l, f)}{2} \tag{5.9}$$

où Δ_{16} est la variation maximale du bruit spectral[6]. Or le nombre de niveaux $2^{R_2(l,f)}$

4. Rappelons que la quantification scalaire uniforme d'un signal x par un QSU de pas de quantification Δ induit une distorsion de x de variance $\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} x^2 dx$

5. Ce type d'observation peut être réalisé de plusieurs façons : on peut calculer directement une décomposition MDCT d'un bruit de quantification uniforme simulé. On peut également réaliser la quantification d'un signal de test. Si ce signal est déjà quantifié sur 16 bits, comme c'est le cas des signaux extraits de CD audio, on peut "déquantifier" au préalable ces signaux en leur appliquant la chaîne de traitement suivante : décomposition MDCT, très légère modification des coefficients, et resynthèse du signal par IMDCT. En pratique nous avons fait les deux types d'expérimentation et elles conduisent au même résultat.

6. σ_{16} étant indépendant de la fréquence f, il en va de même de Δ_{16}. Notons que comme $\Delta_2(l, f) < \Delta_1(l, f)$, l'équation (5.9) implique que la quantification Q_1 est aussi robuste au bruit de quantification 16 bits.

de la quantification de tatouage est liée au pas de quantification de la watermark par la relation :

$$\Delta_2(l,f) = \frac{2\ A(l,f)}{2^{R_2(l,f)}} \tag{5.10}$$

De plus, le bruit de quantification 16 bits ayant dans le domaine des MDCT un comportement gaussien, on choisit de prendre $|\Delta_{16}| < 4\ \sigma_{16}$. En utilisant les équations 5.9 et 5.10, on en déduit la contrainte suivante sur le nombre de niveaux $2^{R_2(l,f)}$ de la quantification de tatouage :

$$2^{R_2(l,f)} < \frac{A(l,f)}{4\ \sigma_{16}} \tag{5.11}$$

Soit

$$R_2(l,f) < \lfloor \log_2(\frac{A(l,f)}{4\ \sigma_{16}}) \rfloor \tag{5.12}$$

où $\lfloor . \rfloor$ désigne la partie entière. De manière à maximiser la capacité $C(l,f)$, on choisit :

$$R_2(l,f) = \lfloor \log_2(\frac{A(l,f)}{4\ \sigma_{16}}) \rfloor \tag{5.13}$$

5.1.3.5 Fermeture de la boucle codage-décodage

Pour que notre procédé de séparation par codage des descripteurs des signaux sources puisse fonctionner, il est indispensable que les grilles de quantification $Q_1(l,f)$ et $Q_2(l,f)$ soient exactement les mêmes au codeur et au décodeur. En effet, la valeur d'un coefficient MDCT affectée à un niveau de $Q_2(l,f)$ à l'encodeur en fonction d'un certain symbole à tatouer (après être passée par un niveau de référence $Q_1(l,f)$), doit être reconnue exactement au même niveau de la même grille $Q_2(l,f)$ au décodeur (après être passée par le même niveau de référence de $Q_1(l,f)$ au décodeur) pour une bonne extraction du symbole tatoué. Or, nous venons de voir que les quantificateurs $Q_1(l,f)$ et $Q_2(l,f)$ utilisés dépendent des caractéristiques du signal traité, et en particulier, qu'elles peuvent varier d'un macro-bloc à l'autre. Par conséquent, les caractéristiques des quantificateurs déterminés au codeur sur un macro-bloc temporel de L trames doivent pouvoir être retrouvées automatiquement sur le même macro-bloc au décodeur. Ceci rappelle le problème de codage en boucle fermée (vs. codage en boucle ouverte) rencontré en compression. En d'autres termes, il s'agit de s'assurer que le tatouage du signal n'influe pas sur la détermination des caractéristiques de $Q_1(l,f)$ et $Q_2(l,f)$.

Contrairement aux algorithmes de codage audio, dans la méthode de SSI que nous proposons dans ce chapitre, les paramètres des quantificateurs du signal de mélange ne sont pas transmis au décodeur, bien qu'ils soient indispensables à l'extraction des descripteurs de sources au décodeur. En revanche, on transmet les échantillons du signal de mélange tatoué, et nous montrons ci-après que les paramètres des quantificateurs peuvent être retrouvés à partir de ces échantillons, et ce malgré la modification des

coefficients MDCT due au tatouage lui-même. Ceci permet d'allouer l'ensemble de la capacité disponible au codage des descripteurs des signaux sources.

Comme la résolution du quantificateur $Q_2(l,f)$ ne varie qu'en fonction du facteur d'échelle $A(l,f)$ (équation (5.13)), la fermeture de la boucle codage-décodage pour $R_2(l,f)$, *i.e.* le fait que $Q_2(l,f)$ soit retrouvé au décodeur identique au codeur, est assurée par la détermination au décodeur de ce seul facteur $A(l,f)$ (cf. Section 5.1.3.4). Cependant, la valeur de $m^x_{max}(l,f)$ qui permet d'obtenir au décodeur la valeur de $A(l,f)$ est modifiée par le processus de tatouage et de conversion au format CD-audio, comme tout coefficient MDCT. Une solution pour résoudre ce problème consiste à quantifier les facteurs d'échelles $A(l,f)$ grâce à un quantificateur $Q_A(f)$ fixé une fois pour toutes (indépendant du signal), et identique au codeur (bloc 2) et au décodeur (bloc 9). C'est la valeur de $m^x_{max}(l,f)$ quantifiée sur $Q_A(f)$ à la valeur directement supérieure, toujours notée $A(l,f)$ pour ne pas alourdir les notations, qui est utilisée comme facteur d'échelle des quantificateurs $Q_1(l,f)$ et $Q_2(l,f)$.

Pour qu'une quantification de $m^x_{max}(l,f)$ sur une grille de quantification $Q_A(f)$ soit robuste au tatouage, il faut, dans un premier temps, que la résolution $R_A(f)$ de $Q_A(f)$ soit plus faible que celle de $Q_1(l,f)$. Cependant, il faut également que $R_A(f)$ soit suffisamment élevée pour assurer la construction d'un ensemble de quantificateurs $Q_1(l,f)$ et $Q_2(l,f)$ de bonne qualité (une résolution trop faible pourrait avoir des conséquences sur la qualité d'écoute des signaux en ne permettant pas un bon suivi de la dynamique des MDCT par les quantificateurs $Q_1(l,f)$ et $Q_2(l,f)$). Pour assurer ce compromis, nous avons donc choisi de fixer la résolution $R_A(f)$ à 6 bits, au regard des 8 bits de la grille $Q_1(l,f)$. Tout comme R_1, cette résolution est indépendante de la fréquence, et on la note par la suite R_A. Notons $A_{max}(f)$ le facteur d'échelle de $Q_A(f)$. Ce facteur représente l'amplitude maximale des coefficients MDCT maximum $m^x_{max}(l,f)$ pour l'ensemble de tous les macro-blocs de tous les signaux traités : il s'agit en quelque sorte d'un maximum des maxima. En pratique, ce paramètre est déterminé (pour chaque fréquence f) à partir du calcul des MDCT sur une très grande base de données de sons de musique et de parole[7]. Si la résolution $R_A = 6$ bits assure la robustesse au tatouage des coefficients MDCT maximum dans la plupart des cas, il existe cependant un cas de figure qui pose problème : c'est celui où le coefficient $m^x_{max}(l,f)$ est juste au dessus d'un niveau de la grille $Q_A(f)$. Le schéma 5.7 illustre ce cas de figure. Sur cet exemple, le coefficient $M_{max}(f)$ est quantifié au codeur par $Q_A(f)$ au niveau $A_2(f)$. Le tatouage peut entraîner une modification de la valeur du coefficient $(m^x_{max}(f))_{R_1}$ (coefficient $m^x_{max}(f)$ quantifié sur la grille $Q_1(l,f)$) d'une valeur maximale de $\pm\frac{\Delta_1(l,f)}{2}$, pour donner $(m^x_{max}(f))_{R_2}$[8]. Si le coefficient $m^x_{max}(f)$ est

7. En réalité, pour plus de sécurité nous prenons $A_{max}(f)$ égal à deux fois la valeur déterminée expérimentalement. A noter que le spectre des divers types de signaux peut être très différent d'un signal à l'autre, et pour quantifier de manière adéquate les coefficients MDCT maximum, il faut déterminer $A_{max}(f)$ sur des signaux ayant une bonne représentativité des signaux à traiter.

8. Remarquons que la connaissance de la grille $Q_2(l,f)$ n'est pas nécessaire pour établir la modification maximale introduite par le tatouage. En effet la grille $Q_2(l,f)$ est déterminée de manière à ne pas modifier la valeur des coefficients MDCT tatoués sur la grille de quantification $Q_1(l,f)$, condition

trop proche [9] du niveau de la grille $Q_A(f)$ auquel il est immédiatement supérieur ($A_1(f)$ sur le schéma 5.7), alors le tatouage peut modifier sa valeur de sorte qu'elle devienne inférieure à ce niveau. Au décodeur, cette valeur tatouée est alors quantifiée par $Q_A(f)$ à la valeur immédiatement supérieure $A_1(f)$ au lieu de $A_2(f)$. On a donc dans ce cas une mauvaise "transmission" du facteur d'échelle. Pour remédier à ce problème, nous avons choisi de modifier la valeur de $M_{max}(f)$ au codeur si ce coefficient est trop proche de $A_1(f)$. On remplace alors $m^x_{max}(f)$ par $m^x_{max}(f) + \frac{\Delta_1(l,f)}{2}$. Notons que seul le maximum des coefficients MDCT est modifié sur tout le canal f du macro-bloc traité, et que cette modification est sans incidence sur la qualité audio du signal puisqu'elle est de l'ordre de grandeur de la quantification $Q_1(l,f)$ qui ne détériore pas le signal. Avec cette précaution, l'exacte correspondance entre les facteurs d'échelle au codeur et au décodeur est alors assurée dans tous les cas. Notons que pour garantir la fermeture de la boucle codage-décodage, codeur et décodeur sont supposés parfaitement synchronisés, c'est-à-dire qu'on sait se positionner dans le signal aussi bien au codeur qu'au décodeur, et en particulier on connaît les frontières des macro-blocs. Ceci ne représente pas une contrainte très forte : on dispose de tous les échantillons de signal au codeur comme au décodeur. On sait donc quand mettre à jour les quantificateurs au décodeur, et assurer que le codage et le décodage sont effectués sur les mêmes blocs temporels.

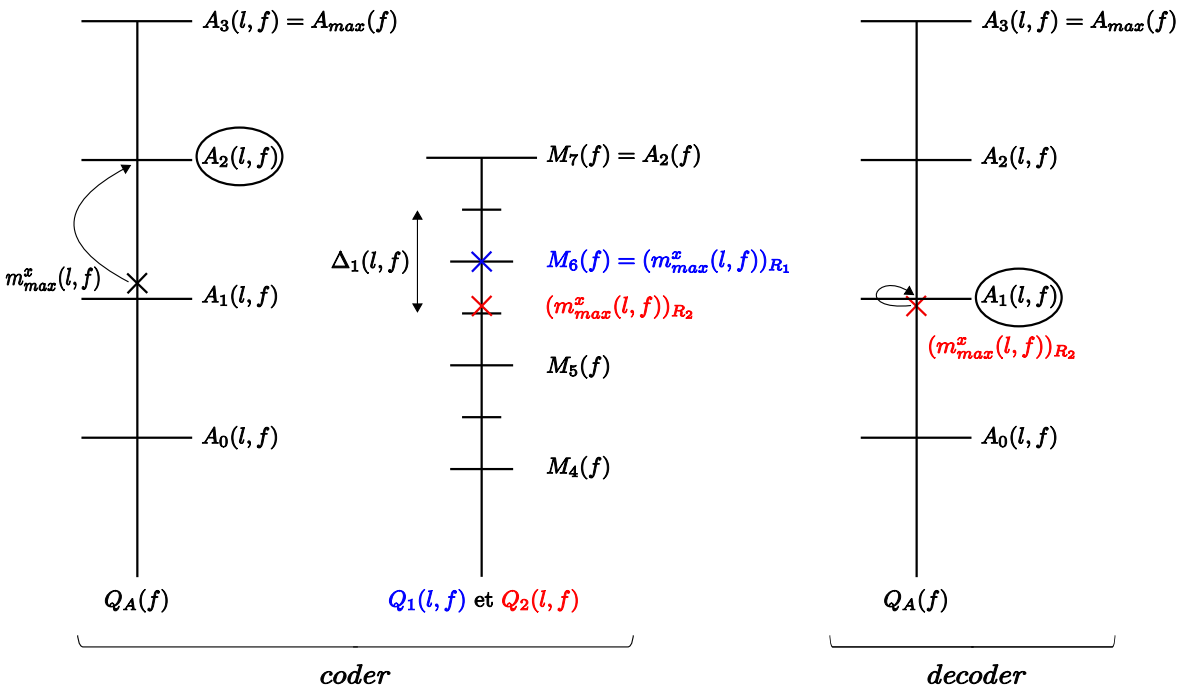


FIGURE 5.7 – Conservation du facteur d'échelle $A(f)$ au décodeur.

qui n'est vérifiée que si la modification de la valeur du MDCT tatoué est inférieure à $\pm\frac{\Delta_1(l,f)}{2}$.

9. A une distance inférieure à $\frac{\Delta_1(l,f)}{2}$ de ce niveau.

5.1.4 Descripteurs des signaux sources et estimation des sources associée

Comme nous l'avons vu à la Section 4.3.5, les descripteurs des signaux permettent de caractériser dans le plan TF les molécules de coefficients MDCT de chacun des signaux sources à séparer de manière à aider à leur séparation. Plusieurs types de descripteurs ont été considérés dans cette étude et leur combinaison permet d'accroître la précision avec laquelle chaque signal source est décrit. Ces descripteurs et la façon dont ils sont utilisés pour la séparation sont détaillés dans les sections 5.1.4.1 et 5.1.4.2. Le problème de séparation de sources informée est multi-paramètres et ces paramètres sont interdépendants : la taille des molécules, le nombre de descripteurs, leur nature, et la précision avec laquelle ils sont codés dépendent de la capacité d'insertion d'une molécule, elle-même fonction de la capacité d'insertion de chacun de ses coefficients MDCT pour un canal fréquentiel f et un bloc temporel l (cf équation (5.8)). Les dimensions d'une molécule influent donc directement sur la capacité disponible pour insérer de l'information relative aux signaux sources, et c'est cette capacité disponible qui dicte quels sont les descripteurs utilisés et la façon dont sont alloués les bits entre ces différents descripteurs. L'allocation de bits précise que nous avons utilisée en fonction de la capacité d'insertion de chaque molécule sera détaillée à la Section 5.1.4.4.

5.1.4.1 Gain moléculaire

Définition du descripteur gain

Nous nous plaçons ici dans le cas d'un pavage régulier du plan TF. La décomposition MDCT, de par sa capacité à concentrer l'énergie du signal sur un faible nombre de coefficients fournit une représentation compacte du signal avec un faible nombre de molécules énergétiques. L'énergie est donc une caractéristique essentielle des molécules de coefficients MDCT, et du fait du domaine de décomposition choisi (transformation MDCT linéaire), on suppose que cette énergie est aussi un facteur discriminant pour caractériser la contribution des sources dans le mélange.

Le premier descripteur des sources choisi est donc basé sur l'énergie de chacune des molécules des signaux sources. Comme on cherche à caractériser la contribution relative des sources dans le mélange, il s'agit plus précisément du rapport entre l'énergie de chaque molécule d'un signal source et l'énergie de la molécule correspondante[10] du mélange. Ainsi, en utilisant les mêmes notations qu'à la Section 5.1.2, l'énergie de la molécule de coordonnées moléculaires (p, q) du signal s_i est définie par :

$$E_{s_i}(p,q) = \sum_{(f,t)\in\{P\times Q\}} |m_t^{s_i}[f]|^2 \tag{5.14}$$

où $P = [(p-1)F, pF-1]$ et $Q = [(q-1)T, qT-1]$. L'information de contribution énergétique relative, ou *gain*, d'une source s_i par rapport au mélange x, pour la molécule de coordonnées (p, q), est donnée par le rapport des énergies suivant :

10. C'est-à-dire de mêmes coordonnées TF.

$$E_{s_i/x}(p,q) = \frac{E_{s_i}(p,q)}{E_x(p,q)} = \frac{\sum\limits_{(f,t)\in\{P\times Q\}} |m_t^{s_i}[f]|^2}{\sum\limits_{(f,t)\in\{P\times Q\}} |m_t^{x}[f]|^2} \tag{5.15}$$

Quantification du descripteur gain

Au décodeur, l'information de gain utilisée pour reconstruire la molécule $\hat{M}_{pq}^{s_i}$, estimée de la molécule de coordonnées (p,q) de la source s_i n'est pas celle calculée au codeur mais une version quantifiée de cette dernière. En effet cette information doit être transmise au décodeur via le tatouage, et doit donc être codée sur un nombre de bits limité (voir la section 4.4). Pour cela on utilise un ensemble de QSU. Les résolutions des quantificateurs des descripteurs sont déterminées grâce à des tables d'allocation de bits ; ce problème sera traité à la Section 5.1.4.4. Il est ici possible de vérifier sur un exemple simple la nécessité d'un codage à l'échelle d'une molécule, et non à l'échelle des coefficients MDCT individuellement. Prenons un exemple concret de codage de la contribution énergétique d'un coefficient MDCT d'une source s_1 par rapport à celle de son mélange linéaire instantané avec s_2, $x = \alpha s_1 + \beta s_2$. La valeur de $E_{s_1/x}(t,f)$ pour le coefficient MDCT situé à la fréquence f et sur la trame t peut prendre une valeur entre 0 et 400% (ordre de grandeur observé sur des signaux test [11]), et de fait, la quantification de cette contribution énergétique à 5% près nécessite 7 bits. Il aurait été impossible d'insérer 7 bits d'information sur certains coefficients MDCT seuls alors que c'est tout à fait possible sur toute molécule à 8 coefficients (2×4) par exemple, étant donné l'ordre de grandeur de capacité d'insertion par coefficient MDCT (cf Figure 5.12). Par la suite, une série de quantificateurs de résolutions différentes sera testée dans la partie expérimentations en liaison avec le processus d'allocation de bits.

Notons que les facteurs d'échelle des quantificateurs des descripteurs sont eux-mêmes quantifiés grâce à un QSU de résolution 6 bits, fixé une fois pour toutes, et connu à la fois au codeur et au décodeur. Contrairement aux quantificateurs des signaux de mélange, ces facteurs d'échelle sont transmis par tatouage. En effet, ils ne peuvent pas être retrouvés au décodeur à partir des coefficients MDCT du mélange car ils caractérisent les signaux sources non mixés, et non le signal de mélange. Le coût de cette information supplémentaire est supposé très faible en comparaison de la capacité allouée au codage des descripteurs. En effet, ces paramètres sont mis à jour moins souvent que ne le sont les quantificateurs du signal de mélange (un rapport de mise à jour de l'ordre de 1 pour 2 à 1 pour 10, en fonction de la taille du L du bloc temporel sur lequel est effectué l'ensemble du traitement, est observé entre la mise à jour des facteurs d'échelle $A(l,f)$ et la mise à jour des quantificateurs des descripteurs).

11. Lors du mélange, il arrive que les contributions des différentes sources en opposition de phase conduisent à un signal de mélange dont l'amplitude est inférieure à celle des signaux sources, d'où une participation énergétique supérieure à 100% dans un tel cas.

5.1.4.2 Information de forme

Principe

Un autre élément important de la structure d'une molécule est l'organisation relative des amplitudes des coefficients MDCT les uns par rapport aux autres au sein de cette molécule. C'est ce que l'on appelle la *forme* d'une molécule. Dans l'éventualité où une source donnée est la seule source présente dans une certaine zone du plan TF, son estimée est reconstruite par la molécule de mélange dont la forme correspond exactement à la forme de la molécule du signal source, d'où une parfaite reconstruction. Cependant, dans le cas général de sources superposées dans le plan temps-fréquence avec des contributions comparables, cette superposition (qui est rigoureusement une addition dans le cas de mélanges linéaires instantanés) fait que la forme d'une molécule de mélange est généralement significativement différente de la forme de la molécule correspondante de chacune des sources composant le mélange. Dans le pire cas, les amplitudes des coefficients des molécules de différentes sources peuvent s'annihiler si les signaux sont localement en opposition de phase. Il en résulte, même après pondération de la molécule de mélange par l'énergie relative de la molécule de source associée, que la forme de la molécule de l'estimée de cette source peut être très différente de la molécule source originale, et par conséquent le signal reconstruit peut être très différent de l'original dans la zone TF considérée.

Pour pallier à ce problème, nous proposons de tatouer, lorsque c'est possible (*i.e.* si $C_M(l,p)$ est suffisamment importante), une information relative à la forme des molécules des sources sur chaque molécule de mélange. Cette information de forme (appliquée sur des molécules normalisées comme nous le verrons par la suite) constitue un second descripteur susceptible d'aider à la séparation des signaux sources à partir du mélange, en plus de la contribution énergétique des sources à l'échelle des molécules (qui permettra de remettre à l'échelle la molécule normalisée). Le tatouage de ces deux informations affine la précision avec laquelle un signal source est décrit au niveau moléculaire. L'information de forme doit être vue comme un raffinement de la simple contribution énergétique : elle va toujours de pair avec cette dernière et ne peut être codée seule pour décrire une molécule de signal source.

Relation avec la séparation à base de modèles

Comme nous l'avons vu dans le Chapitre 2, les résultats les plus encourageants obtenus jusqu'alors en séparation de sources en configuration sous-déterminée sont généralement obtenus à partir de méthodes utilisant des modèles de sources *appris* sur des bases de données de signaux d'apprentissage. Des dictionnaires de données (atomes de décomposition sur une base de coefficients par exemple) adaptés à chacune des sources à séparer sont de plus en plus utilisés en séparation de sources pour améliorer les performances de reconstruction des sources qui peuvent, au sein d'un mélange, être de type très différents. Il est clair que la projection d'un mélange contenant un signal de parole sur une base de coefficients appris sur des modèles de signaux de parole

isolés permet de mieux séparer ce signal de parole du mélange que si le mélange est projeté sur une base de coefficients génériques (atomes de Gabor ou coefficients MDCT du mélange par exemple). Cette observation est cohérente avec notre proposition de tatouer sur chaque molécule de mélange, en plus de l'information de gain de chaque molécule de signal source correspondante du plan TF, sa forme, si la place disponible sur la molécule de mélange le permet. Nous allons voir dans les sous-sections suivantes qu'on peut utiliser des dictionnaires de forme appris sur des sources spécifiques non mixées. Ainsi pour chaque type de source, instrument dans le cas de signaux musicaux ou type de locuteur dans le cas de signaux de parole, et pour chaque zone caractéristique du plan temps-fréquence, un dictionnaire de forme est établi à partir de signaux sources d'apprentissage du même type. Ces formes constituent alors un modèle des sources à séparer. De ce point de vue, on peut dire que la méthode proposée dans cette partie de la thèse rentre bien dans la catégorie (efficace) des méthodes à base de modèles.

Utilisation de dictionnaires de forme et codage d'une molécule

La forme d'une molécule décrit la répartition de l'amplitude des coefficients MDCT les uns par rapport aux autres. Coder cette forme de manière idéale revient donc à coder l'amplitude de chaque coefficient MDCT la composant. Or, coder avec précision chaque coefficient MDCT requerrait un nombre de bits très élevé pour une molécule particulièrement énergétique, et donc supérieur à sa capacité d'insertion par tatouage. C'est pour cette raison que nous utilisons des *dictionnaires de formes prototypes* préalablement appris sur des signaux de même nature que les signaux dont on souhaite coder la forme. Il s'agit en fait des principes de base des techniques dites de quantification vectorielle [12] très largement exploitées en compression des signaux [Gray and Gersho, 1992]. De fait, l'approche que nous envisageons en codant la forme d'une molécule s'apparente à une approche de type codage des sources, et même codage source-canal étant donné que la précision du codage des sources dépend de la capacité du signal hôte, canal de transmission de l'information. Un dictionnaire est constitué de formes prototypes les plus représentatives possibles des formes de molécules pouvant être rencontrées pour un type de signal audio [13]. Le descripteur de la forme d'une molécule sera donc la molécule prototype du dictionnaire qui lui "ressemble" le plus, c'est à dire la plus proche au sens d'une certaine distance. L'utilisation de dictionnaires de forme permet de réduire considérablement le coût du codage d'une telle information puisque seul l'indice de la molécule prototype retenue dans le dictionnaire constitue l'information de forme tatouée sur le mélange (et non l'amplitude de tous les coefficients de la molécule). La richesse des formes d'un dictionnaire croît évidemment avec sa taille. Selon cette taille, le coût de codage de l'indice peut varier de façon importante, mais dans tous les cas, ce coût de codage reste inférieur à celui du codage séparé de chaque

12. Avec un réordonnancement arbitraire des coefficients MDCT, une molécule est bien équivalente à un vecteur de coefficients.

13. Le type du signal peut être une voix d'homme/de femme ou encore un instrument particulier selon que l'on considère des signaux de parole ou de musique.

coefficient MDCT d'une molécule pour une qualité équivalente.

Dans notre application, l'utilisation de l'information de forme et la précision de son codage dépend de la place disponible pour insérer le tatouage. C'est pourquoi nous utilisons un ensemble de dictionnaires de tailles différentes (ces tailles seront précisées dans le paragraphe **Réalisation des dictionnaires** ci-après). Cette banque de dictionnaires sera utilisée de façon adaptative suivant la capacité d'insertion de chaque molécule. Plus la capacité de tatouage est grande, plus le dictionnaire de formes choisi est volumineux et meilleure est l'approximation de la forme d'une molécule de signal source. Par ailleurs, nous avons vu que le comportement des coefficients MDCT et la capacité d'insertion des molécules dépendent du canal fréquentiel. L'ensemble de dictionnaires de taille variable est donc défini pour chaque canal fréquentiel moléculaire. Un dictionnaire d'un canal fréquentiel donné est donc représentatif des molécules de ce canal.

Pour des raisons d'efficacité de codage, c'est-à-dire l'optimisation du rapport qualité/coût binaire, on cherche toujours à utiliser des dictionnaires les plus représentatifs possibles pour un nombre de prototypes donné. Pour cette raison, si des molécules diffèrent d'un facteur additif et/ou multiplicatif, elles doivent être associées au même prototype. Pour cela, on considère des formes normalisées selon ces facteurs additif et multiplicatif. En d'autres termes, on effectue une opération de centrage (soustraction de la moyenne) et de réduction (division par l'écart-type) des molécules avant encodage par leur prototype. La moyenne d'une molécule $M_{pq}^{s_i}$ du signal source s_i de coordonnées (p, q) est définie par (avec les mêmes notations qu'utilisées précédemment) :

$$\mu_{pq}^{s_i} = \frac{\sum\limits_{(f,t)\in\{P\times Q\}} m_t^{s_i}[f]}{T \times F} \tag{5.16}$$

et son écart-type est défini par :

$$\sigma_{pq}^{s_i} = \sqrt{\frac{\sum\limits_{(f,t)\in\{P\times Q\}} (m_t^{s_i}[f] - \mu_{pq}^{s_i})^2}{T \times F}} \tag{5.17}$$

On note $N_{pq}^{S_i}$ la molécule normalisée donnée par :

$$N_{pq}^{s_i} = \frac{M_{pq}^{s_i} - \mu_{pq}^{s_i}}{\sigma_{pq}^{s_i}} \tag{5.18}$$

C'est cette molécule $N_{pq}^{s_i}$ dont on cherche le plus fidèle substitut parmi un dictionnaire lors du codage de la forme de $M_{pq}^{s_i}$ au bloc 4 de la figure 5.1. La moyenne et l'écart-type sont encodés séparément par quantification scalaire. Ce processus est une technique bien identifiée en codage et assure un rapport qualité sur coût du codage optimal (on parle en anglais de "Mean-Gain-Shape quantization", voir par exemple [Baker, 1984] et [Oehler and Gray, 1993] ; l'écart-type est ici assimilé à un gain multiplicatif, et dans la suite du document, on utilise le terme *gain* pour désigner cet écart-type lorsqu'il n'y a pas d'ambiguïté avec le gain de l'équation (5.15)). Le nombre de descripteurs des signaux sources à encoder est donc ici de trois : pour chaque molécule d'un signal

source, sa moyenne, son écart-type et sa forme constituent le message inséré sur la molécule du mélange située aux mêmes coordonnées du plan temps-fréquence si la place disponible est suffisante (la distribution de la ressource binaire de tatouage entre ces trois descripteurs fait partie de l'allocation de bits discutée à la section 5.1.4.4). Les descripteurs $\mu_{pq}^{s_i}$ et $\sigma_{pq}^{s_i}$ sont encodés séparément, comme par exemple dans [Oehler and Gray, 1993], par des QSU similaires à ceux utilisés pour encoder le descripteur de gain seul (cf 5.1.4.1) [14]. On note $\check{\mu}_{pq}^{s_i}$ et $\check{\sigma}_{pq}^{s_i}$ leur version quantifiée. De même, tout comme pour le gain, les paramètres des quantificateurs de ces descripteurs sont transmis par tatouage avec une mise à jour et un coût très faible devant le débit de codage des signaux.

Réalisation des dictionnaires

Les dictionnaires sont déterminés à partir d'une base de données de signaux d'apprentissage. Les dictionnaires de parole sont appris sur des signaux de parole de la base de données TIMIT [Fisher et al., 1986] comportant des voix de femme et des voix d'homme. Ces signaux représentent un panel varié (en anglo-américain) de dialectes et d'accents. Les dictionnaires de musique sont construits à partir de pistes de trois instruments, une *guitare basse*, une *batterie*, et un *piano*, et à partir de pistes d'une *voie chantée* de femme. Ces dictionnaires sont obtenus par un algorithme dérivé du célèbre algorithme de quantification vectorielle de Linde, Buzo et Gray (LBG) [Linde et al., 1980]. Dans notre cas, cependant, il s'agit d'une quantification matricielle et non vectorielle où les éléments des dictionnaires de forme sont des molécules de coefficients MDCT.

Étant donné que les molécules des dictionnaires sont centrées réduites, l'apprentissage se fait sur des molécules elles-mêmes centrées et réduites avant l'application de l'algorithme. De plus, comme on détermine un ensemble de dictionnaire pour les différents canaux fréquentiels moléculaires, pour apprendre un dictionnaire sur un canal donné, on utilise les molécules d'apprentissage de ce canal.

Le principe général de l'algorithme LBG est donné ci-dessous. Pour l'appliquer, on se munit d'une distance dans l'espace des molécules. Cette distance est la distance euclidienne sur les coefficients MDCT des molécules.

1. On se donne un dictionnaire initial $\mathcal{C}'$ composé de N_c molécules.
2. On forme N_c classes à partir de la base de molécules d'apprentissage : chaque molécule m est placée dans la classe S_i si le i-ième mot du dictionnaire initial est le plus proche de m.
3. On calcule les barycentres de chaque nouvelle classe : ce sont les nouveaux prototypes [15].

14. De par la nature de ces descripteurs, des quantificateurs symétriques sont utilisés pour la moyenne, alors que des quantificateurs à valeurs uniquement positives sont utilisés pour le gain.

15. Lors de ce calcul, il faut penser à normaliser la nouvelle molécule prototype, car un barycentre de molécules normalisées n'est pas implicitement normalisé.

4. On itère 2 et 3 et on arrête quand la distorsion totale i.e. la distance moyenne entre les molécules d'apprentissage et leur prototype devient inférieure à un seuil limite. Les derniers prototypes constituent les éléments du dictionnaire.

Cet algorithme itératif tend à optimiser un dictionnaire à partir d'un dictionnaire initial en convergeant vers un minimum local. Le choix du dictionnaire initial est donc important pour les performances de l'algorithme LBG. Linde, Buzo et Gray proposent de construire ce dictionnaire initial par une méthode de divisions successives (*splitting*) à partir de la base d'apprentissage :

1. On calcule le barycentre de la base d'apprentissage m_1 (molécule barycentre dans notre cas).
2. On fait varier très faiblement le barycentre initial par addition de termes aléatoires faibles de sorte à obtenir une deuxième molécule m_2 proche de m_1. On a alors un nouveau dictionnaire de taille 2.
3. On applique les étapes 2 et 3 de l'algorithme du LBG sur ce dictionnaire pour le stabiliser.
4. On fait varier faiblement chaque molécule du nouveau dictionnaire de sorte à doubler la taille de ce dernier.
5. On répète les étapes 3. et 4. jusqu'à obtenir N_c molécules (N_c est une puissance de 2).
6. On applique l'optimisation de l'algorithme F initial sur ces N_c molécules. Les N_c molécules finales constituent le dictionnaire.

L'algorithme LBG décrit ci-dessus permet de générer des dictionnaires de taille croissante et donc de plus en plus précis avec un rapport qualité/coût de codage optimal. Pour des molécules à 8 coefficients MDCT (dont on verra dans la Section 5.2 que cette taille de molécule offre le meilleur compromis entre capacité d'insertion du tatouage et la qualité de la séparation pour la résolution TF choisie), la taille maximale des dictionnaires est de 10 bits, soit 1024 formes, et ce même si la place disponible sur une molécule de mélange pour coder la forme est supérieure à 10 bits. Trois raisons expliquent la taille maximale de 10 bits : la limitation du temps de codage et décodage de l'information de forme, la nécessité d'avoir un ratio suffisant entre la taille des données d'apprentissage et la taille des dictionnaires (il est fixé supérieur à 100), et la limitation du temps de calcul de l'algorithme de génération LBG. De plus, il a été déterminé expérimentalement toujours pour des molécules de 8 coefficients que des dictionnaires de taille inférieure à 6 bits ne permettent pas de représenter de façon correcte les molécules sources. Les molécules sources sont alors mieux représentées par les molécules du mélange que par les molécules des dictionnaires. Pour cette raison, un ensemble de 5 dictionnaires $\mathcal{D}_i = \{D_i^r(p)\}, r \in [6, 10]$ de taille 2^r est calculé pour chaque canal fréquentiel moléculaire $p \in [1, W/2F]$, et pour chaque type de signal $i \in [1, I]$ (voix homme/femme, ou instrument selon le cas).

5.1.4.3 Estimation des signaux sources

L'estimation des signaux sources constitue le coeur du processus de séparation proprement dit. Elle est réalisée au bloc 13 de la Figure 5.1. Deux cas de figure sont à distinguer pour la reconstruction d'une molécule. Dans le cas où le seul descripteur de gain est choisi pour caractériser les signaux sources, chaque molécule de signal source est reconstruite grâce à la molécule de mélange de mêmes coordonnées temps-fréquence pondérée par la contribution énergétique relative de chaque source sur cette molécule. En reprenant les notations précédentes, une molécule $M_{pq}^{s_i}$ du signal source s_i est donc reconstruite à partir de la molécule du mélange M_{pq}^{x} par :

$$\hat{M}_{pq}^{s_i} = M_{pq}^{x} \times \sqrt{\check{E}_{s_i/x}(p,q)} \tag{5.19}$$

où $\check{E}_{s_i/x}(p,q)$ est la version quantifiée décodée du gain $E_{s_i/x}(p,q)$ de l'équation (5.15). Dans le cas de figure où l'on code les informations de forme, moyenne et écart-type de la molécule, une molécule $M_{pq}^{s_i}$ du signal source s_i est donc reconstruite par :

$$\hat{M}_{pq}^{s_i} = \check{\sigma}_{pq}^{s_i} \times \check{N}_{l_q} + \check{\mu}_{pq}^{s_i} \tag{5.20}$$

où $\check{\mu}_{pq}^{s_i}$ et $\check{\sigma}_{pq}^{s_i}$ sont les versions quantifiées des moyenne et écart-type, et $\check{N}_{l_q}$ est la molécule du dictionnaire $D_i^r(p)$ la plus proche de $N_{pq}^{s_i}$. Notons que dans ce dernier cas de figure, la molécule M_{pq}^{x} n'est plus utilisée comme base de reconstruction de la molécule $M_{pq}^{s_i}$, comme c'est le cas lorsque seul le gain est tatoué. La molécule M_{pq}^{x} n'est ici considérée que comme un support de l'information des descripteurs des signaux sources, et c'est la molécule prototype issue du dictionnaire qui est utilisée comme base de la reconstruction (on peut même dire que dans cette configuration le décodage et l'estimation de la molécule $\hat{M}_{pq}^{s_i}$ sont une seule et même étape). Ceci justifie l'appellation "séparation de sources informée par codage des signaux sources" utilisée pour ce premier système.

Finalement, la génération des signaux sources proprement dite est réalisée par transformée IMDCT du pavage TF estimé après ce processus de séparation. Cette opération (décrite à la Section 5.1.1) est réalisée au bloc 14 de la Figure 5.1.

5.1.4.4 Allocation de bits entre descripteurs

La répartition des ressources disponibles pour l'insertion du tatouage est un problème complexe d'optimisation. En effet, elle doit permettre une reconstruction des signaux sources la meilleure possible tout en respectant un grand nombre de contraintes. Cette répartition dépend en effet du nombre de sources, du type de descripteurs encodés, et de la précision avec laquelle chaque descripteur est encodé. Nous ne résolvons pas explicitement ce problème dans cette section, mais nous proposons des bornes au nombre de bits à utiliser pour chaque descripteur, ainsi qu'une solution idoine pour la séparation de deux à quatre sources. L'information de chacun des descripteurs des signaux sources (gain seul, ou triplet moyenne-gain-forme) ne nécessite pas le même nombre de bits pour être insérée sur le signal de mélange. Il faut parvenir, en fonction

du nombre global de bits disponibles sur chaque molécule, à un compromis sur la ressource binaire à allouer à chacune de ces informations. De plus cette allocation de bits varie en fonction du nombre de sources à séparer au sein du mélange. Plus le nombre de sources à séparer est élevé plus le nombre de bits alloués à chaque source décroît. Le nombre de bits alloués par source peut de plus être fonction de la nature même de cette source. Un instrument de musique et un signal de parole ne nécessitent pas forcément des descripteurs de même précision pour être convenablement séparés. Cependant de manière générale, des tests nous ont permis de déterminer quelques bornes en nombre de bits pour chacun des descripteurs. Ainsi la taille minimale d'un dictionnaire est de 64 formes, soit 6 bits, et énergie et moyenne ne sont codables avec une précision suffisante pour notre méthode que sur 4 bits minimum. Un exemple de table d'allocation de bits déterminée à la suite de tests d'écoute est donné à la table 5.1. Cette table présente l'allocation de bits établie pour la séparation de deux à quatre signaux sources de parole et d'instruments, dans le cas d'un pavage régulier du plan temps-fréquence avec des molécules de dimension 2×4. Les mélanges dont ils sont extraits comportent deux à quatre sources (deux à quatre locuteurs pour les mélanges de parole, et une voix chantée plus trois instruments pour la séparation de signaux de musique [16]). Pour des raisons de simplicité d'implémentation et de test, le nombre de bits alloués est ici identique pour chaque source.

Un codage du descripteur *gain* sur 7 à 8 bits, selon le nombre de sources à séparer, semble suffisant (l'amélioration observée par un codage plus précis n'est pas apparue significative). Or, il faut au moins 12 ou 14 bits (idem) pour pouvoir coder le triplet moyenne-gain-forme (rappelons que lorsque la forme est encodée, il faut impérativement tatouer l'information de moyenne et de gain de la molécule car les molécules du dictionnaires sont centrées réduites). Il existe donc un intervalle (entre 8 et 13, ou entre 7 et 13, ou entre 7 et 11 bits selon le nombre de sources à séparer) où il faut mettre en place une variante aux deux choix possibles jusqu'alors, "codage gain" et "codage moyenne-gain-forme". Nous avons choisi de raffiner le codage du gain des molécules quand de 7 à 13 bits sont disponibles pour le tatouage. Il s'agit en fait de diviser chaque molécule en 2 sous-molécules selon l'échelle des fréquences. Ainsi, une molécule de signal source 2×4 sera reconstruite grâce à 2 demi-molécules adjacentes 1×4 de mélange pondérées par les énergies des 2 demi-molécules de sources situées aux mêmes coordonnées du plan temps-fréquence. En d'autres termes, on applique exactement les principes du codage gain sur les deux demi-molécules. La colonne G de la table 5.1 qui désigne en général le nombre de bits alloués au codage du gain d'une molécule entière de signal source désigne, dans ce cas particulier où les demi-molécules sont utilisées, le nombre de bits alloués au codage de la première demi-molécule. G' désigne alors le nombre de bits alloués au codage de la seconde demi-molécule.

En résumé, trois cas de figure existent pour le codage des descripteurs d'une molécule source en fonction de la capacité d'insertion de la molécule du mélange :

16. ce corpus sera plus amplement détaillé dans la Section 5.2.1, et la table d'allocation sera utilisée dans les expérimentations correspondantes

TABLE 5.1 – Table d'allocation de bits (par source) pour la séparation de 2 à 4 sources avec des molécules de taille 2×4. M, G et S, signifient *Mean* (moyenne), *Gain* (écart-type), et *Shape* (forme). Quand les demi-molécules sont encodées G désigne le gain de la demi-molécule de plus haute fréquence, et G', celui de la demi-molécule de plus basse fréquence. *NS* est le nombre de signaux sources à séparer, C_M est la capacité d'une molécule.

	2 signaux				3 signaux				4 signaux			
C_M/NS	S	G	G'	M	S	G	G'	M	S	G	G'	M
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-
4	-	4	-	-	-	4	-	-	-	4	-	-
5	-	5	-	-	-	5	-	-	-	5	-	-
6	-	6	-	-	-	6	-	-	-	6	-	-
7	-	7	-	-	-	4	3	-	-	4	3	-
8	-	4	4	-	-	4	4	-	-	4	4	-
9	-	5	4	-	-	5	4	-	-	5	4	-
10	-	5	5	-	-	5	5	-	-	5	5	-
11	-	6	5	-	-	6	5	-	-	6	5	-
12	-	6	6	-	-	6	6	-	6	3	-	3
13	-	7	6	-	-	7	6	-	6	4	-	3
14	6	4	-	4	6	4	-	4	7	4	-	3
15	7	4	-	4	7	4	-	4	8	4	-	3
16	7	4	-	5	7	4	-	5	9	4	-	3
17	7	5	-	5	8	4	-	5	-	-	-	-
18	8	5	-	5	9	4	-	5	-	-	-	-
19	8	5	-	6	9	5	-	5	-	-	-	-
20	8	6	-	6	10	5	-	5	-	-	-	-
21	9	6	-	6	-	-	-	-	-	-	-	-
22	9	6	-	7	-	-	-	-	-	-	-	-
23	9	7	-	7	-	-	-	-	-	-	-	-
24	10	7	-	7	-	-	-	-	-	-	-	-
25	10	7	-	8	-	-	-	-	-	-	-	-
26	10	8	-	8	-	-	-	-	-	-	-	-

1. le gain (en tant que rapport énergétique source/mélange) de molécules entières
2. le gain (idem) de demi-molécules

3. le triplet moyenne-gain-forme des molécules entières (le gain est ici l'écart-type d'une molécule)

5.1.5 Le streaming du tatouage

Après allocation de bits et encodage des différents descripteurs des signaux sources d'une molécule donnée, la dernière étape de traitement consiste à fixer les codes résultant, qui constituent la watermark de cette molécule, sur la molécule de mélange correspondante dans le plan TF. Le message complet encodé sur la grille de quantification Q_2 consiste en la juxtaposition de chacun des sous-messages, à la fois de chaque descripteur, de chaque source, et éventuellement les mises à jours des quantificateurs des descripteurs (cf Sections 5.1.4.1 et 5.1.4.2). Le streaming du tatouage (*i.e.* le découpage du message global, juxtaposition de l'ensemble des messages élémentaires de chaque molécule) est en soi une étape purement technique, dont les caractéristiques sont fixées une fois pour toutes arbitrairement. La seule contrainte est que ce streaming soit identique au codeur et au décodeur pour que le décodage se fasse correctement. Lors du codage du message, la watermark complète, obtenue par la juxtaposition des sous-messages décrite ci-dessus, est redécoupée pour être insérée sur chacun des coefficients MDCT de la molécule. Chaque coefficient MDCT est tatoué au maximum de sa capacité d'insertion : il porte le message le plus volumineux, en nombre de bits, qui puisse lui être associé sans modification de son niveau de quantification sur la grille Q_1. On rappelle que les niveaux de quantification Q_1 et Q_2 ayant été déterminés de sorte à ne pas nuire à la qualité audio du signal porteur de la watermark, et à ne pas subir les influences de la fixation du signal sur support CD, aucun code correcteur d'erreurs n'est nécessaire. Au décodage, le message complet est lu sur la grille de quantification $Q_2(l, f)$. Une étape de découpage du message recouvré en sous-messages est alors nécessaire. Là encore, la table d'allocation de bits est utilisée pour savoir de combien de bits est constitué chaque sous-message. Une fois les sous-messages correspondant à chaque descripteur déterminés, il est possible de retrouver les valeurs numériques de chacun de ces descripteurs.

5.1.6 Un exemple global de traitement d'une molécule

Dans le but de synthétiser l'ensemble des traitements développés dans les précédents paragraphes, nous présentons dans cette section un exemple concret de la boucle complète de traitement. Tout au long de cette section, les différentes étapes du codeur et du décodeur présentés Figure 5.1 sont effectuées et illustrées. Au moment opportun on se focalisera sur l'étude d'une molécule spécifique d'un macro-bloc temporel. Nous nous plaçons dans cet exemple dans le cas d'un mélange linéaire instantané x de deux sources échantillonnées à 44.1 kHz jouant le même morceau de musique : un piano qui constitue la source s_1, et une voix chantée de femme qui représente la source s_2. Ces deux sources ont des supports temporels et fréquentiels partiellement superposés.

La première étape du codeur consiste comme nous l'avons vu dès la Section 5.1.1 en

une décomposition dans le plan temps-fréquence, par MDCT, des signaux sources et du signal de mélange, avec une fenêtre d'analyse de 512 échantillons. Après découpage des signaux en macro-blocs temporels de quelques secondes, l'ensemble des traitements de la séparation de sources informée est effectué sur chaque macro-bloc. Dans la suite nous nous plaçons sur un macro-bloc donné d'indice l. Les caractéristiques des signaux et de la fenêtre temporelle d'analyse utilisée sont celles décrites dans la Section 5.1.3.2. Le pavage temps-fréquence de ce macro-bloc est donc constitué de 256 canaux fréquentiels. En utilisant la grille de quantification $Q_A(f)$ introduite en détails dans la Section 5.1.3.5, les coefficients MDCT maximum de chaque canal fréquentiel $M_{max}(l, f)$ sont quantifiés sur 6 bits afin d'obtenir les facteurs d'échelle $A(l, f)$ des quantificateurs $Q_1(l, f)$ et $Q_2(l, f)$. L'ensemble des MDCT est ensuite quantifié sur la grille de référence $Q_1(l, f)$ et chacun prend une valeur, quantifiée sur 8 bits, comprise entre $-A(l, f)$ et $A(l, f)$. En ce qui concerne le troisième bloc du codeur de la Figure 5.1, nous nous plaçons dans cet exemple dans le cas d'un groupement moléculaire par pavage régulier du plan temps-fréquence, avec des molécules de taille 2×4. La Figure 5.8 représente une portion de la décomposition du signal de mélange dans le plan temps-fréquence et le rectangle rouge matérialise la localisation de la molécule $M^x_{3,4}$ de coordonnées (3,4) (ce qui correspond aux canaux fréquentiels 5 et 6 et aux canaux temporels 13 à 16) qui va maintenant nous servir d'exemple dans la suite de cette section.

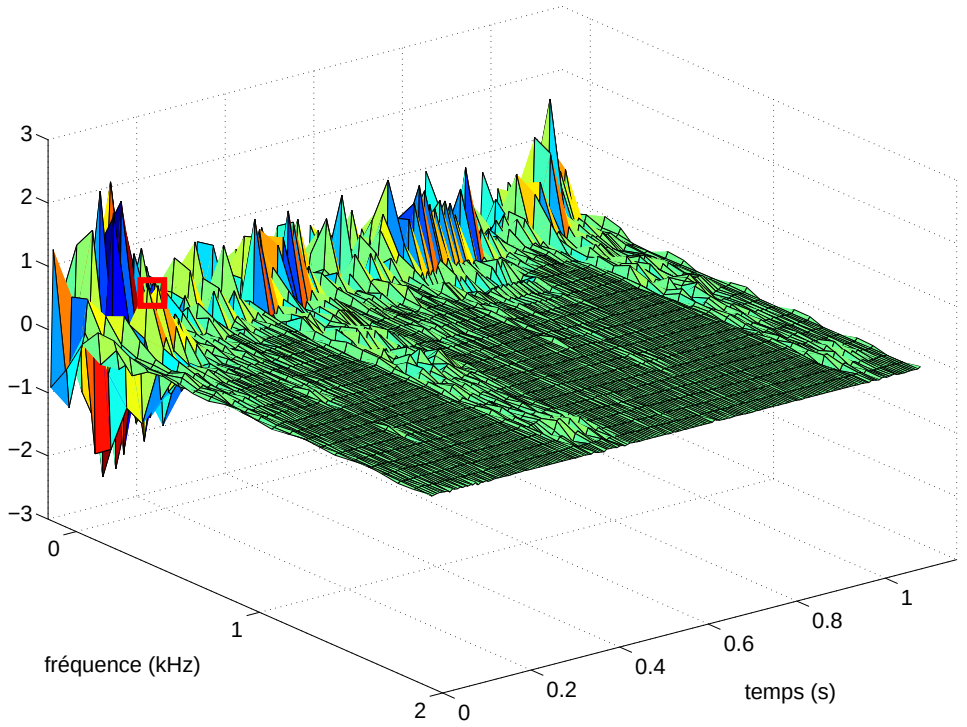


FIGURE 5.8 – Décomposition MDCT du mélange.

La résolution de la quantification de tatouage et donc la capacité d'insertion sur chaque canal fréquentiel sont données par les équations (5.8) et (5.13). En fonction

de la capacité d'insertion de la molécule, la lecture de la table d'allocation de bits 5.1 permet de savoir quels descripteurs peuvent être tatoués et de connaître le nombre de bits à allouer au codage de chacun. Dans cet exemple, 7 bits étant disponibles par coefficient MDCT sur le canal fréquentiel 5, et 6 bits sur le canal 6, la capacité d'insertion totale de la molécule est donc de 52 bits soit 26 bits par source à séparer. On déduit de la table d'allocation que les trois descripteurs de moyenne, forme et gain peuvent être utilisés, et que 8 bits sont alloués pour coder la moyenne, 10 pour la forme et 8 pour le gain. A noter que l'on peut vérifier l'importante concentration d'énergie sur les coefficients MDCT des signaux dans les basses fréquences, ce qui se traduit par de fortes capacités d'insertion de tatouage. Lors du codage de la forme, $N_{3,4}^{s_1}$ est comparée à chacune des 2^{10} molécules du dictionnaire $D_{10}^{s_1}(3)$ et la molécule du dictionnaire la plus "proche" de $N_{3,4}^{s_1}$ (forme normalisée de $M_{3,4}^{s_1}$) est retenue (notons-la $N_l^{D_{10}}(3)$). $M_{3,4}^{s_1}$ sera reconstruite à partir de $N_l^{D_{10}}(3)$, et non pas grâce à $M_{3,4}^{x}$ comme ce serait le cas avec le tatouage du gain seul. L'information de forme (pour s_1) à tatouer sur le signal de mélange correspond à l'indice l de $N_l^{D_{10}}(3)$ dans le dictionnaire $D_{10}^{s_1}(3)$. Dans notre exemple l'indice l est le *809*. Ce qui correspond au message binaire *1100101001* sur 10 bits. Les descripteurs de moyenne et de gain sont quant à eux codés par des QSU de résolution 8 bits, respectivement signé et non-signé. Une fois les informations sur les trois descripteurs codées, les sous-messages de code correspondants sont juxtaposés de la manière suivante : la forme, le gain et la moyenne, et ce pour chaque source. La figure 5.9 offre l'exemple du message tatoué sur la molécule $M_{3,4}^{x}$, le sous-message relatif à la source s_2 ayant été déterminé de la même façon que celui de la source s_1. Le message complet présenté Figure 5.9 est ensuite "découpé" pour être inséré sur chaque coefficient MDCT de la molécule $M_{3,4}^{x}$, en fonction de la place disponible sur chacun (ici 7 ou 6 bits par coefficient). Le découpage du message est illustré par les flèches situées au dessous du train binaire, qui indiquent le message porté par chacun des 8 coefficients de la molécule $M_{3,4}^{x}$ (les coefficients 1 à 4, situés sur le canal fréquentiel 5 ont une capacité d'insertion de 7 bits alors que les coefficients 5 à 8, situés sur le canal fréquentiel 6 ont une capacité d'insertion de 6 bits).

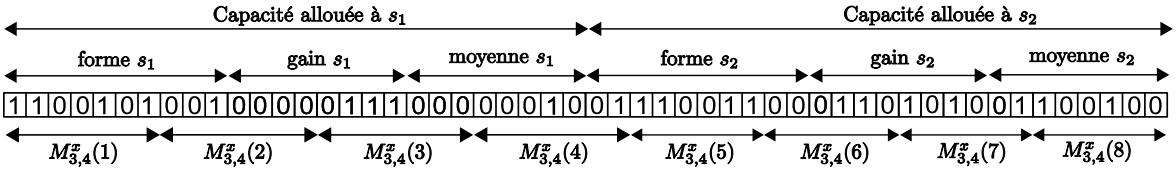
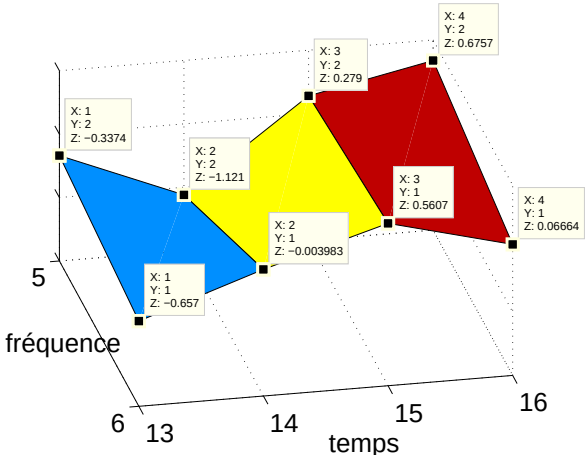
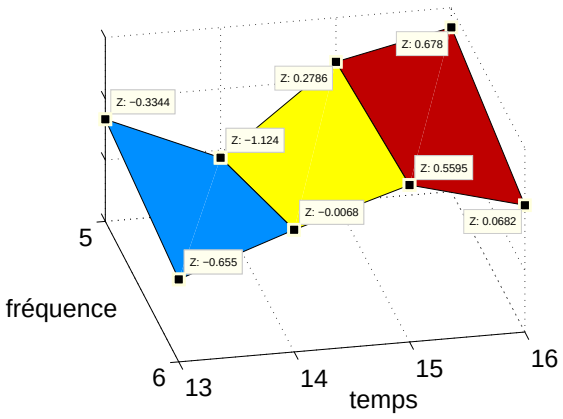


Figure 5.9 – Exemple de message tatoué sur une molécule de mélange. $M_{3,4}^{x}(i)$ représente les coefficients MDCT de la molécule $M_{3,4}^{x}$ renumérotés de façon arbitraire pour simplifier la présentation.

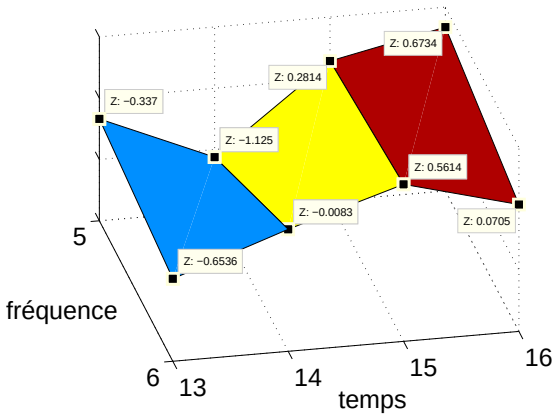
L'insertion de la watermark sur chaque coefficient MDCT est faite en modifiant la valeur de ce coefficient sur la grille $Q_2(l, f)$ (par rapport au niveau de quantification de référence sur $Q_1(l, f)$). La Figure 5.10 illustre l'ensemble des traitements subis par la molécule $M_{3,4}^{x}$ et montre les changements d'amplitude dont chaque coefficient MDCT de la molécule a été l'objet.



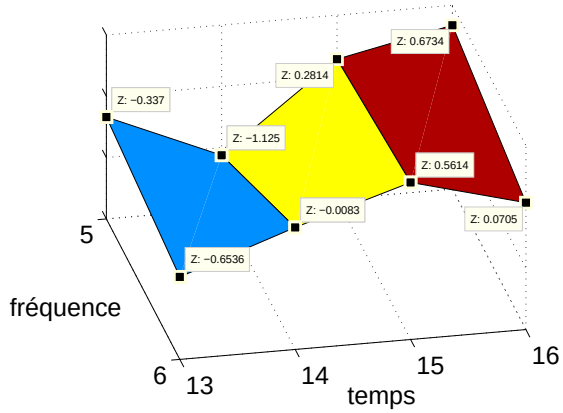
(a) Valeurs initiales des coefficients MDCT



(b) Quantification sur la grille $R1$



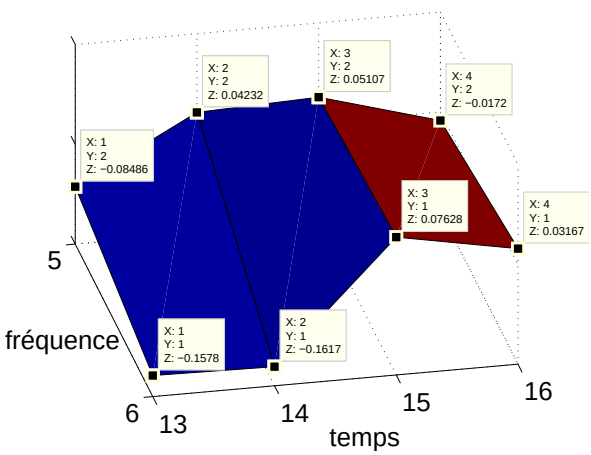
(c) Quantification sur la grille $R2$



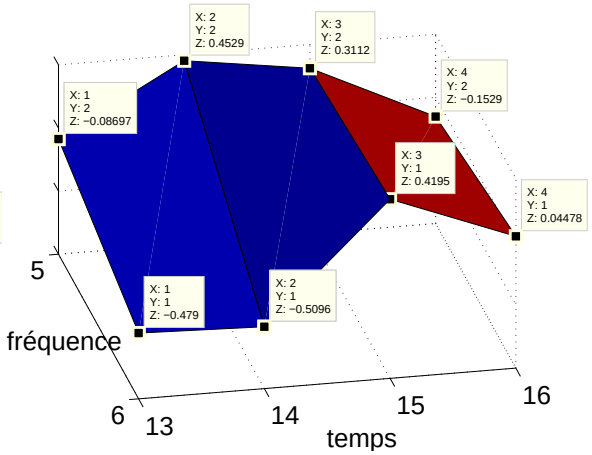
(d) Quantification 16-bits

FIGURE 5.10 – Les deux étapes de quantification subies par la molécule $M^x_{3,4}$ lors de l'insertion du tatouage, et influence de la conversion au format CD-audio.

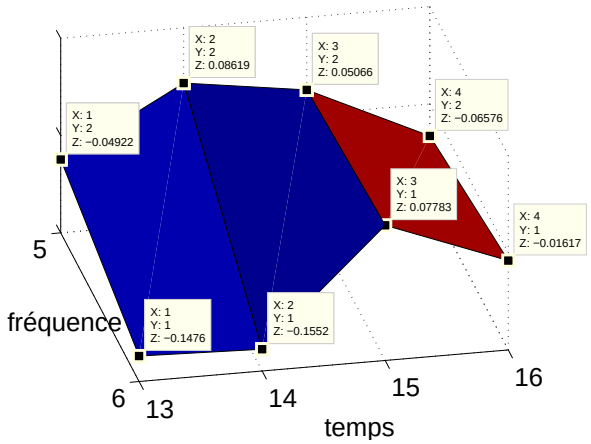
La Figure 5.11 représente la molécule $M^{s_1}_{3,4}$ initiale, la molécule $N^{D_{10}}_{809}(3)$ utilisée pour la reconstruction de la molécule $M^{s_1}_{3,4}$, et enfin la molécule reconstruite $\hat{M}^{s_1}_{3,4}$. La molécule $M^x_{3,4}$ située dans les basses fréquences est relativement énergétique, ce qui permet de coder l'information de forme à l'aide d'un dictionnaire de grande taille (ici 10 bits), d'où une forte similarité entre la molécule $N^{s_1}_{3,4}$ et $N^{D_{10}}_{809}(3)$, et un très bon codage de $M^{s_1}_{3,4}$. Dans cet exemple, il y a bien superposition significative des deux sources, et la molécule de mélange Figure 5.10a a une forme significativement différente de celle de la molécule de s_1 Figure 5.11a, justifiant l'utilisation d'un dictionnaire de formes.



(a) Valeurs initiales des coefficients MDCT



(b) Molécule normalisée du dictionnaire retenue



(c) Molécule reconstruite

Figure 5.11 – Estimation d'une molécule de signal à partir d'une molécule prototype d'un dictionnaire de forme.

5.2 Expérimentations

5.2.1 Données et plans expérimentaux

Les tests présentés dans cette section d'expérimentation sont effectués à la fois sur des signaux de parole échantillonnés à 16 kHz, et des signaux de musique échantillonnés à 44.1 kHz. Les mélanges utilisés sont de type voix+voix, et voix chantée+instruments, avec un nombre de sources variant de deux à quatre. Les signaux de parole sont issus de la base de données TIMIT. Les données de quatre locuteurs anglo-saxons, deux hommes et deux femmes, sont utilisés pour les tests. 279 locuteurs (122 locuteurs femme et 157 locuteurs hommes) prononçant 10 phrases différentes sont utilisées pour la construction des dictionnaires de forme. Pour ce qui est des instruments de musique, les instruments sont une guitare basse, une batterie et un piano, et la voix chantée est celle d'une chanteuse. Six morceaux de musique où les quatre sources musicales jouent en harmonie sont utilisées. Chacune des quatre sources a été enregistrée séparément dans des conditions de studio d'enregistrement. Quatre des six morceaux sont utilisés pour construire les dictionnaires de forme, et des extraits des deux autres morceaux sont utilisés pour tester les performances de séparation. Un morceau rapide et un autre plus lent ont été choisis pour les tests, de manière à augmenter la diversité des signaux tests. Les détails à propos des données d'apprentissage des dictionnaires sont données à la Table 5.2 [17]. Le ratio entre la taille des données d'apprentissage et la taille maximale des dictionnaires est toujours supérieur à 100 de manière à assurer la représentatitivé des dictionnaires générés. Dans la suite, les notations $kI + 1S$ font référence à un mélange musical de k instruments et une voie chantée, alors que nS fait référence à un mélange de parole constitué de n locuteurs.

TABLE 5.2 – Caractéristique du corpus d'apprentissage utilisé pour la génération des dictionnaires de forme.

signal source	durée (minutes)	taille (nombre de molécules)	taille du corpus / taille du dictionnaire
guitare basse	18.1	186732	182.3
voix chantée	10.8	109568	107.1
batterie	16.7	171080	161.1
piano	15.1	154836	151.2
locuteurs femme	64.1	239961	234.3
locuteurs homme	80.1	300580	293.5

Afin de tester notre méthode de séparation de sources, nos expérimentations portent sur trois grands axes : en premier lieu, la qualité de la chaîne de codage est évaluée,

17. Notons que la différence de durée entre les différentes sources musicales utilisées pour l'apprentissage résulte de la suppression des portions de silence.

pour vérifier que les résolutions de quantification R_1 et R_2 permettent d'assurer une inaudibilité de l'insertion d'une watermark test en même temps qu'une parfaite détection de celle-ci au décodeur. L'influence de la taille des molécules sur la qualité de reconstruction des signaux sources, indépendamment de toute problématique de tatouage est ensuite testée. Enfin, le schéma complet codeur/décodeur, combinaison de la phase de codage des descripteurs des signaux, de la phase de tatouage par quantification, et de la phase de séparation est vérifié. Les performances de séparations sont calculées avec les outils de la sous-section suivante.

5.2.2 Les mesures de performances

Les mesures de performances du système de séparation de sources informée ont été réalisées par deux grands type de mesures, les mesures informelles et subjectives d'écoute avec un casque audio de qualité, et des mesures mathématiques objectives.

Concernant les tests d'écoute, nous n'effectuons dans ce chapitre qu'une comparaison informelle de type (A,B), où A représente le signal original, et B le signal watermarqué. Des tests formels plus poussés sur un nombre significatif de sujets seront présentés dans le Chapitre 7 concernant les derniers résultats de séparation de sources obtenus avec une approche raffinée de SSI.

Un deuxième type de mesures, objectives, permet de mesurer la qualité des signaux aux différents stades du traitement en s'affranchissant de tests d'écoute. Ces mesures, plus faciles à mettre en œuvre que les tests d'écoute sont souvent utilisées en amont de tests subjectifs. Récemment, Vincent et al. ont proposé un ensemble de critères de mesure de performances de séparation de sources [Vincent et al., 2005]. Dans le cas de figure où les signaux sources originaux sont disponibles, ces mesures permettent de quantifier la qualité d'un processus de séparation de sources. Dans le cas d'un mélange non-bruité, le principe de ces mesures de performances est le suivant : on considère que l'estimée $\hat{s}_i(t)$ d'un signal $s_i(t)$ peut être décomposée selon la somme

$$\hat{s}_i(t) = s_i^{cible}(t) + e_i^{interf}(t) + e_i^{art}(t) \tag{5.21}$$

où $s^{cible}(t)$ représente une dégradation acceptable du signal source $s_i(t)$, $e_i^{interf}(t)$ représente une dégradation due aux autres signaux sources considérés alors comme des interférences, et $e_i^{art}(t)$ représente les artefacts dûs au processus de séparation tels que le bruit musical. Les mesures de performances introduites sont alors les Signal-to-Distorsion Ratio (SDR), Signal-to-Interference Ratio (SIR), et Signal-to-Artefacts Ratio (SAR) définis respectivement par

$$\mathrm{SDR} = 10\log_{10}\frac{\|s_i^{cible}\|^2}{\|e_i^{interf} + e_i^{art}\|^2} \tag{5.22}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\|s_i^{cible}\|^2}{\|e_i^{interf}\|^2} \tag{5.23}$$

$$\mathrm{SAR} = 10 \log_{10} \frac{\|s_i^{cible} + e_i^{interf}\|^2}{\|e_i^{art}\|^2} \tag{5.24}$$

Nous utilisons également deux autres mesures objectives de la qualité de la séparation, notées respectivement SNR_{out} (rapport signal-sur-bruit en sortie) et ISNR (amélioration du rapport signal-sur-bruit entre la sortie et l'entrée). Le SNR_{out} est le rapport Signal-to-Noise où le Signal est le signal source original s_i et Noise consiste en la différence entre le signal source cible original et son estimé par le processus de séparation. Considérons un mélange $\sum_{k=1}^{I} s_k$ de I signaux sources. Le SNR_{out} peut alors être défini par

$$\mathrm{SNR}_{out}^{i} = 10 \log_{10} \left(\frac{\sum_n (s_i\,[n])^2}{\sum_n (s_i\,[n] - \hat{s}_i\,[n])^2} \right) \tag{5.25}$$

Notons que le SNR_{out} ainsi défini est voisin du SDR de Vincent et al (si $s_i(t) = s^{cible}(t)$, SNR_{out}=SDR). Pour tenir compte de la difficulté à séparer une source du mélange, nous introduisons le rapport Signal-to-Noise d'entrée du mélange, noté SNR_{in}, dans lequel Signal est le signal source cible s_i et Noise le bruit consistant en la somme des autres signaux sources originaux dans le mélange $\sum_{k=1,k\neq i}^{I} s_k$. Le SNR_{in} est ainsi défini par

$$\mathrm{SNR}_{in}^{i} = 10 \log_{10} \left(\frac{\sum_n (s_i\,[n])^2}{\sum_{k\neq i} \sum_n (s_k\,[n])^2} \right) \tag{5.26}$$

L'amélioration du rapport signal-sur-bruit, ISNR, pour un signal cible s_i entre l'entrée et la sortie est alors donnée par

$$\mathrm{ISNR} = \mathrm{SNR}_{out} - \mathrm{SNR}_{in}. \tag{5.27}$$

L'ISNR permet de mesurer effectivement l'efficacité du processus de séparation, ce que ne permet pas le SNR_{out} seul. En effet, si la mesure SNR_{in}^{i} pour une source donnée s_i est relativement élevée (signifiant que cette source n'est pas beaucoup dégradée dans le mélange), alors la valeur de SNR_{out}^{i} correspondant a de fortes chances d'être aussi élevée sans que la qualité audio de la source estimée $\hat{s}_i$ ne soit garantie. Dans ce même cas de figure, la valeur de ISNR^{i} sera, elle, faible, indiquant ainsi que s_i n'est pas convenablement rehaussée. L'amélioration du rapport signal sur bruit est d'autant plus élevée que la reconstruction est fidèle à l'original mais aussi que l'original est fortement "dilué" parmi les autres sources dans le mélange.

5.2.3 Résultats

5.2.3.1 Le processus de tatouage : inaudibilité et fiabilité

L'objectif de cette section est tout d'abord de vérifier l'inaudibilité de l'insertion du tatouage sur le signal de mélange, puis, de mesurer la fiabilité du système de transmission de ce tatouage.

Des tests objectifs et subjectifs ont été réalisés sur les signaux tests de parole et de musique introduits à la Section 5.2.1. Après décomposition dans le plan temps-fréquence de ces signaux, les coefficients MDCT de leur pavage temps-fréquence ont été quantifiés à une résolution $R_1 = 8$ bits avec une mise à jour des facteurs d'échelle toutes les 1.5 secondes, avant de reconstruire un signal temporel par transformée MDCT inverse. On se place ainsi dans des conditions très similaires à celles du tatouage. On rappelle que la grille Q_2, support du tatouage, étant une sous-grille de la grille Q_1, si l'effet d'une quantification à la résolution R_1 du signal de mélange est inaudible, alors l'inaudibilité du tatouage est assurée, même si Q_2 n'est pas une quantification à proprement parlé. C'est entre ces signaux reconstruits et leur version temporelle originale que sont effectués les tests objectifs sur les effets de la quantification fréquentielle à la résolution 8 bits. La Table 5.3 présente un échantillon des valeurs de rapport signal sur bruit obtenues. L'ensemble des mesures fournissant des valeurs de rapport signal sur bruit supérieures à 35dB, nous pouvons considérer que la quantification à la résolution R_1 fixe sur l'ensemble du signal n'affecte pas la qualité audio des signaux audio en sortie de codeur. Aux vues des résultats obtenus Table 5.3, il est possible de conclure que le tatouage n'a pas d'effet perceptible sur la qualité d'écoute des signaux testés.

Signaux	Musique	Parole	Musique + Parole
1	35,4	38,8	42,1
2	36,4	37,8	39,9
3	35,2	37,6	40,1
4	35,6	39,3	42,2
5	37,3	39,5	42,1
6	36,0	38,1	40,0
7	37,9	39,0	39,9
8	36,1	38,6	40,9
9	36,4	38,0	40,2
10	35,0	38,7	40,3
moyenne	**36,1**	**38,5**	**40,8**

Table 5.3 – Rapport signal sur bruit entre un signal et sa version quantifiée sur 8 bits pour 3 types de signaux

Une fois l'inaudibilité du tatouage assurée, la fiabilité du système de transmission associée est ensuite testée, pour vérifier si le processus de décodage permet de retrouver la watermark au décodeur. Pour cela un message aléatoire de la taille maximum en

regard de la capacité d'insertion est tatoué sur le signal de mélange puis décodé. Le taux d'erreurs binaires est alors calculé sur chaque canal fréquentiel. Le taux d'erreur binaire est le ratio entre le nombre de bits erronés et le nombre de bits total émis. Pour l'ensemble de signaux testés (tous genres confondus), ce taux d'erreur est de zéro : le message inséré au codeur est parfaitement retrouvé. Le système de tatouage en tant que chaîne de transmission d'information utile à la séparation fonctionne donc parfaitement dans les conditions prévues (pas d'attaques intentionnelles autres que la conversion PCM 16 bits en sortie de codeur).

5.2.3.2 La taille des molécules

Nous avons vu aux sections 4.3.3 et 5.1.2 que les molécules de mélange constituent le support élémentaire de l'information issue des descripteurs des signaux sources à séparer. Pour cette raison, la taille des molécules est un paramètre qui occupe une place particulièrement importante au sein de notre méthode de SSI par codage des signaux sources. Elle influe sur la place disponible pour insérer la watermark et donc le nombre de sources séparables et conditionne la qualité de la reconstruction. L'objectif est ici d'établir une taille de molécule satisfaisante pour la séparation d'un nombre fixé de signaux sources dans le cas de mélanges de parole et de musique. Pour se faire nous testons l'influence de la taille d'une molécule sur la qualité de la séparation.

Les tests sur des signaux de parole sont effectués sur deux mélanges, de deux et quatre locuteurs, d'une durée de 30 secondes. Concernant les signaux de musique, des tests sont effectués sur 60 secondes d'un mélange 3I+1S. Les conditions d'étude de la taille des molécules sont ici restreintes aux mesures de performances obtenues lors de l'utilisation du seul gain. Même si un test complet dans la configuration moyenne-gain-forme serait en toute rigueur préférable à la détermination d'une taille optimale de fenêtre, nous choisissons ici de limiter notre étude à une validation d'une taille de molécule satisfaisante. En effet, déterminer une taille de molécule optimale implique de générer, pour chacune des tailles testées, des dictionnaires de forme et des tables d'allocations de bits correspondantes, ce qui alourdit considérablement les tests, en comparaison de ceux à effectuer dans le cas de l'utilisation du gain seul. Une deuxième caractéristique de ce test est qu'il ne vise pas à tester la qualité du processus de watermarking, c'est pourquoi la reconstruction des signaux sources est ici testée du simple point de vue de la séparation sans prendre en compte le processus de tatouage et de quantification des descripteurs. Les signaux sources sont donc ici reconstruits par les molécules du signal de mélange pondérées de la vraie valeur du gain de chaque source, et non pas à partir de son approximation par quantification comme c'est le cas à l'équation (5.19).

Une étape préalable aux tests sur la taille des molécules est la détermination de bornes d'étude. Pour cela, nous effectuons le raisonnement suivant : si une molécule est trop petite, la capacité de cette molécule est insuffisante pour permettre de coder l'information de forme de la molécule (seul le gain de la molécule peut être codé). Or l'utilisation des dictionnaires de forme a pour but d'augmenter la qualité de codage.

Ainsi, sous la double contrainte du choix de la configuration de codage (gain seul ou moyenne-gain-forme) et de la résolution avec laquelle sont encodés les descripteurs, une limite inférieure à la taille d'une molécule a par conséquent été fixée. Dans la configuration mono-canal décrite dans cette section, la taille minimale d'une molécule est fixée à 8 coefficients MDCT (le cas de molécules 1×4 est introduit en prévision de la Section 5.3 traitant de mélange stéréophoniques). Une deuxième contrainte, relative à la structure même des signaux à séparer, fixe une limite supérieure à la taille des molécules. Pour limiter le risque de superposition entre les différents signaux sources, il est logique de penser que la taille des molécules ne doit pas être trop importante. C'est pourquoi les tests effectués se limitent à une dimension maximale de molécule de 16 coefficients. Des tests sont donc effectués pour différentes tailles de molécules, de 4 à 16 coefficients et pour divers types de mélanges. Les résultats présentés pour chaque taille de molécules sont les résultats moyennés sur l'ensemble des I sources composant le mélange. Les Tables 5.4, 5.5 et 5.6 présentent les performances moyennes de reconstruction pour différents types de mélanges.

Table 5.4 – Influence de la taille des molécules sur la qualité de reconstruction des signaux pour des mélanges de 2 sources de parole.

F×T	SNR	ISNR	SDR	SIR	SAR
1×4	12,9	12,9	12,8	21,2	13,5
2×4	12,2	12,2	12,0	19,7	12,8
4×2	11,7	11,7	11,5	19,2	12,3
3×4	11,4	11,4	11,2	18,3	12,2
4×3	11,4	11,4	11,2	18,4	12,1
4×4	11,1	11,1	10,8	17,7	11,9

Table 5.5 – Influence de la taille des molécules sur la qualité de reconstruction des signaux pour des mélanges de 4 sources de parole.

F×T	SNR	ISNR	SDR	SIR	SAR
1×4	7,5	12,4	6,9	14,2	8,0
2×4	6,8	11,6	6,0	12,6	7,4
4×2	6,7	11,6	6,0	12,5	7,3
3×4	6,2	11,1	5,4	11,4	6,9
4×3	6,4	11,3	5,6	11,8	7,1
4×4	6,0	10,9	5,1	10,9	6,8

Il apparaît clairement pour chacun des trois types de mélange testés que plus la taille d'une molécule croît, moins la séparation est bonne. Les meilleurs résultats de séparation sont obtenus pour une molécule 1×4, mais cette taille de molécule n'est pas suffisante pour garantir un codage performant des signaux sources, c'est pourquoi, dans cette première approche de SSI par codage des signaux sources à partir d'un mélange

TABLE 5.6 – Influence de la taille des molécules sur la qualité de reconstruction des signaux pour des mélanges de 3 instruments et une voix chantée.

F×T	SNR	ISNR	SDR	SIR	SAR
1×4	5,9	11,8	4,7	9,5	7,1
2×4	5,5	11,3	4,1	8,3	7,0
4×2	5,1	11,0	3,6	7,6	6,9
3×4	5,1	11,0	3,6	7,5	7,0
4×3	5,0	10,8	3,4	7,2	6,9
4×4	4,9	10,7	3,3	7,0	6,9

monophonique, une taille de molécule de 2 canaux fréquentiels (*i.e.* 62.5Hz pour des signaux échantillonnés à $F_e = 16kHz$ et 172.5Hz si $F_e = 44.1kHz$) en ordonnées et 4 canaux temporels (*i.e.* 46ms si $F_e = 44.1kHz$, pour une fenêtre temporelle de décomposition MDCT de 512 échantillons) en abscisses est celle choisie. C'est cette dimension de molécule qui permet, pour chacun des trois mélanges testés, d'obtenir la meilleure reconstruction des signaux sources dans le cas où seul le gain des molécules de ces signaux est tatoué sur le signal de mélange. À nombre de coefficients égal, les performances de séparation d'une molécule privilégiant la résolution fréquentielle à la résolution temporelle sont légèrement plus élevées, ce qui confirme l'importance de la dynamique spectrale pour les signaux audio.

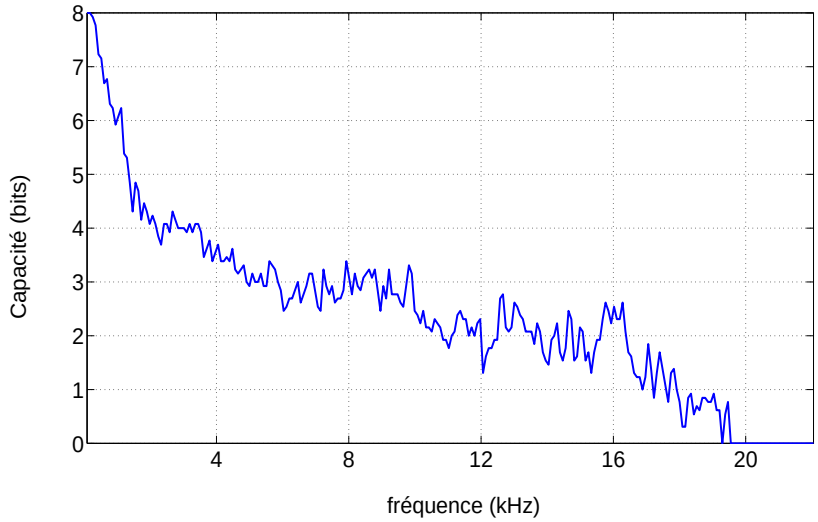
5.2.3.3 Capacité d'insertion de l'information, et allocation de bits

L'objectif de cette section est de vérifier la capacité d'insertion d'information sur des molécules de taille fixée. Nous souhaitons également établir la corrélation entre l'importance énergétique et donc perceptive des coefficients MDCT et la place disponible pour insérer le tatouage.

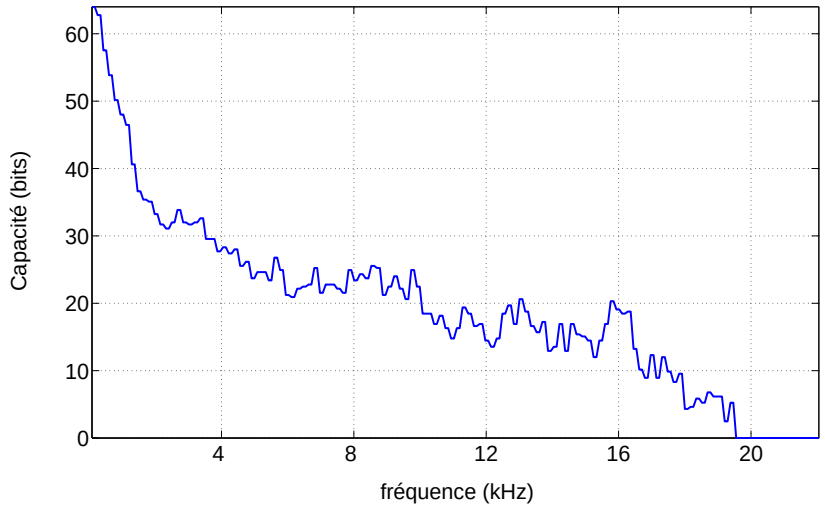
Pour cela nous nous plaçons pour ces tests dans le cas où les molécules sont de dimension 2×4. Les signaux tests sont toujours des mélanges de deux à quatre locuteurs pour les mélanges de parole, et des mélanges 3I+1S pour les signaux de musique. Les performances sont mesurées sur des portions de 30 secondes de signaux, avec des capacités mises à jour tous les macro-blocs de 1.5 secondes, puis moyennées sur chaque canal fréquentiel.

La figure 5.12a indique les valeurs moyennes (calculées sur les 30 secondes de signal) prises par la capacité $C(f)$ sur chaque canal fréquentiel f (cf équation (5.7)). La Figure 5.12b obtenue à partir de $C(f)$ (cf équation (5.8)) indique le nombre de bits disponibles par molécule en fonction de son canal fréquentiel moléculaire (une molécule étant ici de taille 2×4, un canal fréquentiel de molécule équivaut à deux canaux fréquentiels de coefficient MDCT).

Avec cette première configuration de calcul de la capacité d'insertion, il apparaît qu'une relativement grande quantité d'information peut être tatouée sur le signal tout en respectant les deux contraintes indispensables : inaudibilité et parfait décodage de



(a) Nombre de bits disponibles par coefficient MDCT



(b) Nombre de bits disponibles par molécule

Figure 5.12 – Exemple du nombre de bits disponibles pour le tatouage en fonction du canal fréquentiel. Valeurs moyennées sur une durée de 30 secondes d'un signal de musique *jazz* composé de 3 instruments et une voix chantée.

la watermark. Comme pour de nombreux signaux de musique, la plupart de l'énergie du signal est localisée dans une portion spécifique du spectre de fréquence, principalement en basses fréquences. Il en résulte une capacité d'insertion allant jusqu'à plus de 60 bits par molécule pour les zones de plus forte énergie des signaux, puis entre 2 et 9kHz, de 20 à 30 bits sont disponibles, entre 12 et 16kHz, environ 15 bits, et à partir de 16kHz, la faible énergie des coefficients MDCT se traduit par une chute de la capacité disponible. La capacité d'insertion est donc confortable pour coder les descripteurs des signaux sources, particulièrement en basses fréquences. En hautes fréquences, même si la capacité disponible est significativement inférieure à celle disponible en plus basse fréquence, ceci n'est pas forcément très pénalisant car, l'oreille humaine étant moins sensible en hautes fréquences, il y a moins d'information à encoder sur ces molécules

(la contribution énergétique des sources, par exemple, n'a pas besoin d'être quantifiée aussi précisément pour ces molécules car leur participation énergétique est plus faible dans le signal de mélange) d'où la possibilité de conserver un ratio "information à insérer/place disponible" satisfaisant. De manière générale, l'information la plus utile à la qualité d'un signal audio est localisée en basse fréquence. Les coefficients MDCT basses fréquences sont de plus fortes amplitudes que les coefficients hautes fréquences, et selon l'équation (5.13), le nombre de bits disponibles sur chaque canal fréquentiel moléculaire pour insérer la watermark est proportionnel à l'amplitude $A(l, f)$ des coefficients MDCT sur ce canal (par l'intermédiaire de la résolution R_2). Plus l'information d'un signal est importante pour sa qualité audio, plus la place disponible pour tatouer l'information relative aux sources est grande, et de fait plus les sources peuvent être décrites avec précision. Or le nombre de bits disponibles est particulièrement élevé en basses fréquences, d'où la possibilité de coder l'information de forme précisément grâce à des dictionnaires de grande taille, et de fait une relativement bonne reconstruction des signaux sources dans cette zone du plan temps-fréquence [18]. Ceci vérifie ce que nous annoncions en 4.4 et le parallèle fait avec les modèles psychoacoustiques utilisés en compression.

La Figure 5.13 fournit le détail de l'allocation de bits, par source, entre les différents descripteurs obtenue à partir de la capacité présentée Figure 5.12 et de la table d'allocation 5.1, dans le cas de deux sources à séparer. Les dictionnaires de forme sont utilisés dans les plus basses fréquences, jusqu'à environ 4.5kHz. De 4.5 à 16.5kHz, le gain de demi-molécules 1×4 peut être encodé, alors que de 16.5 à 19.5kHz, seul le gain d'une molécule est utilisé pour aider à la séparation. Dans les plus hautes fréquences, il arrive que moins de 4 bits soient disponibles par source, auquel cas aucune information ne peut être encodée. Notons qu'au delà de 16.5kHz, les coefficients MDCT du mélange sont vraisemblablement de très faible amplitude, et de fait quasiment inaudible. La séparation des diverses sources ne semble pas indispensable dans cette portion du spectre.

Les débits [19] d'insertion moyens (sommés sur l'ensemble des canaux fréquentiels et moyennés sur une grande durée de signal) de plusieurs mélanges de parole et de musique, de deux à quatre sources, sont présentés Table 5.7. Des débits d'insertion de 115.3 à 134 kbits/s sont obtenus pour des signaux de musique, et de 46.6 à 52.3 kbits/s pour des signaux de parole. Ces taux élevés confirment la possibilité d'insérer une quantité importante d'information sur des mélanges de signaux audio, particulièrement de musique. Nous travaillons avec des signaux de mélange normalisés en amplitude. Lorsque les signaux sources composant le mélange changent, la valeur normalisée du mélange est également modifiée, et donc celle de $A(l, f)$, ce qui peut expliquer que la

18. Toutefois, nous verrons dans la Section 5.3 qu'une méthode plus raffinée de calcul de la capacité tenant compte du contenu spectro-temporel du signal hôte permet d'augmenter significativement la capacité d'insertion en moyennes et hautes fréquences, ce qui résulte en une amélioration des performances de séparation dans ces portions fréquentielles, mais aussi en une amélioration globale de la qualité du signal.

19. Le débit d'insertion correspond à la capacité d'insertion par seconde.

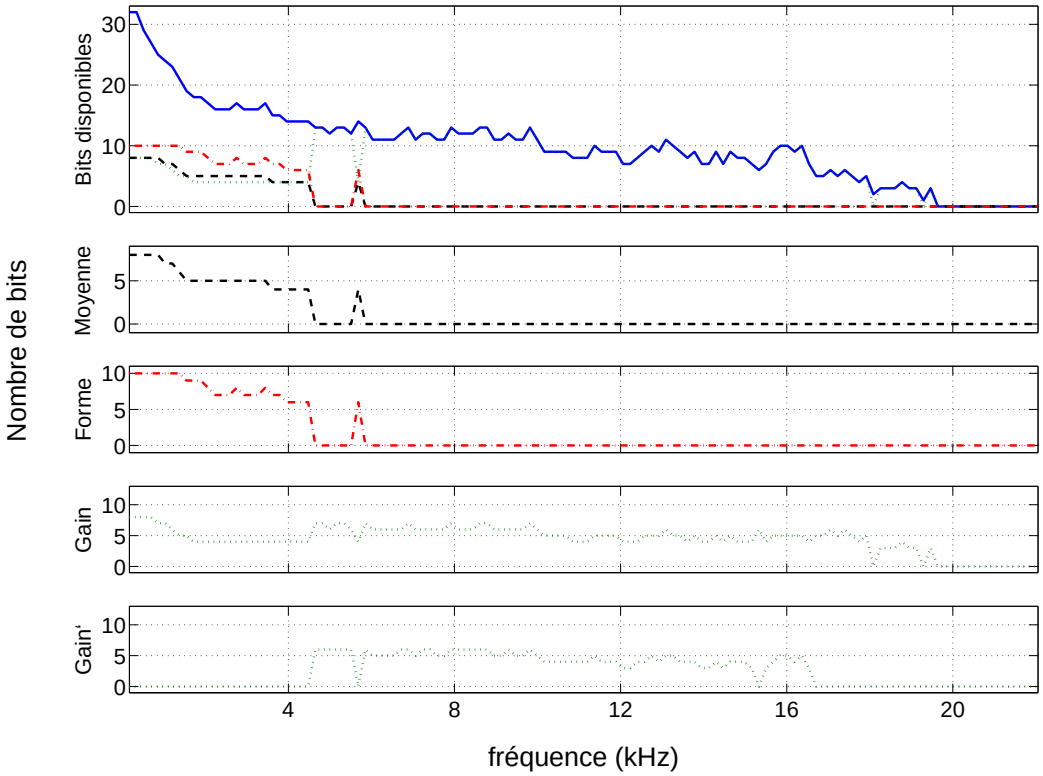


Figure 5.13 – Bits disponibles par source (courbe bleue) pour la séparation de deux signaux de musique, et allocation des descripteurs moyenne (courbe noire), forme (courbe rouge) et gain (molécule entière ou demi-molécule, courbe verte).

capacité d'insertion pour un mélange de deux instruments puisse être supérieure à celle d'un mélange de trois instruments.

Mélange	3I+1S	3I	2I+1S	2I	1I+1S	4S	3S	2S
Capacité (kbits/s)	122.0	115.4	121.8	134.0	115.3	52.3	51.1	46.6

Table 5.7 – Débit d'insertion pour divers mélanges de signaux de parole et de musique, calculée sur approximativement 60 secondes de signal. Les mélanges sont composés de k instruments et une voie chantée (kI+1S) avec k=1,2 ou 3, m instruments (mI) avec m=2 ou 3, et n locuteurs (nS), avec n=2, 3 ou 4.

5.2.3.4 Boucle complète codeur/décodeur et résultats de séparation

Dans cette section sont présentés les résultats obtenus pour la boucle complète du système codeur/décodeur introduit à la Figure 5.1, dans cette première configuration de SSI par codage des signaux sources.

Une première série de tests est effectuée pour mesurer les effets de l'utilisation des dictionnaires de formes sur les performances de séparation. La comparaison est faite

entre deux configurations : dans la première seul le descripteur de gain est utilisé pour l'ensemble des molécules, même lorsque la capacité disponible est élevée (le descripteur de gain défini formule (5.15) est alors quantifié avec une forte précision). Dans une seconde configuration, les trois descripteurs de moyenne, gain et forme sont utilisés, et la ressource binaire disponible est allouée selon les règles définies à la Table 5.1. Nous nous plaçons pour ces tests dans le cas de molécules de taille fixe 2×4. La Table 5.8 présente les résultats obtenus pour les deux configurations (gain seul (*Gain*) et moyenne-gain-forme (*MGF*)) pour la séparation de 4 signaux musicaux de leur mélange de type 3I+1S, et la séparation de deux signaux de parole à partir d'un mélange 4S de deux locuteurs femme et deux locuteurs homme. La Table 5.8 montre des scores d'ISNR variant de 7.3 à 12.7 dB dans la configuration *Gain* et de 11.7 à 19.7dB dans la configuration *MGF*, ce qui traduit une bonne séparation aussi bien pour des signaux de parole que pour des signaux de musique. Des tests d'écoute réalisés sur les signaux séparés en configuration *Gain* révèlent la présence d'interférences résiduelles, ainsi que d'un bruit musical significatif. L'amélioration de performances de séparation obtenue par l'utilisation de dictionnaires de forme est évidente, avec une augmentation du score d'ISNR de 4.5dB en moyenne (de 2.2dB pour le piano qui possède la distribution temporelle la plus parcimonieuse, à 7 et 7.2dB pour la batterie et la guitare basse respectivement). Les défauts rencontrés dans la configuration *Gain* sont très largement atténués par l'utilisation de dictionnaires dans la configuration *MGF*.

TABLE 5.8 – Performances pour la séparation de 4 signaux d'un mélange de 4 signaux de musique, et la séparation de 2 signaux de parole à partir d'un mélange de 4 signaux de parole avec et sans dictionnaires de forme.

Mix	Signaux	ISNR (dB)	
		Gain	MGF
Musique	guitare basse	9.6	16.8
	voix chantée	7.3	11.7
	batterie	12.7	19.7
	piano	12.1	14.3
Parole	voix de femme	12.0	15.0
	voix d'homme	9.3	12.1

Maintenant que l'intérêt des dictionnaires de forme a été mis en évidence, nous présentons de plus amples résultats obtenus avec la configuration *MGF* et avec la table d'allocation 5.1. La Table 5.9 donne les performances de séparation pour le codage des signaux sources sur un signal de mélange monophonique pour différents types de mélanges de signaux de musique, et un nombre variable de sources extraites. De deux à quatre signaux sont séparés à partir de mélanges de trois ou quatre sources. La Table 5.9 illustre l'influence sur les performances de séparation du nombre de sources dans le mélange, et du nombre de sources à séparer. A noter que plus le nombre de sources

Table 5.9 – Performances de séparations pour des mélanges de signaux de musique en terme d'ISNR (dB).

sources à séparer	mélange de 4 sources (g,c,b,p)			mélange de 3 sources (g,b,p)		
	SNR_{in}	SNR_{out}	ISNR	SNR_{in}	SNR_{out}	ISNR
guitare basse (g)	-5.8	11.1	16.8	-	-	-
voix chantée (c)	-1.1	10.6	11.7	-	-	-
batterie (b)	-9.0	10.7	19.7	-	-	-
piano (p)	-5.7	8.6	14.3	-	-	-
guitare basse (g)	-5.8	13.1	18.9	-1.9	13.1	15.0
voix chantée (c)	-1.1	14.4	15.5	-	-	-
batterie (b)	-9.0	13.3	22,3	-5.7	13.1	18.9
piano (p)	-	-	-	-1.8	10.3	12.1
guitare basse (g)	-5.8	13.4	19.2	-1.9	13.4	15.3
voix chantée (c)	-	-	-	-	-	-
batterie (b)	-	-	-	-	-	-
piano (p)	-5.7	11.3	17.0	-2.7	11.3	14.0
guitare basse (g)	-	-	-	-	-	-
voix chantée (c)	-1.1	15.1	16.2	-	-	-
batterie (b)	-9.0	14.7	23.7	-	-	-
piano (p)	-	-	-	-	-	-

composant un mélange est important, plus le SNR d'entrée (SNR_{in}) est faible, et plus grandes sont les chances de recouvrements entre les signaux sources. Pour le mélange de trois sources, le SNR d'entrée est de -1.9dB, -5.7dB et -1.8dB respectivement pour la *guitare basse*, la *batterie*, et le *piano*, mais quand on ajoute au mélange un signal de voix chantée avec un SNR d'entrée de -1.1dB, les SNR d'entrée des autres sources décroissent jusqu'à -5.8, -9.0 et -5.7dB respectivement. Le principal résultat ressortant de la Table 5.9, est une bonne séparation, indépendamment du type de source ou du mélange, y compris dans le cas le plus ardu de la séparation de quatre sources de musique à partir d'un mélange de quatre sources. Dans ce dernier cas, les valeurs de ISNR obtenues s'échelonnent de 11.7dB pour le signal de Chant (cette valeur la plus faible de ISNR s'explique par le fait que le Chant soit la source ayant le SNR d'entrée le plus élevé des quatre sources), à 19.7dB pour la batterie (qui possède le SNR d'entrée le plus faible). Le score d'ISNR des autres instruments est de 14.3 pour le piano et 16.8dB pour la guitare basse. Dans le cas de la séparation de trois sources à partir d'un mélange de quatre sources musicales, les ISNR s'échelonnent de 15 à 22.3dB. Cet

accroissement significatif des deux mesures de SNR de sortie et d'ISNR entre la séparation de quatre sources et celles de trois sources (à partir du même mélange) révèle l'influence du partage de la ressource de capacité d'insertion d'information entre les sources. En effet, la ressource allouée au piano dans le cas d'un mélange à 4 sources est distribuée entre basse, voix et batterie dans le cas d'un mélange à 3 sources, d'où une augmentation des performances de séparation lorsque le mélange est composé de seulement 3 sources.

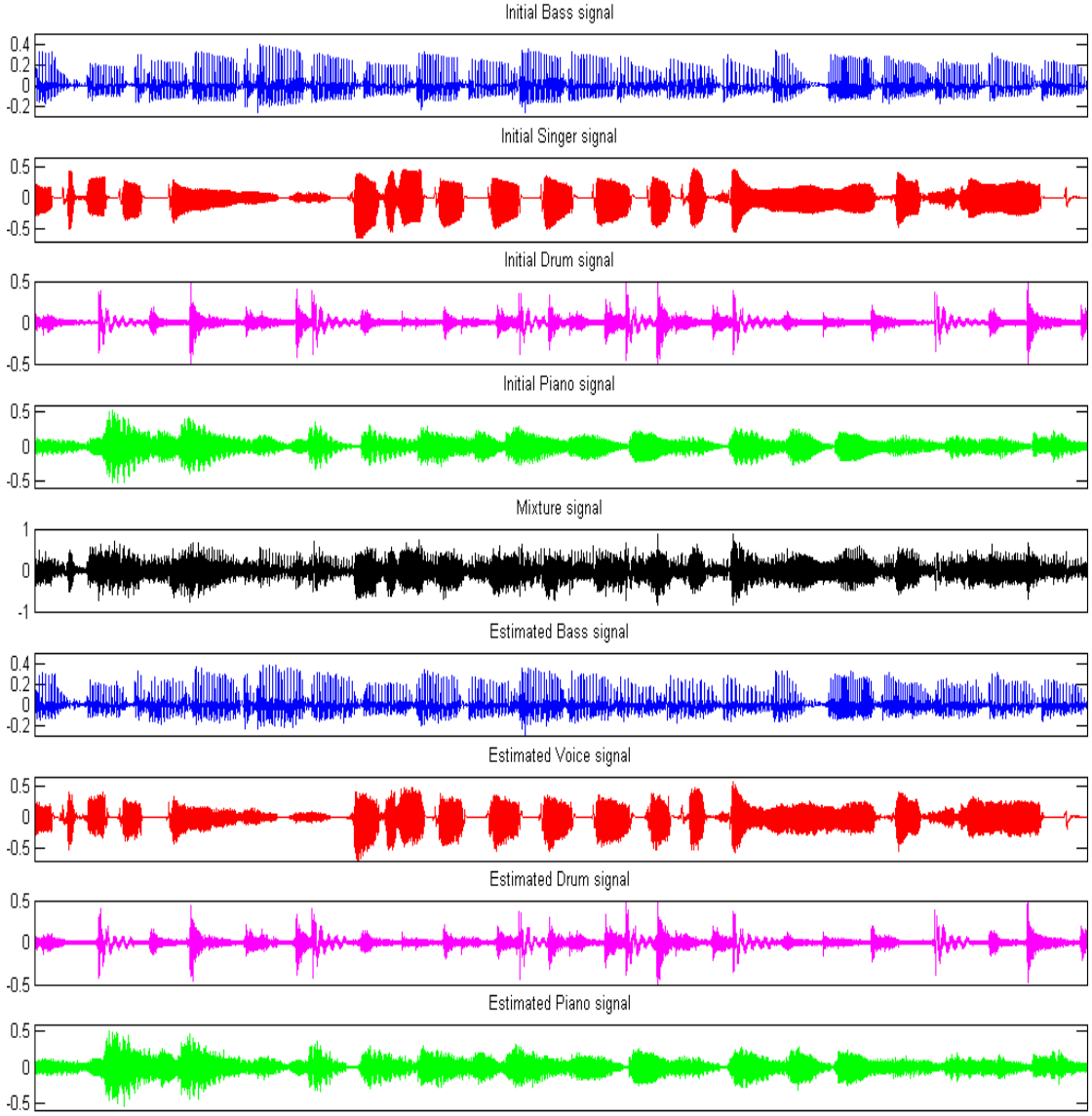


FIGURE 5.14 – Tracé temporel des quatre signaux sources du mélange 3I+1S, du mélange linéaire instantané correspondant, et des sources estimées.

Une autre illustration des bons résultats de séparation est donnée à la Figure 5.14 qui montre clairement la très forte similarité entre le tracé temporel des signaux sources originaux et leur estimées par séparation de sources informée par codage des signaux sources. Cette bonne séparation est par ailleurs confirmée par des tests d'écoute qui font apparaître une très bonne réjection des autres signaux sources pour chaque es-

timée : chaque source musicale reconstruite est clairement isolée. Bien sûr la qualité sonore n'est pas parfaite et un certain niveau de bruit musical demeure, plus ou moins significatif en fonction du signal source considéré et du nombre de sources à séparer. Ainsi, le bruit musical est plus faible lors de l'extraction de trois sources que lors de l'extraction de quatre sources. Les résultats fournissent une très bonne base pour une application de type *écoute active* : en effet, les signaux séparés peuvent être additionnés et même soustraits du mélange directement dans le domaine temporel. Les sources isolées peuvent ainsi être clairement rehaussées ou annulées, et le bruit musical ressortant sur chaque signal source individuellement est partiellement supprimé par le remixage à partir du mélange tatoué. Des exemples de sons sont disponibles à l'adresse suivante : `http ://www.icp.inpg.fr/~girin/WB-ISS-demo.rar`. Le package inclut les sources originales, les mélanges tatoués, et les sources estimées.

Une analyse plus détaillée de la Table 5.9 montre une augmentation des scores de ISNR entre l'extraction de deux ou trois sources à partir d'un mélange de trois sources de musique et à partir d'un mélange de quatre sources. Ceci est principalement du à une baisse des SNR d'entrées avec l'augmentation du nombre de sources composant un mélange, alors que dans le même temps, les SNR de sortie sont approximativement les même pour un mélange de trois ou de quatre sources. Deux explications peuvent être données : soit les capacités d'insertion de l'information sont similaires pour les mélanges de trois et quatre sources, soit la capacité est plus grande pour le mélange de quatre sources, mais le codage des descripteurs atteint une limite de précision (entre autre la limite de 10 bits des dictionnaires de forme). Chacune explique la faible diminution du SNR de sortie, et de qualité audio obtenus entre la séparation de deux et trois sources. Dans le cas de l'application CD-audio à laquelle nous nous intéressons, la séparation d'un grand nombre de sources sera privilégiée. C'est pourquoi nous ne nous attacherons pas plus à améliorer la qualité de séparation d'un faible nombre de sources, mais plutôt à augmenter le nombre de sources à séparer. À noter de plus que la séparation de sources informée par codage des sources devrait s'affranchir du type de mélange considéré, et permettre des résultats de séparation de qualité équivalente que le mélange soit instantané, anéchoïque, voire convolutif.

5.3 Compléments : la SSI-C appliquée à un mélange LIS stéréophonique avec un système de tatouage amélioré

5.3.1 Historique

Après avoir développé la SSI-C à partir d'un mélange monophonique, une autre technique de SSI a été étudiée durant cette thèse, basée non plus sur le codage des sources mais sur l'exploitation de la parcimonie des sources au sein d'un mélange stéréo-

phonique [20]. Cette nouvelle technique de SSI, nommée Séparation de Source Informée par Indexation (SSI-I) est développée dans le Chapitre 6. Parallèlement Jonathan Pinel a proposé, durant son stage de Master au GIPSA-lab de février à juin 2009 [Pinel et al., 2009], une amélioration du système de tatouage utilisé en SSI-C par l'utilisation d'un modèle psycho-acoustique (MPA) conduisant à une amélioration sensible des débits de tatouage. Le travail de thèse présenté dans ce manuscrit s'est limité à l'exploitation et non au développement de ce système de tatouage amélioré. Nous nous limitons ci-dessous à en donner le concept général. Pour plus de détails, le lecteur pourra se reporter à l'Annexe C et à [Pinel et al., 2009]. Nos recherches s'étant orientées vers une SSI à partir d'un mélange stéréophonique, et l'utilisation d'un MPA permettant d'accroître la capacité d'insertion de l'information, nous avons choisi de nous intéresser, brièvement, à une extension directe de la SSI-C au cas de signaux stéréophoniques avec le système de tatouage amélioré. Il s'agit de mesurer avant tout le gain de performances permis par l'utilisation d'un MPA. Le principe de la SSI-C développé dans le cadre monocanal est alors appliqué séparément sur les deux voies du signal de mélange, les différentes sources étant simplement distribuées sur ces deux voies. Nous présentons dans cette section de "nouvelles" performances de séparation obtenues par cette technique de SSI-C "étendue" couplant la multi-dimensionnalité du mélange à un système de tatouage plus performant.

5.3.2 Une brève présentation du système de tatouage basé sur un modèle psycho-acoustique (MPA)

5.3.2.1 Principe du modèle psycho-acoustique

L'utilisation d'un modèle psychoacoustique dans notre système de SSI a pour but d'accroître la quantité d'information de tatouage insérable de manière inaudible sur le signal de mélange. Pour ce faire, le MPA permet d'exploiter au plus près les caractéristiques TF du signal de mélange. L'utilisation d'un MPA permet en particulier de tenir compte du phénomène de masquage fréquentiel exploité notamment dans le domaine de la compression audio. Le masquage fréquentiel peut se résumer ainsi : lorsque deux sons de fréquences proches sont émis au même instant, en fonction de la puissance des deux sons, il est possible que seul le son de plus forte puissance soit entendu, et ce même si séparément chacun des deux sons est parfaitement audible [21].

Pratiquement, l'implémentation du MPA est réalisée à chaque trame temporelle en divisant le spectre en 32 *bandes* régulièrement espacées. Sur chacune des ces sous-bandes, un rapport signal à masque [22] est calculé, définissant une courbe de masquage globale sur l'ensemble du spectre. Toute modification par tatouage du signal de mélange située en dessous de cette courbe de masquage globale est supposée imperceptible

20. La très grande majorité des pistes audio (CD audio ou fichier wav) sont des pistes stéréophoniques.

21. Les deux sons sont situés au dessus de la courbe d'audition absolue qui désigne le volume minimal en dB nécessaire pour qu'un son pur à une fréquence donnée soit entendu.

22. prenant en compte le seuil de masquage absolu ainsi que le seuil de masquage fréquentiel

par l'oreille humaine. Il est donc possible de calculer, sur chaque trame et chaque sous-bande, le nombre maximal de bits insérables sur le mélange de manière inaudible. De plus, la capacité calculée par MPA offre l'avantage d'être ajustable aux besoins en ressources par un décalage de la courbe de masquage global (sur chaque trame temporelle). Un compromis peut ainsi être trouvé entre capacité d'insertion de l'information et qualité sonore du signal de mélange (plus la capacité d'insertion est grande, plus les coefficients MDCT du signal de mélange sont quantifiés grossièrement par la QIM, ce qui pénalise la qualité d'écoute des signaux).

5.3.2.2 Le système de tatouage amélioré

Le principe du tatouage par QIM reste similaire à celui introduit en 5.1.3. L'utilisation d'un MPA en amont de la phase de tatouage permet désormais de déterminer, sur chaque trame temporelle, et pour chaque sous-bande de fréquence, la résolution R_1 minimale avec laquelle peuvent être tatoués les coefficients MDCT de la trame et de la sous-bande en question, sans que cela soit audible. Alors que la résolution R_1 était statique en 5.1.3, elle est donc rendue dynamique (dépendante de la trame et de la sous-bande fréquentielle) par utilisation du MPA, en fonction de l'évolution spectro-temporelle du signal de mélange, et toujours sous contrainte d'inaudibilité de la watermark. La résolution fournie par le MPA étant (en particulier en moyennes et hautes fréquences) inférieure à 8 bits, la capacité d'insertion de l'information est augmentée en comparaison de celle obtenue avec $R_1 = 8$ bits dans nos travaux de la Section 5.2.

Rappelons que pour que le tatouage soit correctement retrouvé au décodeur, il est indispensable que les quantificateurs au décodeur soient identiques à ceux utilisés au codeur. Il faut donc, sur chaque trame, que les résolutions de quantificateurs dans les différentes sous-bandes puissent être retrouvées. Cette contrainte était assurée par la résolution R_1 fixe de 5.1.3. Cependant, l'insertion de l'information de tatouage au codeur modifie les coefficients MDCT de façon trop importante pour que le MPA désormais utilisé permette de retrouver, au décodeur, des résolutions des quantificateurs identiques à celles du codeur. La solution trouvée consiste à transmettre, à chaque trame, les valeurs de ces résolutions. L'insertion de cette information est faite dans les dernières sous-bandes fréquentielles avec une capacité fixe et modérée (connue au décodeur), l'oreille humaine étant peu sensible en hautes fréquences. Ainsi les 24 premières sous-bandes fréquentielles sont utilisées pour insérer sur le mélange le codage des signaux sources, tandis que les 8 dernières sous-bandes sont utilisées pour transmettre valeurs des capacités des quantificateurs utilisés dans les premières sous-bandes. En conséquence, le bloc de quantification des maxima des MDCT par canal fréquentiel du schéma 5.1 a disparu dans le schéma 5.15 présentant le système de SSI-C stéréo utilisant un MPA. Par volonté de compacité, seul le schéma pour la voie de mélange gauche x_L est présenté. Le schéma est rigoureusement identique pour la voie de droite.

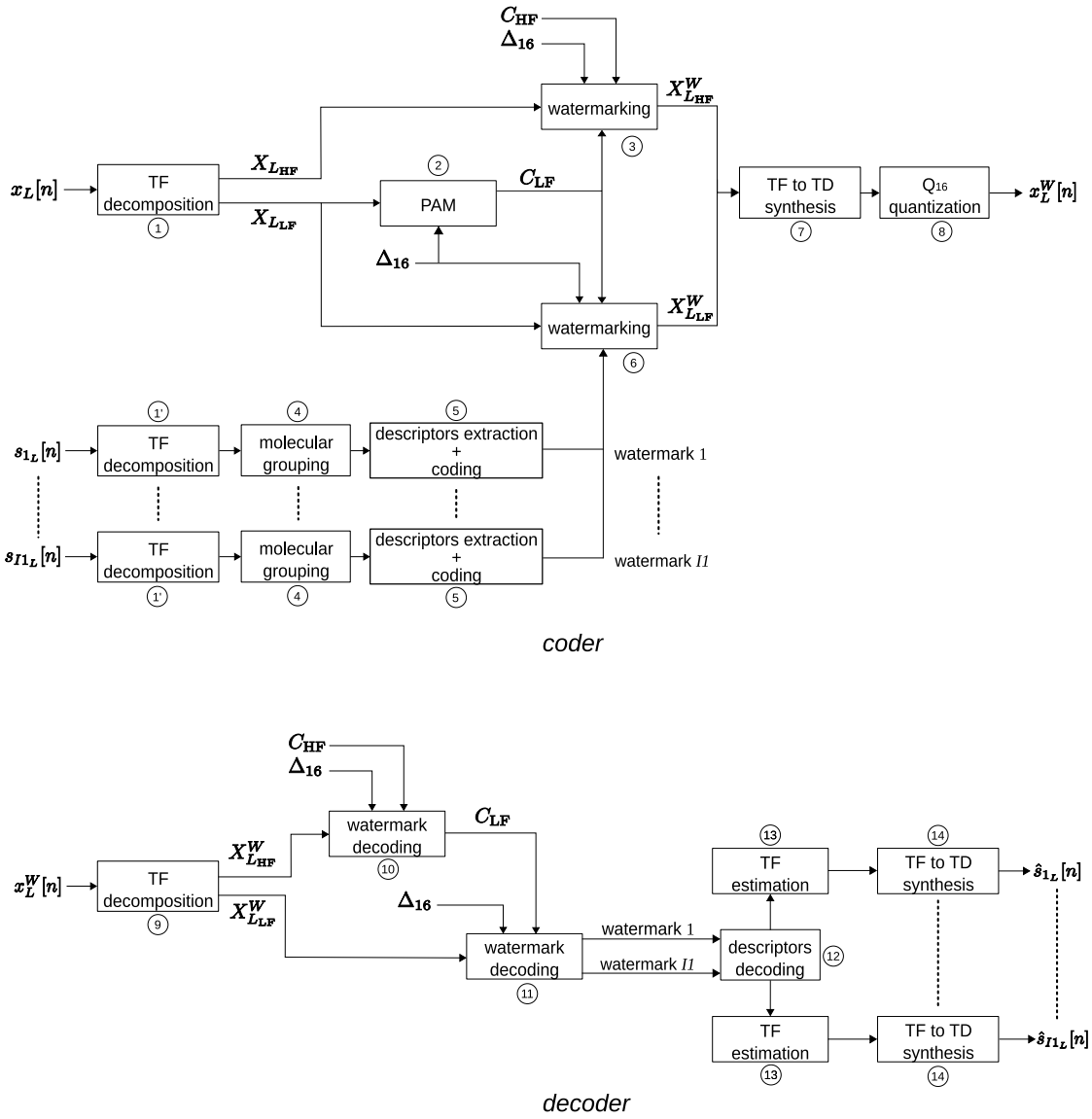


FIGURE 5.15 – schéma du système de SSI partir d'un mélange stéréophonique par codage des signaux sources, et utilisant le modèle psychoacoustique développé dans [Pinel et al., 2009].

5.3.3 Nouveaux résultats

Nous détaillons maintenant les performances de séparation obtenues pour la séparation de signaux musicaux composés de quatre sources, comme introduit à la Section 5.2, mais cette fois à partir d'un mélange linéaire instantané stationnaire stéréophonique. La matrice normalisée de mélange utilisée est :

$$\mathbf{A} = \begin{bmatrix} 0.93 & 0.8 & 0.6 & 0.37 \\ 0.37 & 0.6 & 0.8 & 0.93 \end{bmatrix} \tag{5.28}$$

Les sources $s_1, \ldots, s_4$ sont respectivement une *guitare basse*, une *voix*, une *batterie*, et un *piano*. De par le choix de la matrice $\mathbf{A}$, les sources s_1 et s_2 sont plus accentuées sur la voie de gauche tandis que les sources s_3 et s_4 sont plus accentuées sur la voie de droite. Les tests sont ici effectués sur des portions de signal de 5 secondes environ.

5.3.3.1 Une capacité de tatouage accrue

L'utilisation d'un MPA couplée au codeur au tatouage QIM permet, en adaptant la résolution de quantification R_1 au signal hôte, d'accroître la capacité d'insertion du signal tout en préservant la contrainte d'inaudibilité de la watermark. Une comparaison entre les capacités d'insertion est faite entre trois configurations :

- Une configuration où, comme à la Section 5.1.3.3, R_1 est constante au cours du temps et de la fréquence, et fixée à 8 bits, notée SSI-C_f .
- Une configuration intermédiaire où R_1 est fixée a priori par paliers (3 paliers sur l'ensemble du spectre), de manière à tenir très grossièrement compte de la sensibilité de l'oreille humaine en fonction de la fréquence, notée SSI-C_p . De 0 à 1.5kHz, $R_1 = 8$ bits, de 1.5 à 15kHz, $R_1 = 7$ bits, et enfin après 15kHz, $R_1 = 5$ bits.
- Une configuration où le MPA décrit en 5.3.2.1 est utilisé. Les notations SSI-C_{m10} et SSI-C_{m6} désignent deux variantes de cette configuration avec un décalage respectivement de -10dB et -6dB par rapport au seuil de masquage global (la capacité d'insertion obtenue avec SSI-C_{m6} est supérieure à celle obtenue avec la configuration SSI-C_{m10}). Si dans les deux cas de figure l'inaudibilité du tatouage est assurée, la configuration SSI-C_{m10} offre une "marge de sécurité" supplémentaire.

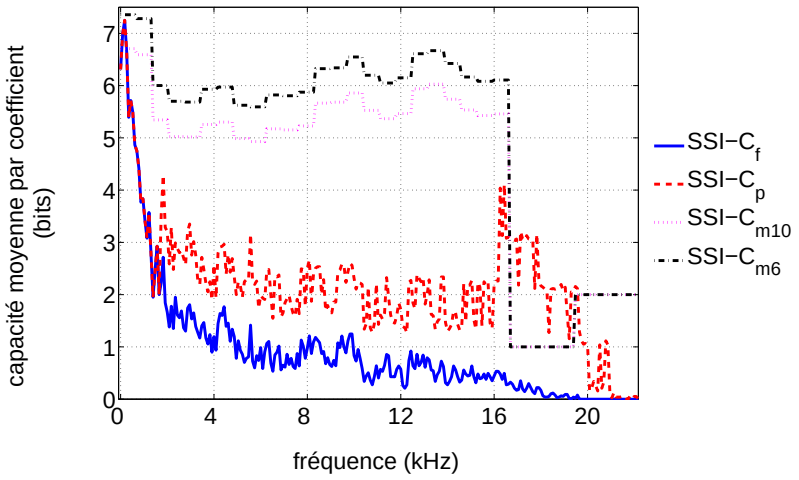


Figure 5.16 – Capacité moyenne par coefficient en fonction de son canal fréquentiel pour les quatre méthodes de calcul de R_1 décrites en 5.3.3.1, calculées sur 5 extraits de 10 secondes de mélanges de musique grand public.

Les différentes capacités obtenues sont présentées Figure 5.16. Le gain de capacité obtenu par une résolution R_1 variable y apparaît clairement : alors que la capacité obtenue avec R_1 fixe chute très rapidement avec la fréquence, la capacité calculée grâce à un MPA est approximativement constante jusqu'à 16kHz [23]. Pour des moyennes et

23. À partir de 16kHz sont encodées les capacités des 24 premières sous-bandes, avec une capacité fixe de 1 puis 2 bits.

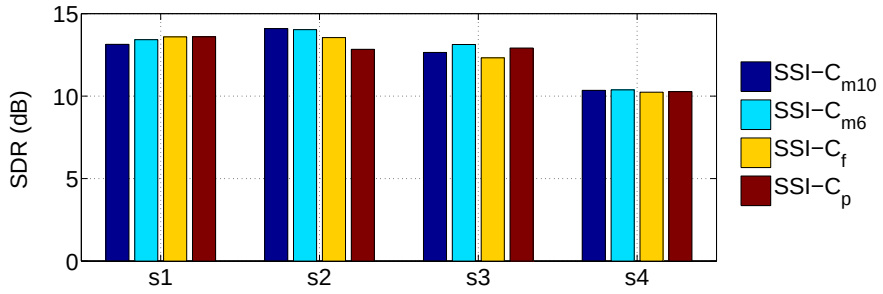
hautes fréquences, une résolution de 8 bits apparaît comme beaucoup trop fine par rapport aux contraintes d'inaudibilité : certaines zones hautes fréquences, influant peu sur la qualité audio globale d'un signal ne nécessitent pas d'être encodées avec une telle précision. À l'inverse, au niveau des très basses fréquences (les tous premiers canaux fréquentiels), là où les capacités les plus élevées sont mesurées, la capacité obtenue par MPA (SSI-C$_{m10}$) est ici potentiellement inférieure à celle obtenue pour $R_1 = 8$ bits, ce qui traduit le fait que dans cette partie du spectre, la résolution R_1 de 8 bits peut être, cette fois, plus faible que celle retournée par le MPA. Quantifier ces très basses fréquences sur 8 bits peut engendrer une altération de la qualité audio du signal correspondant à cette portion du spectre. Notons que la configuration R_1 fixe par paliers, offre une évolution intermédiaire de la capacité, qui, bien que chutant rapidement avec la fréquence, atteint en moyennes fréquences des capacités plus élevées que celles obtenues pour la configuration R_1=8bits (1 bit entre 1.5 et 16.5kHz, puis 3 bits après 16.5kHz). Les brusques sauts de capacité observés sur la courbe de la configuration R_1 fixe par paliers correspondent aux changements de résolution (saut d'un palier à l'autre à 1.5 et 16.5kHz). Cependant, ces niveaux, fixés empiriquement, sont généraux et n'offrent pas l'adaptabilité au contenu audio que peut offrir un MPA.

5.3.3.2 Des performances accrues

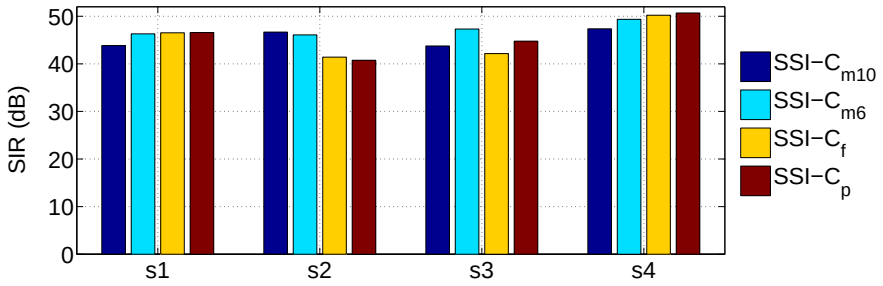
Le schéma de codeur/décodeur de la Figure 5.15 étant identique au schéma 5.1 du Chapitre 5 pour les étapes ne concernant pas le calcul de la capacité ou l'insertion/extraction du tatouage, nous présentons directement les performances de séparation obtenues pour la séparation de quatre sources musicales à partir de leur mélange LIS stéréophonique.

Les performances de séparation sont données pour les deux approches de SSI-C avec R_1 fixe, et R_1 déterminée par MPA (et deux réglages de paramètres sont testés pour chaque méthode).

Les mesures objectives de qualité introduites à la Section 5.2.2 sont effectuées pour chacune des quatre configurations, et les résultats sont présentés Figure 5.17. Les scores de SIR étant particulièrement élevés, les mesures de SDR et SAR sont quasi identiques (cf équations (5.22) et (5.24)), et en conséquence, les courbes de SAR ne sont pas présentées. Notons avant tout, pour l'ensemble des mesures, des scores élevés traduisant la qualité de séparation par SSI-C pour un mélange stéréo. Les SDR d'entrée sont respectivement de -11.9, -7.1, -15.5 et -12.2dB pour la guitare basse, le chant, la batterie et le piano. Des SDR de sortie compris entre 10 et 14dB traduisent une (très) bonne séparation de l'ensemble des sources. Une forte réjection des interférences dues aux autres sources transparaît avec des scores de SIR supérieurs à 40dB. Enfin des scores de SAR compris entre 10 et 14dB, sont concordants avec la qualité de séparation globale obtenue pour chacun des quatre signaux du mélange, même si de légers artefacts demeurent sur les signaux estimés individuels. Cependant, l'analyse comparative des performances entre les quatre configurations ne révèle pas, sur des mesures objectives globales (sur l'ensemble du spectre), de différences significatives entre les deux mé-



(a) SDR.



(b) SIR.

Figure 5.17 – Comparaison des performances de séparation de 4 sources à partir d'un mélange linéaire instantané de ces 4 sources avec ou sans MPA. s_1, s_2, s_3 et s_4 sont respectivement une *guitare basse*, une *voix*, une *batterie*, et un *piano*.

thodes SSI-C classique et SSI-C couplée à un MPA. Ceci est d'autant plus vrai pour les instruments *guitare basse* et *piano*. Ces deux instruments sont ceux possédant la plus faible variabilité spectrale. La majorité de leur énergie est concentrée en basses voire très basses fréquences pour la basse. Or dans ces gammes de fréquences, les deux méthodes avec ou sans MPA offrent des capacités élevées et relativement similaires. A contrario, l'apport d'un MPA transparaît d'autant plus que les sources présentent une large palette spectrale. C'est le cas des signaux *chant*, et *batterie*. La décomposition MDCT de ces signaux présentent des coefficients d'énergie plus importante en moyennes voire hautes fréquences, là où le MPA, en permettant d'obtenir une plus forte capacité d'insertion, permet d'utiliser des dictionnaires de formes, ce que ne permet pas la méthode SSI-C$_f$. Cependant, ces mesures de SDR sur la globalité du signal ne permettent pas d'apprécier à sa juste valeur l'apport du MPA sur l'amélioration de la qualité (en particulier audio) des signaux séparés. Du fait du spectre décroissant en fréquences, l'allure générale d'un signal audio temporel est largement déterminée par ses composantes basses fréquences. Par conséquent, toute modification significative des coefficients MDCT basses fréquences d'un signal audio altère fortement sa forme d'onde, alors qu'une détérioration proportionnellement identique des coefficients MDCT en hautes fréquences n'a que de faibles conséquences sur la forme d'onde temporelle du signal. Autrement dit, une modification des coefficients MDCT basses fréquences a de fortes conséquences sur toute mesure de type rapport-signal-à-bruit, alors qu'une modification du même ordre de grandeur des coefficients hautes fréquences n'a que peu

d'effet sur une telle mesure. Pour percevoir l'apport d'une capacité plus adaptée au signal hôte en moyenne et haute fréquence, il est nécessaire de faire une étude des performances de la séparation par sous-bande fréquentielle. Ce sont ces résultats qui sont présentés Figure 5.18 pour le signal source *chant*.

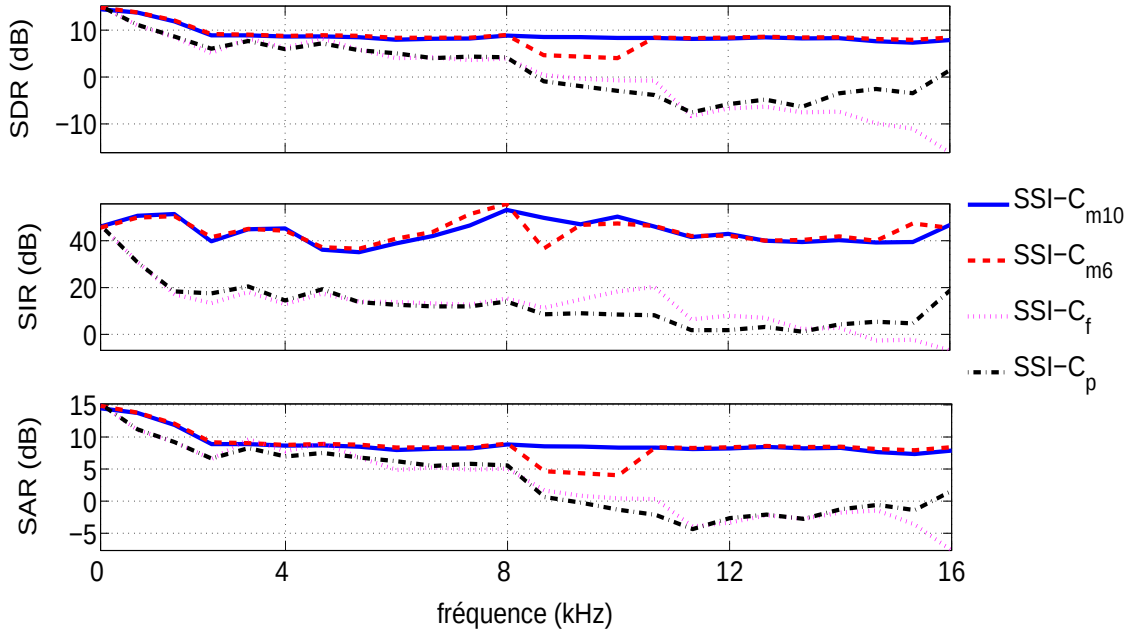


FIGURE 5.18 – Performances de séparation par sous-bande fréquentielle des 4 méthodes SSI-C$_f$, SSI-C$_p$, SSI-C$_{m10}$ et SSI-C$_{m6}$, pour le signal de *chant*.

Ainsi, les résultats globaux de la Figure 5.17 qui semblaient similaires entre les différentes configurations, apparaissent Figure 5.18 très différents, et révèlent que l'utilisation d'un MPA améliore considérablement les performances de séparation en moyennes et hautes fréquences. Alors que lorsqu'un MPA est utilisé pour calculer la capacité d'insertion du tatouage les performances sont approximativement constantes en fonction de la fréquence, ces performances chutent rapidement dès 2kHz dans le cas d'une résolution R_1 fixe (soit tout au long du spectre soit par paliers). Les scores de performances entre la configuration SSI-C$_f$ et les configurations utilisant un MPA ne sont semblables que pour les toutes premières sous-bandes (capacités similaires avec ou sans MPA). Ceci s'explique par le fait que les capacités disponibles à partir de la 3 ou 4-ième sous-bande décroissent très rapidement dans le cas de la configuration SSI-C$_f$, ne permettant pas le codage du descripteur de forme pour les sous-bandes de fréquences supérieures. À partir de la 3 ou 4-ième sous-bande fréquentielle, les molécules des signaux sources sont alors très majoritairement reconstruites à partir de molécules de mélange (seul le descripteur de gain peut être encodé), qui peuvent être très différentes des molécules du signal source.

Les courbes de performances sont logiquement corrélées aux courbes de capacités présentées Figure 5.16. L'intérêt des méthodes de SSI-C utilisant un MPA apparaît plus encore lors de tests d'écoute. En effet c'est la configuration SSI-C$_{m10}$ qui offre les

meilleurs résultats quant à la qualité des signaux séparés. Les signaux estimés sont plus fidèles aux signaux sources, ils présentent moins de distorsions et une plus grande réjection des interférences, tout en minimisant le bruit musical nettement plus présent dans la configuration SSI-C$_f$.

À l'inverse, la meilleure réjection des interférences et la minimisation des distortions pour les méthodes utilisant un MPA s'explique par l'utilisation plus fréquente de dictionnaires de forme dans les configurations SSI-C$_{m10}$ et SSI-C$_{m6}$. La Figure 5.19 indique les molécules où le descripteur de forme est encodé (carré blanc dans le cas où le descripteur de forme n'est pas encodé), et la taille du dictionnaire dont la molécule est alors extraite. Alors que pour la configuration SSI-C$_f$ le descripteur de forme n'est encodé que pour les premiers canaux fréquentiels, il l'est systématiquement sur (quasiment) l'ensemble du spectre dans le cas de la configuration SSI-C$_{m10}$. À noter qu'une majorité des dictionnaires utilisés sont de taille 10 bits, ce qui correspond à la qualité maximale de codage du descripteur de forme (cf. Table 5.1).

Notons de plus que même si l'utilisation d'un MPA implique une quantification plus grossière du signal, et donc un signal de mélange tatoué plus altéré qu'après tatouage par la méthode SSI-C$_f$, le recours aux dictionnaires de forme permet de s'affranchir de la moindre qualité du signal de mélange tatoué (les signaux estimés sont reconstruits à partir de prototypes de forme des dictionnaires).

	SSI-C$_f$	SSI-C$_p$	SSI-C$_{m10}$	SSI-C$_{m6}$
Left mixture	43,8	42,2	33,4	29,4
Right mixture	44,2	41,9	32,8	28,9

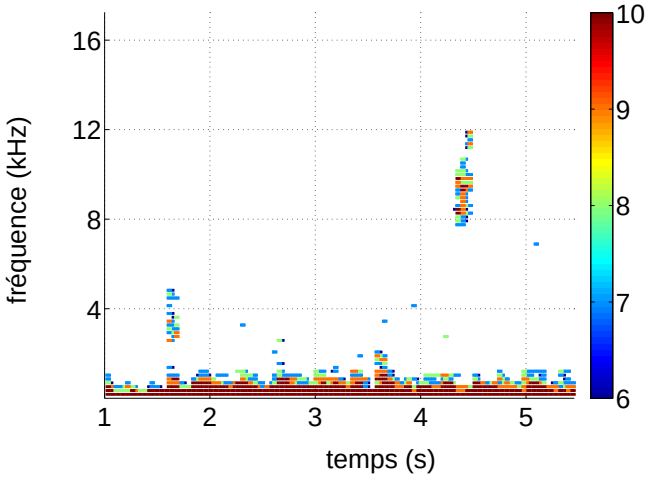
Table 5.10 – Comparaison de la dégradation du signal de mélange par tatouage pour deux méthodes de séparation, avec ou sans MPA. Rapport-signal-à-bruit entre les mélanges tatoués et les mélanges originaux (dB).

La Table 5.10 fournit des mesures de rapport-signal-à-bruit entre les mélanges (gauche et droit) avant et après tatouage, pour les quatre configurations présentées plus tôt. Notons que même si les valeurs de SNR sont plus faibles pour les configurations SSI-C$_{m6}$ et SSI-C$_{m10}$ que pour les configurations à résolution fixe, l'utilisation même d'un MPA assure l'inaudibilité des dégradations subies par le signal, compte tenu de valeurs de SNR autour de 30dB et surtout du fait que le bruit est dans ce cas "mis en forme" par le MPA.

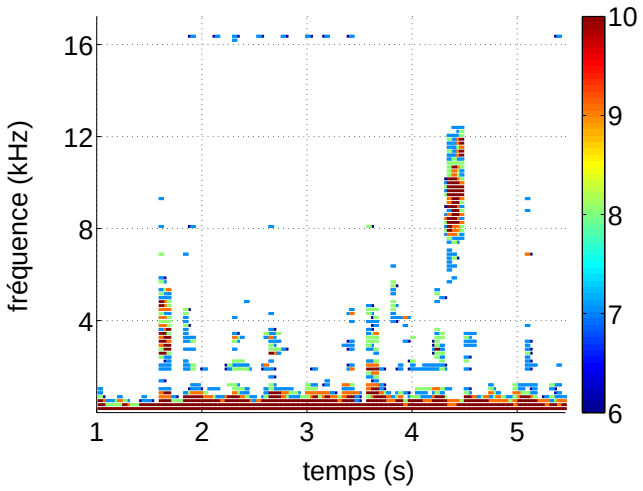
5.4 Conclusion sur la séparation de sources informée par codage des signaux sources

Nous avons vu au cours de ce chapitre une première implémentation d'une technique de séparation de sources informée : la SSI par codage des signaux sources. Une première approche à partir d'un mélange LIS monocanal et une capacité de tatouage

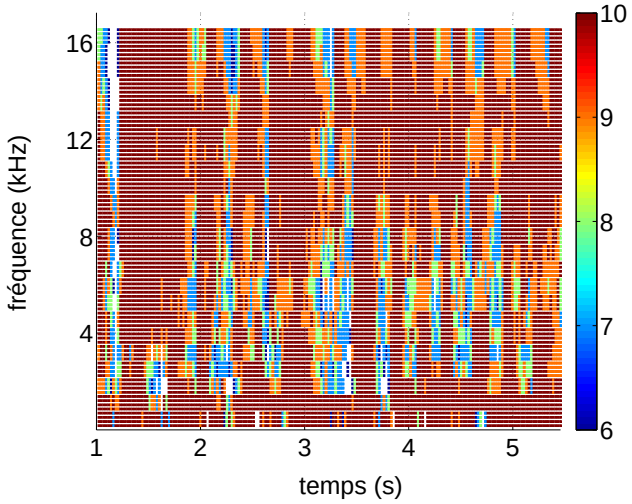
sous-optimale a permis de mettre en évidence les perspectives prometteuses de cette technique de séparation de sources utilisant une importante information a priori sur les signaux sources et le processus de mélange. Un raffinement du calcul de la capacité d'insertion du tatouage par l'emploi d'un MPA parallèlement à l'utilisation d'un mélange LIS stéréophonique a permis de confirmer ces premiers résultats encourageants. En effet, l'augmentation de la ressource par recours à un MPA et le doublement du nombre de voies, et donc de la capacité par rapport à un mélange monocanal, ont permis de significativement augmenter les performances de séparation par un meilleur codage des signaux sources. Les signaux obtenus, s'ils ne sont pas toujours "utilisables" individuellement car ils présentent encore quelques artefacts ou interférences de sources concurrentes, sont cependant d'une qualité globale très supérieure à la qualité généralement obtenue par des méthodes de séparation *aveugle* de sources. L'intérêt de l'approche *informée* est indiscutable pour des mélanges sous-déterminés de signaux présentant un large recouvrement spectro-temporel tels que les signaux de musique. Une des améliorations possibles de cette technique de SSI-C est très vraissemblablement la résolution même de l'analyse TF du signal. En effet, il semble que le choix d'une plus grande fenêtre d'analyse (jusqu'à présent cette fenêtre est choisie de taille 512 échantillons temporels), qui apporterait une meilleure résolution fréquentielle, soit nécessaire pour capter plus en détails la structure des signaux de musique et exploiter leur parcimonie dans le plan TF. Parallèlement, de manière à limiter encore la superposition des sources, et comme les capacités d'insertion de l'information peuvent être augmentées par l'utilisation d'un MPA, il semble pertinent de réduire la taille des molécules. Ces deux points seront plus amplement étudiés au cours du chapitre suivant dans le cadre d'une nouvelle approche de SSI, basée sur la parcimonie des signaux sources, et sur l'exploitation de la stéréophonie.



(a) Configuration SSI-C$_f$



(b) Configuration SSI-C$_p$



(c) Configuration SSI-C$_{m10}$

Figure 5.19 – Molécules où le descripteur de forme est encodé pour la source *chant* dans un mélange à 4 sources, lorsque la capacité d'insertion est calculée avec ou sans MPA. La couleur indique la résolution des dictionnaires de forme.

Chapitre 6

La séparation de sources informée par indexation des sources prédominantes

Dans ce chapitre nous introduisons une nouvelle forme de séparation de sources informée qui, contrairement à celle développée au Chapitre 5, ne repose pas sur le codage des signaux sources. L'élément clé de cette nouvelle méthode de SSI est l'utilisation de la parcimonie des sources dans le domaine temps-fréquence, telle que décrite à la Section 2.5.1.3, pour permettre une estimation des signaux sources par inversion locale d'un mélange initialement sous-déterminé (cette inversion est habituellement limitée à la configuration déterminée). Pour ce faire, l'information guidant la séparation, et transmise par tatouage, est ici l'index des sources localement prédominantes dans le mélange. Ceci justifie l'appellation de "Séparation de Sources Informée par Indexation des sources" que l'on note SSI-I. Alors que la SSI par codage est particulièrement adaptée au cas d'un mélange mono-canal, la SSI-I met à profit la dimension multi-canal du mélange pour séparer les sources localement. En conséquence, il est possible de séparer globalement significativement plus de sources qu'il n'y a de signaux de mélange. Nous continuons dans ce chapitre et plus que jamais à viser tout particulièrement l'application du CD-audio, et plus généralement le format audio PCM 16-bits, et nous nous plaçons donc dans la configuration d'un mélange linéaire instantané stationnaire *stéréophonique* (LISS) sous-déterminé. La forme d'un tel mélange est :

$$\begin{bmatrix} x_L(t) \\ x_R(t) \end{bmatrix} = \begin{bmatrix} a_{L1} & a_{L2} & \cdots & a_{LI} \\ a_{R1} & a_{R2} & \cdots & a_{RI} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_I(t) \end{bmatrix} \tag{6.1}$$

Nous détaillons dans les sections suivantes les différentes étapes de la technique de SSI-I, en particulier la phase de sélection des sources prédominantes et le procédé de séparation. Enfin des résultats obtenus pour des signaux de musique sont donnés à la Section 6.3 puis comparés à ceux obtenus par une méthode de séparation aveugle de sources développée dans [Bofill and Zibulevski, 2001], et aux performances optimales

atteignables calculées au moyen d'estimateurs oracles développés dans [Vincent et al., 2007] (dont nous détaillerons les principes).

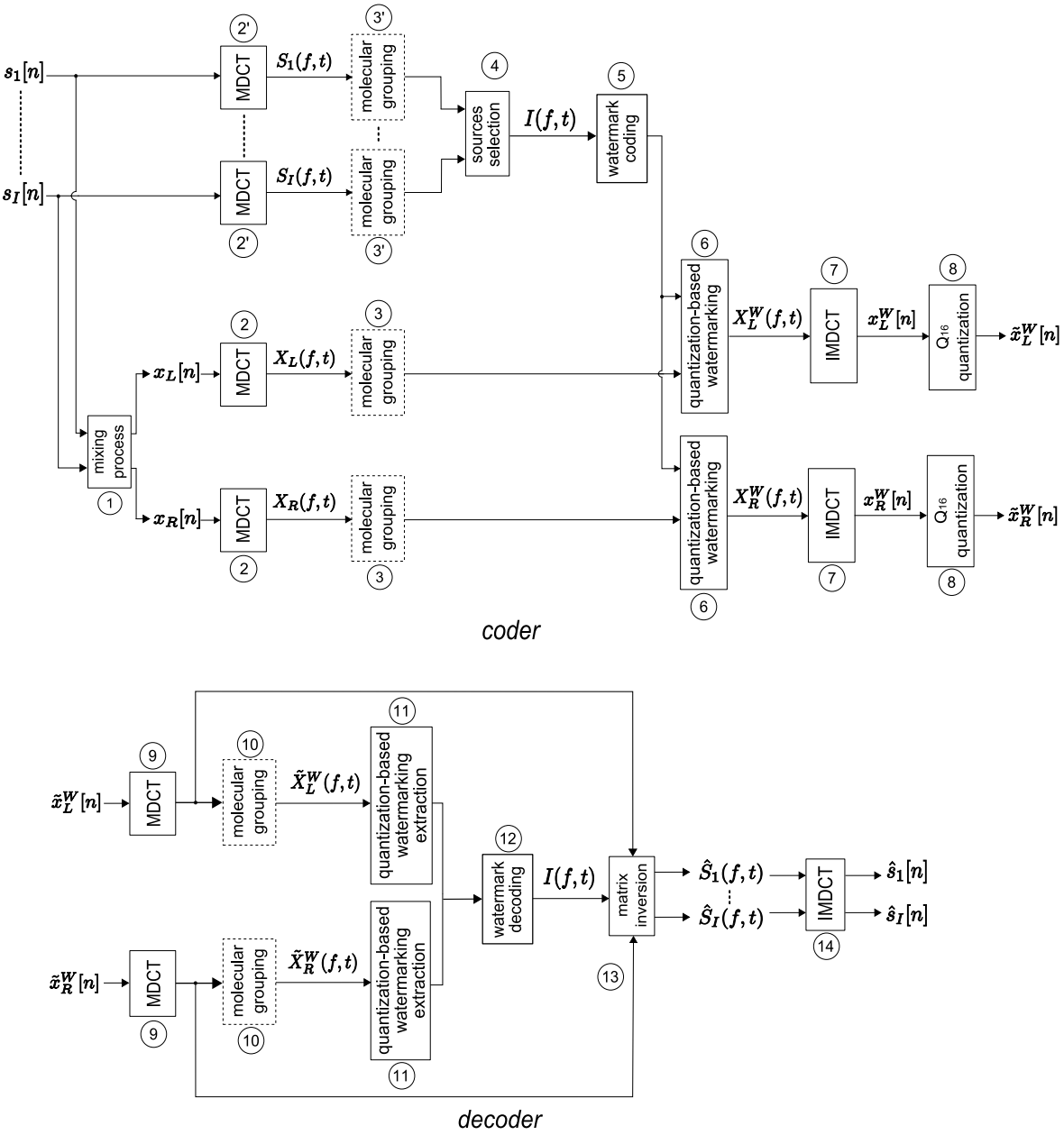


Figure 6.1 – Le bloc Codeur/Décodeur du procédé de SSI par tatouage de l'index des sources actives.

6.1 Principes de la SSI par tatouage des index des sources actives

L'architecture duale codeur/décodeur introduite au Chapitre 4, développée au Chapitre 5 dans le cadre de la SSI par codage des sources, et propre à la séparation informée en général, est bien évidemment conservée dans le cadre de la SSI-I. Ce principe est ici illustré par le nouveau codeur/décodeur de la Figure 6.1. Dans cette section, nous ne

présentons que les grands principes de la SSI-I, les différents blocs du schéma seront quant à eux présentés dans la section suivante, en détaillant particulièrement les blocs propres à cette nouvelle approche par indexation.

En SSI-I, les signaux sources sont toujours supposés connus avant le processus de mélange, réalisé au bloc 1, et le traitement a toujours lieu dans le domaine TF (transformée MDCT aux blocs 2, 2' et 9). Cependant l'information qui est extraite des signaux sources, et la manière dont elle est exploitée, diffère de celle extraite dans le cas de la SSI par codage des signaux sources. Alors que précédemment un "codage source-canal" permettait de décrire localement et "individuellement" le contenu de chaque signal source, la SSI-I se fonde sur l'estimation de ces sources à partir du mélange par un processus d'inversion locale. Pour ce faire, la SSI-I s'appuie essentiellement sur l'hypothèse de parcimonie de signaux sources dans le plan temps-fréquence pour réduire localement la dimension du mélange. Au codeur, une analyse temps-fréquence de chaque signal source permet de sélectionner localement les sources prédominantes, dans chaque région du plan temps-fréquence. Cette analyse est réalisée au bloc 4 de la Figure 6.1 selon un critère discuté en détails à la Section 6.2.2. Le mélange, initialement de taille $J \times I$ dans le cas où l'on dispose de J observations d'un mélange de I sources, avec $J < I$, est réduit, en chaque région du plan temps-fréquence, à une taille $J \times I'$, avec $I' \in [1, \cdots, J]$. I' représente le nombre de sources participant le plus activement au mélange dans chaque région temps-fréquence, et on suppose donc que ce nombre est toujours inférieur ou égal au nombre d'observations. Ces I' sources sont alors estimées par un processus d'inversion locale du mélange. Dans notre application cible audio, $J = 2$ mais le principe vaut pour d'autres configurations de mélange LISS.

Ce principe est directement inspiré des méthodes de séparation aveugles ou semi-aveugles telles que mentionnées à la Section 2.5 basées sur la parcimonie dans le plan temps-fréquence. On peut à nouveau citer en particulier l'étude de Bofill et Zibulevski [Bofill and Zibulevski, 2001] et les travaux dérivés sur des principes similaires [1]. L'intérêt fondamental de la SSI-I par rapport à ces méthodes est la connaissance exacte du contenu du mélange (la contribution des différentes sources) et donc la garantie de la bonne sélection des sources prédominantes. Ceci n'est pas garanti par un critère semi-aveugle de type "plus court chemin" (cf Section 2.5.2) utilisé dans [Bofill and Zibulevski, 2001] : le plus court chemin n'est pas toujours le bon. De plus, pour réaliser l'inversion, la matrice de mélange doit être connue ou estimée au décodeur avec le plus de précision possible. L'estimation de cette matrice constitue une première phase très délicate du processus de séparation pour les méthodes aveugles. À l'inverse, dans l'approche informée, on s'affranchit de cette contrainte car on peut supposer que la matrice de mélange est connue au décodeur : on peut soit la transmettre directement par tatouage, soit transmettre par tatouage des informations peu coûteuses permettant

1. L'hypothèse d'un nombre de sources actives inférieur ou égal au nombre d'observations du mélange semble plus réaliste que l'hypothèse d'une seule source active qui est à la base des méthodes de séparation par masquage binaire telle que [Yilmaz and Rickard, 2004] [Araki et al., 2007] [Araki et al., 2007]. Pour des exemples de séparation avec deux sources actives, voir aussi [Linh-Trung et al., 2005].

de la reconstruire (typiquement, si les coefficients de la matrice suivent un modèle paramétrique, on transmet les paramètres de ce modèle). Dans tout ce chapitre, comme le mélange est instantané, le nombre de coefficients est réduit (coefficients réels constants en temps et en fréquence), et leur coût de transmission est négligeable devant les autres informations transmises. Dans tout ce chapitre, la matrice est donc supposée connue au décodeur.

Dans le système proposé, un message encodant l'index des sources prédominantes sélectionnées dans chaque région TF est formé au bloc 5, inséré sur les signaux de mélange par tatouage (bloc 6), puis les signaux de mélange tatoués sont transformés dans le domaine temporel (bloc 7). Enfin, les signaux temporels tatoués sont convertis au format PCM 16-bits lors de la dernière étape du codeur (bloc 8), comme dans le système des chapitres précédents.

L'estimation des signaux sources au décodeur est réalisée dans le domaine temps-fréquence (bloc 13) en exploitant l'index des sources extrait (bloc 11) et décodé (bloc 12) à partir du signal de mélange, et en appliquant l'inversion locale du mélange sur les sources sélectionnées. Enfin, les signaux estimés dans le domaine temps-fréquence sont transformés dans le domaine temporel (bloc 14). Nous verrons en détails par la suite que l'étape de *molecular grouping* introduite au Chapitre 4 et développée au Chapitre 5 est ici optionnelle (bloc 3, 3' et 10, ce caractère optionnel étant symbolisé par le tracé en pointillés).

6.2 Implémentation

6.2.1 Décomposition MDCT vs. décomposition STFT

La méthode de séparation proposée dans ce chapitre s'appuyant largement sur la parcimonie des sources, une représentation permettant de concentrer l'énergie des différentes sources sur un nombre limité de coefficients est particulièrement utile. Or le domaine temps-fréquence offre, comme nous l'avons vu dans les chapitres précédents, une représentation des signaux beaucoup plus parcimonieuse que le domaine temporel. C'est pourquoi l'ensemble des traitements de sélection des sources prédominantes et de séparation est réalisé en SSI-I dans le domaine temps-fréquence, avec une transformation linéaire permettant de conserver la structure LISS du mélange.

La SSI-I exploitant la parcimonie des sources, la superposition des signaux est un des principaux facteurs limitants des performances de séparation en SSI-I, bien plus qu'en SSI-C où le signal de mélange n'est pas systématiquement utilisé pour décoder les signaux sources (dans les zones de forte énergie, des prototypes de formes sont alors utilisés). Dans le but de limiter au maximum le recouvrement des signaux sources, le premier paramètre sur lequel interagir est le choix de la transformée temps-fréquence utilisée au codeur et au décodeur. Plus la transformée utilisée possède des capacités de concentration de l'énergie du signal élevées, plus le risque de superposition des sources est limité. Nous reprenons donc la transformée MDCT utilisée dans le chapitre 5, et

nous réalisons (seulement) ici une expérience préliminaire permettant de vérifier que la MDCT est plus efficace que la transformée de Fourier à court terme (STFT) en terme de limitation du recouvrement des sources pour des signaux de musique (une telle expérience a déjà été réalisée pour des signaux de parole dans [Aoki et al., 2001]). Rappelons cependant que la MDCT est une transformée orthogonale alors que la STFT est redondante d'un facteur deux, et que par conséquent, les scores de superposition de la MDCT seront d'autant meilleurs que le nombre de source sera élevé. Au contraire les meilleurs scores de superposition de la STFT seront obtenus pour un faible nombre de sources. Pour cette expérience, nous utilisons un mélange de $N = 4$ sources (3 instruments de musique et une voie chantée). Notons que la fenêtre d'analyse utilisée pour la transformée MDCT est une fenêtre de Kaiser-Bessel dérivée de taille $W = 2048$ et une fenêtre de Hanning de $W = 8192$ pour la STFT, en accord avec les dimensions optimales déterminées dans [Aoki et al., 2001] et [Vincent et al., 2007] et [Nesbit and Plumbley, 2008] pour limiter la superposition des sources musicales dans le cadre d'un mélange linéaire instantané. La parcimonie des deux transformées est mesurée par le critère de norme l_ε^0 introduit dans [Karvanen and Cichocki, 2003] [Rickard, 2006], également utilisé dans [Araki et al., 2007] et défini par

$$\| s(f,t) \|_{0,\epsilon(f)} = \text{card}(\{i, |s_i(f,t)| \geq \epsilon(f)\}) \tag{6.2}$$

La norme l_ε^0 représente donc le nombre de signaux source d'amplitude supérieure à ε. La Figure 6.2 fournit, sur chacun des graphes, le pourcentage moyen de trames temporelles pour lesquelles $\| s(f,t) \|_{0,\epsilon(f)} = 0, \cdots, I$, avec $\epsilon(f) = \frac{1}{100} \max_i \max_t |s_i(f,t)|$, soit un seuil à -40dB qui permet d'assurer une qualité audio des signaux après seuillage très proche de celle des signaux sources initiaux. Il apparaît clairement Figure 6.2 que la MDCT offre des propriétés de compacité de l'énergie spectrale supérieures à la transformée de Fourier, ce qui se traduit par un nombre de sources ayant simultanément une énergie significative inférieure à I dans la plupart des cas. Cette propriété est évidemment extrêmement intéressante dans le cas où l'on souhaite réduire localement la dimension du mélange, ce qui est à la base de la méthode proposée dans ce chapitre. Alors que pour la STFT, le nombre de bin temps-fréquence pour lequel trois sources sont actives est de l'ordre de 40%, il est seulement de 10 à 15% pour la MDCT. Or, une estimation des signaux sources par inversion exacte du mélange ne peut être faite qu'aux bins temps-fréquence où le mélange est déterminé. Dans le but de maximiser le nombre de régions temps-fréquence où le mélange est déterminé, l'usage de la MDCT plutôt que de la STFT est pleinement justifié.

6.2.2 Détermination des sources actives : combien et lesquelles ?

6.2.2.1 Principe

En SSI-I tout comme dans la méthode de séparation semi-aveugle de sources [Bofill and Zibulevski, 2001], l'estimation des signaux sources est réalisée par inversion locale du processus de mélange. *Locale* signifie que le processus est réalisé à chaque portion du

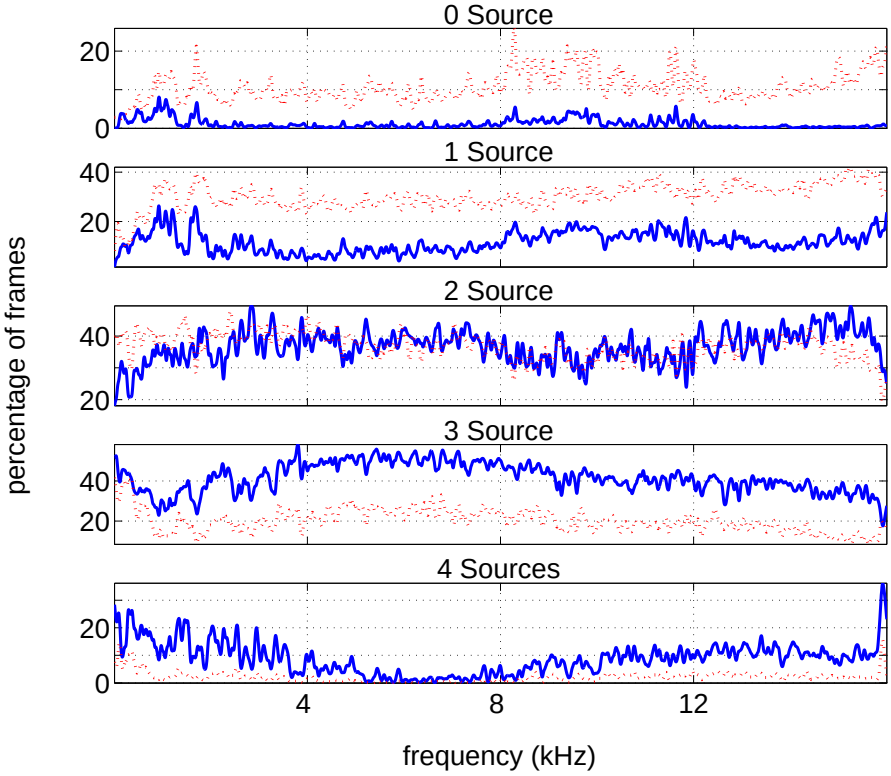


Figure 6.2 – Comparaison des propriétés de concentration énergétique des transformées de Fourier court-terme (ligne continue et bleue) et MDCT (ligne en pointillés et rouge).

plan TF (bin TF (f,t) ou molécule TF (p,q)), et que, dans la portion TF considérée, seules I_{ft} sources parmi les I sources originales sont considérées comme prédominantes. Ces I_{ft} sources sont les sources participant le plus activement au mélange au bin TF (f,t). Le mélange qui initialement s'écrivait :

$$\mathbf{X}(f,t) = \mathbf{A}\mathbf{S}(f,t) \tag{6.3}$$

peut être approximé par

$$\mathbf{X}(f,t) \approx \mathbf{A}_{\mathcal{I}_{ft}}\mathbf{S}_{\mathcal{I}_{ft}}(f,t) \tag{6.4}$$

où $\mathcal{I}_{ft}$ représente le I_{ft}-uplet des sources actives au bin temps-fréquence (f,t), $\mathbf{A}_{\mathcal{I}_{ft}}$ désigne la $J \times I_{ft}$ matrice composée des vecteurs colonnes $\mathbf{A}_i$ de $\mathbf{A}$ d'index $i \in \mathcal{I}_{ft}$, et $\mathbf{S}_{\mathcal{I}_{ft}}$ représente le vecteur composé des I_{ft} sources actives au bin (f,t). Si $\overline{\mathcal{I}_{ft}}$ représente le $(I-I_{ft})$-uplet des sources non actives (ou les moins actives) au bin (f,t), les signaux sources sont estimés en (f,t) par :

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) &= \mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}\mathbf{X}(f,t) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}}_{ft}}(f,t) &= 0 \end{cases} \tag{6.5}$$

où $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}$ représente la matrice (pseudo-)inverse de $\mathbf{A}_{\mathcal{I}_{ft}}$. (Dans le cas où $I_{ft} = J$, $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}$ est en fait l'inverse exacte de $\mathbf{A}_{\mathcal{I}_{ft}}$ en supposant que la matrice de mélange est de rang

plein). Précisons que les notations introduites conservent un caractère général valable pour $J \neq 2$ (même si dans notre application cible stéréo on a $J = 2$).

C'est au bloc 4 de la Figure 6.1 qu'a lieu l'étape clé du codeur, à savoir la détermination du nombre et de l'"identité" des sources participant le plus activement au mélange, encodés au bloc 5 comme une combinaison d'index (par exemple $\{1, 3\}$ si les sources s_1 et s_3 sont sélectionnées comme les deux sources actives). Contrairement aux descripteurs utilisés dans le chapitre 5 pour le codage des signaux sources, qui constituent une information riche en terme de contenu, l'information sur les signaux sources utilisée dans la configuration SSI-I est donc beaucoup plus simple et elle-même parcimonieuse. Elle n'en reste pas moins fondamentale pour permettre la séparation.

Il s'agit maintenant de définir un critère et un processus d'optimisation associé permettant de déterminer la combinaison de sources optimale pour chaque région temps-fréquence. En fait, ce problème est similaire au problème de l'estimateur dit *oracle* proposé par [Vincent et al., 2007] dans le cadre général de l'évaluation des méthodes de séparation de sources, notamment dans la configuration sous-déterminée avec exploitation de la parcimonie. Les auteurs de [Vincent et al., 2007] établissent un ensemble de classes d'estimateurs oracle qui offrent une référence de manière à quantifier la qualité des algorithmes de séparation de sources. Pour un type de mélange et d'algorithme de séparation donnés, les estimateurs oracles introduits permettent de déterminer les meilleures performances de séparation atteignables, calculées à partir des signaux sources supposés connus. Le but recherché par les auteurs est d'établir une limite supérieure des performances de séparation atteignables, de comparer les performances de différentes approches de séparation en regard de cette limite supérieure, et éventuellement de quantifier la difficulté de séparabilité des sources.

Si les signaux sources sont connus, une mesure des performances de séparation peut être faite en mesurant la distorsion $d(\hat{\mathbf{s}}, \mathbf{s})$ entre un signal source original $\mathbf{s}$ et sa version estimée $\hat{\mathbf{s}}$. Les auteurs de [Vincent et al., 2007] choisissent la distance Euclidienne comme mesure de distorsion : $d(\hat{\mathbf{s}}, \mathbf{s}) = \| \hat{\mathbf{s}} - \mathbf{s} \|^2$, où $\| \mathbf{s} \|^2 = \sum_{i=1}^{I} \sum_{t=0}^{T} s_i^2(t)$ pour un signal $\mathbf{s}$ à I voies et T échantillons sur chaque voie. Pour une catégorie de méthodes donnée, calculer l'estimateur oracle consiste à exploiter toute l'information disponible pour réaliser la séparation optimale, *i.e.* celle qui minimise $d(\hat{\mathbf{s}}, \mathbf{s})$.

Dans notre cas, pour chaque bin temps-fréquence (f, t), l'estimateur oracle consiste à déterminer la combinaison optimale de I_{ft} sources parmi les I sources originales. Cette combinaison optimale est notée $\tilde{\mathcal{I}}_{ft}$. $\tilde{\mathcal{I}}_{ft}$ vérifie :

$$\tilde{\mathcal{I}}_{ft} = \arg \min_{\mathcal{I}_{ft} \in \mathcal{P}} \sum_{i=1}^{I} \left(\hat{S}_i(f,t) - S_i(f,t) \right)^2 \tag{6.6}$$

où $\mathcal{P}$ dénote l'ensemble des $\mathcal{C}_{I_{ft}}^{I}$ combinaisons $\mathcal{I}_{ft}$ possibles. Trouver, à chaque bin temps-fréquence, la combinaison de I_{ft} sources qui minimise la distorsion est un problème de combinatoire qui peut être résolu en effectuant une recherche exhaustive parmi l'ensemble des $\mathcal{C}_{I_{ft}}^{I}$ combinaisons possibles $\mathcal{P}$, si le nombre I de sources reste

raisonnable.

6.2.2.2 Cas $I_{ft} = J$

Considérons d'abord le cas de la séparation par inversion exacte du mélange, *i.e.* lorsque le nombre de sources actives I_{ft} est égal à la dimension du mélange. La matrice inverse vaut alors $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}} = \mathbf{A}^{-1}_{\mathcal{I}_{ft}}$.

La sélection des sources prédominantes lorsque $I_{ft} = J$ est illustrée plus en détails à l'Annexe B.1 dans le cas de figure d'un mélange LISS de quatre signaux sources.

6.2.2.3 Cas $I_{ft} < J$

Si la matrice inverse $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}$ est mal conditionnée, il peut être préférable de considérer un nombre de sources localement actives inférieur à J. En effet, les sources considérées comme non actives ne sont pas exactement nulles et elles sont assimilables à un bruit qui lors de l'estimation peut être amplifié par une matrice $\mathbf{A}^{-1}_{\mathcal{I}_{ft}}$ mal conditionnée. Dans ce cas, il peut être préférable de réduire le mélange à un mélange sur-déterminé. Dans le cas $J = 2$, on a donc qu'une source localement active, et alors $\tilde{\mathcal{I}}_{ft} = \{k\}$. Le signal source $S_k(f,t)$ est alors obtenu par $\hat{S}_k(f,t) = \mathbf{A}_k^{\mathrm{T}}\ \mathbf{X}(f,t) / \parallel \mathbf{A}_k \parallel_2^2$. Une illustration de l'implémentation de la détermination de la source active est donnée à l'Annexe B.2 dans le cas d'un mélange LISS de quatre signaux sources.

6.2.2.4 Cas $I_{ft} > J$

Lorsque le principe de la SSI-I a été introduit à la Section 6.1, pour des raisons de simplicité, il a été supposé que $I_{ft} \leq J$. En fait, il peut être préférable de supposer plus de J sources simultanément actives. En effet, dans le cas de figure où K sources sont en réalité simultanément actives (d'énergie comparable dans le même bin temps-fréquence) avec K supérieur au nombre d'observations J du mélange, considérer que seulement $J < K$ sources sont actives implique que, 1) $K - J$ sources ne seront pas reconstruites correctement au décodeur, leur énergie étant mise à zéro alors qu'elles étaient originellement d'énergie comparable à celle des J sources reconstruites, et 2) les J sources actives pourront ne pas être correctement reconstruites du fait de l'interférence des $K - J$ autres sources actives (mais non considérées comme telles) dans le processus d'inversion. Il apparaît alors judicieux de chercher à estimer l'ensemble des $K > J$ sources effectivement actives. Ceci peut se faire par le calcul de la pseudo-inverse au sens de Moore-Penrose [Trefethen and Bau, 1997] de la $J \times K$ matrice de mélange $\mathbf{A}_{\mathcal{I}_{ft}}$. Le mélange est ici localement réduit à un mélange sous-déterminé dont la résolution revient à calculer une solution à l'équation (6.4) au sens des moindres carrées.

6.2.2.5 Sélection de la combinaison optimale

Au final, différentes configurations peuvent donc être considérées : on peut autoriser $I_{ft} < J$, $I_{ft} = J$ ou $I_{ft} > J$, ou n'importe quelle association de ces configurations. Quelle que soit la configuration choisie, $\tilde{I}_{ft}$ est donnée par la combinaison des sources minimisant la distorsion entre les sources estimées et les sources originales (équation (6.6)) : la meilleure estimation locale possible du vecteur source est assurée par le critère de sélection a posteriori des sources prédominantes.

Dans la pratique, des tests effectués sur des morceaux de musique occidentale grand public composés de 5 signaux sources ont montré qu'en moyenne, $I_{ft} = 2$ pour environ 60% des bins TF, $I_{ft} > 2$ pour 35% des bins, et $I_{ft} = 1$ pour moins de 5% des bins TF. Cependant, si l'on tient compte de l'énergie du signal sur chacun des bins TF, il apparaît que la très grande majorité de l'énergie d'un signal est concentrée sur les bins TF où $I_{ft} = 2$.

6.2.2.6 Traitement moléculaire

L'ensemble des équations des sous-sections précédentes est donné à l'échelle d'un atome temps-fréquence, mais est directement applicable à l'échelle moléculaire sous la contrainte que la molécule soit de dimension 1 en fréquence (un canal fréquentiel d'une molécule correspond alors à un canal fréquentiel d'un atome). Cette dernière contrainte sera vérifiée en pratique dans ce chapitre. Il est alors possible d'exprimer l'équation (6.6) pour un canal fréquentiel f donné et pour l'ensemble des T trames temporelles d'une molécule (dans toute cette sous-section, on suit les notations introduites au Chapitre 5, Section 5.1.2). Dans le cas présent, une molécule $M_{pq}^{\mathbf{X}}$ de taille $F \times T$ avec $F = 1$ est donnée par [2]

$$M_{pq}^{\mathbf{X}} = \{\mathbf{X}(f,t)\}_{\substack{f=(p-1)F \\ t\in Q=[(q-1)T,qT-1]}} \tag{6.7}$$

À l'échelle moléculaire, l'équation du mélange initial s'écrit :

$$M_{pq}^{\mathbf{X}} = \mathbf{A} M_{pq}^{\mathbf{S}} \tag{6.8}$$

et l'équation d'estimation (6.5) devient :

$$\begin{cases} M_{pq}^{\hat{\mathbf{S}}_{\mathcal{I}_{pq}}} &= \mathbf{A}_{\mathcal{I}_{pq}}^{\dagger} M_{pq}^{\mathbf{X}} \\ M_{pq}^{\hat{\mathbf{S}}_{\overline{\mathcal{I}}_{pq}}} &= 0 \end{cases} \tag{6.9}$$

la combinaison $\mathcal{I}_{pq}$ étant ici identique pour chacun des T atomes des molécules considérées. La combinaison de sources optimale est obtenue, de façon similaire à l'équation (6.6), par

2. Étant donné que $F = 1$, on pourrait directement noter une molécule $M_{fq}^{\mathbf{X}}$, mais il est choisi de conserver les notations propres aux molécules introduites au Chapitre 5, afin de ne pas introduire de nouvelles notations.

$$\tilde{\mathcal{I}}_{pq} = \arg\min_{\mathcal{I}_{pq}\in\mathcal{P}} \sum_{i=1}^{I} \left(M_{pq}^{\hat{\mathbf{S}}} - M_{pq}^{\mathbf{S}} \right)^2 \tag{6.10}$$

On obtient ainsi une combinaison optimale *en moyenne* sur l'ensemble d'une molécule (cette combinaison peut ne pas être optimale à l'échelle de chaque coefficient). C'est donc une approche sous-optimale par rapport au calcul de la combinaison de sources actives à l'échelle d'un coefficient MDCT. On testera les conséquences du moyennage de la combinaison de sources optimale sur les performances de séparation à la Section 6.3.

6.2.3 Procédé de séparation

L'étape d'estimation des signaux sources est réalisée au bloc 13 de la Figure 6.1. Nous revenons ici à des notations à l'échelle des atomes MDCT mais ces traitements peuvent être directement appliqués à l'échelle moléculaire comme vu dans la sous-section précédente. À chaque bin temps-fréquence (f,t), les signaux sources sont estimés, en accord avec l'équation (6.5) par

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) &= \mathbf{A}^{\dagger}_{\mathcal{I}_{ft}} \tilde{\mathbf{X}}^{W}(f,t) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}_{ft}}}(f,t) &= 0 \end{cases} \tag{6.11}$$

où $\tilde{\mathbf{X}}^{W}$ représente la décomposition temps-fréquence du signal de mélange tatoué et converti au format PCM 16-bits. On note donc que toute dégradation du signal de mélange par le tatouage ou la conversion 16-bits a une incidence potentielle directe sur l'estimée du vecteur sources. Ce point est discuté plus amplement à la section suivante, et est étudié expérimentalement à la Section 6.3. Notons d'ores et déjà que l'influence de la conversion PCM 16 bits est négligeable devant les effets de l'insertion du tatouage. Ceci est établi de deux manières. Tout d'abord d'un point de vue théorique : le tatouage est implémenté de manière à être robuste à la conversion PCM 16 bits, ce qui implique que les dégradations sur le signal de mélange causées par l'insertion du tatouage sont nettement plus importantes que celles produites par la conversion PCM 16 bits. Enfin, une validation expérimentale de cette hypothèse a également été apportée : les performances de séparation obtenues à partir d'un mélange non-quantifié sont quasi identiques à celles obtenues avec un signal quantifié sur 16 bits (Les écarts de SNR sont de l'ordre de 10^{-2}dB).

Dans le cas où seulement une source est supposée active, $\mathcal{I}_{ft}$ est réduite à un singleton $\{k\}$, la pseudo inverse $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}$ vaut $\mathbf{A}_k^{\mathrm{T}} / \parallel \mathbf{A}_k \parallel_2^2$ et la source $S_k(f,t)$ est estimée par

$$\hat{S}_k(f,t) = \frac{\mathbf{A}_k^{\mathrm{T}}\, \tilde{\mathbf{X}}^{W}(f,t)}{\parallel \mathbf{A}_k \parallel_2^2} \tag{6.12}$$

les autres sources étant mises à zéro.

6.2.4 Codage et allocation de l'index des sources prédominantes

6.2.4.1 Codage de l'index

Le processus de séparation par inversion locale du mélange choisi en SSI-I impacte directement la nature de l'information à tatouer sur le signal de mélange. Comme nous l'avons déjà mentionné, cette dernière est beaucoup plus parcimonieuse en SSI-I qu'en SSI par codage des signaux sources (SSI-C). Or, le procédé de watermarking utilisé en SSI-I pour transmettre cette information (bloc 6 de la Figure 6.1) est similaire au tatouage utilisant un MPA présenté au Chapitre 5, Section 5.3. Cependant la contrainte sur la capacité nécessaire pour insérer l'information utile à la séparation est moins exigeante que celle de la séparation par codage des signaux sources. Il en résulte un dimensionnement du tatouage et des contraintes de groupement moléculaire différents de ceux choisis dans le cas de la SSI-C.

En effet, dans le cadre de la SSI-I, l'information tatouée doit, comme nous l'avons vu, simplement permettre d'identifier, parmi les I sources, quelles sont celles qui participent le plus activement au mélange. En fonction du nombre maximal de sources supposées simultanément actives, le nombre de combinaisons $\text{card}(\mathcal{P})$ prend différentes valeurs : si une seule source est supposée active à chaque bin temps-fréquence, $\text{card}(\mathcal{P}) = I$; si exactement deux sources sont supposées simultanément actives, $\text{card}(\mathcal{P}) = \frac{I(I-1)}{2}$; si au plus deux sources sont supposées simultanément actives, $\text{card}(\mathcal{P}) = \frac{I(I+1)}{2}$; et enfin si le nombre de sources actives est libre, et potentiellement supérieur au nombre d'observations, $\text{card}(\mathcal{P}) = 2^I$. La combinaison optimale peut être codée avec un code de taille $\lceil \log_2(I_{ft}) \rceil$. Dans le cas le plus complexe, cette taille de code vaut donc I bits. Dans le cas d'un nombre de sources actives $I_{ft} \leq 2$ (nous sommes toujours dans le cas d'un mélange stéréophonique), le nombre de combinaisons possibles de sources est de 10 pour un mélange à 4 sources, et de 15 pour un mélange à 5 sources ; 4 bits sont donc suffisants (même si cette taille de code n'est pas optimale) pour encoder l'information I_{ft}. Comme nous le verrons par la suite, les effets du watermarking sur les performances de séparation sont, pour des tailles de watermark de l'ordre de 4 bits par coefficient MDCT, très limités. Pour cette raison, un message de 4 bits est utilisé pour encoder $I_{ft} \leq 2$, que le mélange soit composé de 4 ou 5 sources. De plus, la dimension stéréo du mélange permet de tatouer simultanément les deux voies du signal, divisant ainsi par deux la ressource nécessaire par canal. De fait, si le processus de séparation est mené à l'échelle de chaque coefficient MDCT, la capacité moyenne d'insertion de la watermark est de 2 bits par coefficient. En pratique, le processus de séparation est limité à la bande fréquentielle [0-16kHz]. Le nombre de coefficients MDCT étant identique au nombre d'échantillons temporels d'un signal, le débit de l'information à insérer est de $2 \times F_s \times 16,000/(F_s/2) = 64$kbits/s/voie, ($F_s$ est la fréquence d'échantillonnage). Cela représente environ $1/4$ de la capacité maximale du système de tatouage [Pinel et al., 2010b]. Les paramètres de calcul du MPA du système de tatouage sont alors réglés de manière à ce que seule la capacité totale nécessaire soit disponible pour insérer la watermark.

6.2.4.2 Effets du watermarking sur les performances de séparation

Comme nous l'avons vu à la Section 6.2.3, le processus de séparation en SSI par inversion du mélange est appliqué sur le signal de mélange tatoué. Toute modification du signal de mélange par le tatouage est donc susceptible d'impacter directement les performances de séparation. Les blocs 3, 3' et 4 de la Figure 6.1 sont en effet étroitement liés : plus la taille de la molécule est grande, moins le signal de mélange est altéré par l'insertion du tatouage (le message $\mathcal{I}_{ft}$ devient alors $\mathcal{I}_{pq}$ et est réparti sur $F \times T$ atomes), mais plus la superposition des sources risque d'interférer avec la séparation. D'un autre côté, plus la dimension de la molécule est réduite, plus les sources sont disjointes, mais au détriment possible de la dégradation du signal de mélange par le tatouage. Dans le but de mesurer les effets du tatouage sur les performances d'estimation des sources par inversion, plusieurs configurations sont testées. La configuration décrite dans la sous-section précédente, dans laquelle tout le processus de SSI-I est mené à l'échelle d'un coefficient TF (f, t), et où un tatouage moyen de 2 bits/coefficient MDCT est inséré, est appelée SSI-I basique. Deux autres configurations sont testées :

- Une configuration nommée light-watermark SSI-I dans laquelle le taux d'insertion de tatouage est volontairement réduit à la moitié du taux de la configuration SSI-I basique, *i.e.* 32kbits/s/voie, dans le but de limiter l'influence du tatouage sur le processus de séparation. Pour cela, la courbe de masquage obtenue par le MPA du système de tatouage est abaissée, et le processus d'estimation des sources est réalisé à l'échelle d'une molécule 1×2 de coefficients MDCT (le groupement moléculaire est effectué au bloc 3 de la Figure 6.1) : Une seule valeur moyenne de $\mathcal{I}_{ft}$ est utilisée pour les bins TF (f, t) et $(f, t+1)$ de chaque molécule. Dans cette configuration, les coefficients MDCT au décodeur sont plus proches des coefficients MDCT du mélange initial, mais ceci au détriment de la résolution de séparation.
- Une seconde configuration nommée full-watermark SSI-I, dans laquelle le débit de tatouage est volontairement augmenté (en positionnant la courbe de masquage à son niveau maximal). Ce débit est alors significativement plus élevé que celui effectivement nécessaire à la transmission du paramètre $\mathcal{I}_{ft}$, qui est de nouveau transmis, dans cette configuration, à l'échelle d'un bin TF. Le but de cette configuration est de tester si le processus de séparation par inversion du mélange est robuste à l'insertion d'un tatouage haute capacité. Un tel tatouage pourrait être utilisé pour insérer une plus grande quantité d'information pour guider la séparation (par exemple dans l'optique d'un système hybride SSI-C/SSI-I tel que celui présenté au Chapitre 7). Dans la pratique, un message aléatoire est ajouté au message de $\mathcal{I}_{ft}$, de sorte que le débit de tatouage atteigne le niveau maximum permis par le MPA sous contrainte d'inaudibilité. La différence entre la configuration SSI-I basique et la configuration full-watermark SSI-I est illustrée par la Figure 6.3 qui donne un exemple de la capacité d'insertion par bin TF dans chacune des deux configurations.

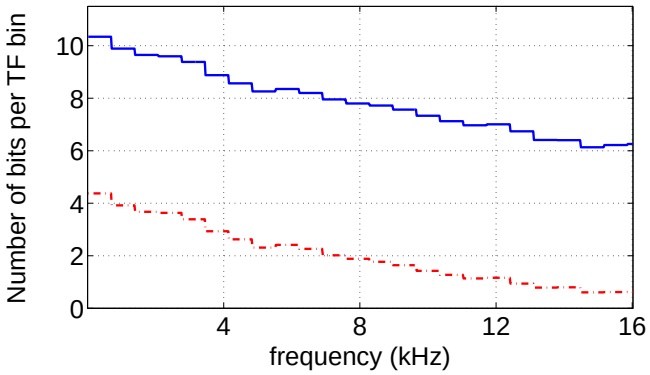


FIGURE 6.3 – Exemple de capacité d'insertion par coefficient TF sur une trame MDCT donnée. Ligne continue : le MPA est réglé de sorte à fournir la capacité d'insertion maximale sous contrainte d'inaudibilité ; Ligne pointillés : le MPA est réglé de sorte à fournir une capacité de 2 bits par coefficient en moyenne correspondant au codage de $\mathcal{I}_{ft}$ pour $I_{ft} < 2$ avec I=4 ou 5.

Les caractéristiques des trois configurations mentionnées ci-dessus sont résumées à la Table 6.2. Pour chaque configuration, un code de 4 bits est utilisé pour représenter $\mathcal{I}_{ft}$, hormis pour la configuration $I_{ft} < I$ avec I=5 où un code de 5 bits est nécessaire, ce qui correspond à 2.5bits/coefficient MDCT, soit un débit de tatouage de 80kbits/s/voie. Le compromis entre échelle de traitement et performances de séparation reste un point crucial de la technique de SSI-I qui sera étudié expérimentalement en détails au cours des sections suivantes. Toutefois, les contraintes échelle-performances sont différentes de celles rencontrées en SSI-C où l'échelle de traitement conditionnait l'utilisation de descripteurs de sources spécifiques.

6.2.4.3 Allocation de l'information à tatouer

Au codeur, un réglage adéquat des paramètres lors du calcul de la capacité (décalage de la courbe de masque par rapport à la courbe d'inaudibilité dans le cas de l'utilisation d'un MPA) permet d'obtenir une capacité d'insertion globale du signal de mélange qui corresponde au mieux au débit, fixe [3], de l'information de tatouage. L'ensemble des codes locaux correspondant aux index des sources prédominantes à chaque atome/molécule sont ensuite concaténés, formant un message global de taille inférieure ou égale à la capacité disponible, puis ce message est découpé et inséré sur le signal de mélange en fonction de la capacité de chaque bin TF [4]. Au décodeur, un redécoupage du message décodé permet de retrouver l'index *local* des sources prédominantes nécessaire à l'estimation des signaux sources.

3. Le nombre de bits par atome/molécule est fixe et connu en fonction du nombre de sources I dans le mélange et du nombre de sources simultanément actives autorisé I_{ft} voir Section 6.2.4.1.

4. L'ensemble du processus (calcul de la capacité, concaténation des codes, insertion du tatouage) est réalisé soit au niveau de chaque trame MDCT, soit sur une portion plus large de signal (plusieurs trames MDCT). La première solution est privilégiée dans une perspective d'implémentation temps réel

En SSI-I, la nature même de l'information de tatouage propre à cette technique de séparation a une conséquence essentielle sur la technique de tatouage : la taille du message utile à l'estimation des signaux sources guide ici le calcul de la capacité d'insertion et non l'inverse comme c'était le cas en SSI-C. L'index des sources prédominantes constituant l'information à tatouer en SSI-I est en effet de taille identique en tout point du plan TF. Or, quel que soit le réglage du MPA utilisé pour estimer la capacité disponible à l'insertion du tatouage, celui-ci fournit une capacité plus importante dans les zones de forte énergie du signal (typiquement en basses fréquences) comme le montre la Figure 6.3. Une répartition ponctuelle de la ressource, *i.e.* sur une molécule (p, q) du signal de mélange étaient insérées les informations des molécules (p, q) des signaux sources à coder, comme celle choisie en SSI-C n'est donc pas adaptée à un message de taille constante. C'est pourquoi, il est choisi de ne plus répartir l'information de tatouage ponctuellement entre les coefficients MDCT. Les messages correspondants à la combinaison optimale $\tilde{\mathcal{I}_{ft}}$ pour chaque molécule, ou chaque bin TF, sont répartis sur le signal là où la capacité le permet. La contrainte sur la capacité du signal hôte n'est donc plus locale, mais devient globale : le message global correspondant à l'ensemble des messages des combinaisons optimales doit pouvoir être inséré, mais il n'est pas nécessaire que chaque combinaison (*i.e.* chaque message local) soit tatouée sur la molécule (ou l'atome) à laquelle elle est appliquée pour la séparation.

6.3 Expérimentations

6.3.1 Données

Les résultats de séparation par SSI-I présentés dans cette section sont obtenus sur des signaux de musique échantillonnés à 44100Hz avec des mélanges voix chantée+instruments. Les signaux de mélanges stéréo sont obtenus par mélange linéaire instantané de 4 à 5 signaux sources jouant en harmonie (il s'agit de morceaux grand public de genre *jazz*, *pop-rock*, *rock*, *funk* et *new wave*). Les résultats sont moyennés sur environ 50 secondes de musique (5 extraits de morceaux de différents styles de 10s chacun). Les signaux sources sont, selon le morceau considéré : s1 = une guitare ou un piano, s2 = une batterie (une piste pour l'ensemble de la batterie), s3 = une voie chantée (homme ou femme), s4 = une guitare basse, s5 = une section de cuivres, des choeurs, ou un synthétiseur. Différentes matrices de mélange LIS ont été utilisées de manière à fournir un résultat de mélange le plus proche possible du mélange commercial initial (répartition de la puissance des instruments sur les deux voies du mélange). Voici un exemple de matrice pour un mélange de 5 signaux sources :

$$\mathbf{A} = \begin{bmatrix} 0.95 & 0.82 & 0.71 & 0.58 & 0.32 \\ 0.32 & 0.58 & 0.71 & 0.82 & 0.95 \end{bmatrix} \tag{6.13}$$

qui correspond au vecteur d'azimuts (en degrés) $\boldsymbol{\theta} = [-30, -10, 0, 10, 30]$. Pour un mélange de 4 sources, la matrice de mélange $\mathbf{A}$ est la sous-matrice formée des 4 premières

colonnes (normalisées) de la matrice de l'équation (6.13), et seules les sources s_1 à s_4 sont utilisées.

6.3.2 Superposition des sources

Nous étudions dans cette section la superposition des différentes sources en fonction de l'énergie de ces signaux. Il s'agit d'une extension des tests effectués à la Section 6.2.1 où seule la superposition des signaux sources était mesurée : à chaque bin TF était compté le nombre de sources d'énergie significative, mais aucune information n'était fournie sur la portion de l'énergie du signal où ce signal était la $k^{\text{ième}}$ source la plus énergétique du mélange, $k \in [1, \ldots, I]$. Il est possible d'affiner cette mesure de la superposition des sources en calculant, pour chaque atome (f, t) (ou molécule (p, q)) du plan temps-fréquence, le ratio énergétique entre cet atome (ou molécule) d'une source donnée et les atomes (ou molécules) correspondants des autres sources. Ce ratio énergétique est défini à l'échelle d'un atome TF par

$$R_i(f,t) = \frac{|S_i(f,t)|^2}{\sum_{j \neq i} |S_j(f,t)|^2} \tag{6.14}$$

et à l'échelle d'une molécule, par

$$R_i(p,q) = \frac{\sum\limits_{(f,t)\in\{P\times Q\}} |S_i(f,t)|^2}{\sum_{j\neq i} \sum\limits_{(f,t)\in\{P\times Q\}} |S_j(f,t)|^2} \tag{6.15}$$

avec $P \times Q = [(p-1)f, pf-1] \times [(q-1)t, qt-1]$. Ainsi, pour chaque atome (ou molécule) du plan temps-fréquence, il est possible de connaître le classement énergétique des I sources, puis de calculer pour chaque source, à partir de ce résultat, quel pourcentage de l'énergie de cette source correspond au cas où elle est classée k-ième source la plus énergétique, avec $k \in [1, \ldots, I]$.

La Table 6.1 a été établie à l'échelle de l'atome TF, à partir des décompositions MDCT de signaux sources de musique de différents styles pour des mélanges à cinq sources. La distribution énergétique des signaux sources présentée Table 6.1 laisse apparaître que la plus grande partie de l'énergie des différentes sources est localisée sur des molécules où cette source est parmi les deux sources les plus énergétiques. De 82.0% (pour s_5 à la Table 6.1e) à 99.7% (pour s_3 à la Table 6.1b) de l'énergie totale de chaque signal source est concentrée sur les molécules où cette source est une des deux plus énergétiques des cinq sources composant le mélange. De plus, le pourcentage d'énergie où une source arrive en seconde position est particulièrement significatif. Ainsi, si seule la source la plus active est considérée, jusqu'à 66% de l'énergie totale d'un signal peut être négligée dans le cas d'une source à large spectre (typiquement s_5 qui représente le synthétiseur dans le mélange de la Table 6.1b). Il apparaît donc que considérer plus d'une source active est nettement préférable à l'hypothèse d'une seule source active utilisée en séparation de sources par masquage binaire [Yilmaz and Rickard, 2004]. À noter, pour la plupart des sources, un pourcentage énergétique très différent selon

rang	s1	s2	s3	s4	s5
1	76.3	82.1	79.0	87.0	85.5
2	16.3	13.4	17.6	10.3	12.6
3	5.5	3.3	2.7	2.2	1.6
4	1.6	1.0	0.7	0.4	0.2
5	0.2	0.2	0.1	6.10^{-2}	3.10^{-2}

(a) Jazz

rang	s1	s2	s3	s4	s5
1	82.5	96.7	97.8	84.8	33.6
2	15.2	2.9	1.9	12.4	52.2
3	2.0	0.3	0.3	2.4	13.3
4	0.3	4.10^{-2}	6.10^{-2}	0.4	0.9
5	1.10^{-2}	4.10^{-3}	1.10^{-2}	3.10^{-2}	2.10^{-2}

(b) New wave

rang	s1	s2	s3	s4	s5
1	47.4	93.0	95.5	87.9	93.7
2	37.4	6.2	3.7	10.9	5.6
3	12.4	0.6	0.6	0.9	0.5
4	2.4	0.1	0.2	0.2	8.10^{-2}
5	0.3	2.10^{-2}	3.10^{-2}	1.10^{-2}	1.10^{-2}

(c) Rock

rang	s1	s2	s3	s4	s5
1	93.9	93.6	93.6	91.8	86.7
2	5.2	5.6	5.3	6.7	11.7
3	0.8	0.6	1.0	1.2	1.4
4	0.1	0.1	0.2	0.2	0.2
5	2.10^{-2}	2.10^{-2}	1.10^{-2}	2.10^{-2}	1.10^{-2}

(d) Funk

rang	s1	s2	s3	s4	s5
1	81.0	71.6	93.3	88.7	46.5
2	16.9	20.7	5.3	9.9	35.5
3	1.8	5.7	1.1	1.1	15.0
4	0.3	1.7	0.2	0.2	2.7
5	2.10^{-2}	0.3	5.10^{-2}	4.10^{-2}	0.2

(e) Pop-rock

Table 6.1 – Pourcentage de l'énergie totale des signaux sources en fonction de leur rang énergétique dans le mélange. Étude à l'échelle de l'atome temps-fréquence.

qu'une source appartient aux deux sources les plus énergétiques, ou aux trois dernières sources. Cette dernière configuration ne représente en général que quelques pourcents

de l'énergie totale du signal [5]. Dans un première implémentation, les trois sources les moins énergétiques pourront alors être considérées comme un *bruit* de faible puissance par rapport aux deux sources prépondérantes (la configuration est alors déterminée et l'estimation des sources se fait par inversion exacte du mélange). Si la matrice inverse introduite à l'équation (6.5) n'est pas mal dimensionnée, la distribution énergétique des signaux sources dans le mélange laisse à penser que les sources séparées seront relativement fidèles aux sources originales (inversion exacte de la matrice de mélange).

6.3.3 Qualité des signaux de mélange au décodeur

Avant de donner les performances de séparation à proprement parler, nous confirmons dans cette sous-section que le tatouage du signal de mélange (bloc 6 de la Figure 6.1) n'a pas d'influence audible sur la qualité du mélange. L'influence du tatouage a été étudiée par de nombreux tests informels et des mesures d'*Objective Difference Grade* (ODG [6]) reportés dans [Pinel et al., 2010b] [Pinel et al., 2010]. En réalité, le tatouage est inaudible dans la configuration full-watermark SSI-I définie dans la Section 6.2.4.2, dans laquelle les réglages du MPA sont les plus contraignants. De fait, l'inaudibilité du tatouage dans les configurations SSI-I basique et plus encore light-watermark SSI-I est parfaitement assurée, car dans ces deux configurations, la courbe de masquage est significativement abaissée par rapport à la configuration full-watermark SSI-I (voir Figure 6.3).

Des mesures objectives complémentaires de type SNR confirment également les faibles dégradations subies par le signal de mélange à la suite de l'insertion du tatouage. La Figure 6.4 donne un aperçu de la dégradation moyenne du signal de mélange en terme de SNR dans le domaine des MDCT après l'insertion d'un tatouage à un taux de 64kbits/s correspondant à la configuration SSI-I basique. On note que, dans la zone fréquentielle d'intérêt, le SNR ne décroît pas en dessous de 45dB, ce qui illustre la qualité audio du signal de mélange à la sortie du codeur.

6.3.4 Résultats de séparation

Les trois configurations présentées à la Section 6.2.4.2 et résumées à la Table 6.2 ont été testées afin d'évaluer les performances de séparation de la méthode SSI-I, ainsi que les effets de la résolution à laquelle est réalisé l'ensemble du processus (bin TF

5. Cependant, pour une source peu parcimonieuse et de faible énergie par rapport aux autres sources(typiquement les choeurs), le pourcentage énergétique où cette source appartient aux trois sources les moins énergétiques peut représenter jusqu'à environ 18% de son énergie totale, mais ce cas de figure ne semble pas général.

6. Métrique de mesure de la qualité d'un signal audio selon des critères perceptuels de l'oreille humaine. La note globale d'ODG permet de comparer la qualité d'un signal par rapport à un signal de référence. Elle est obtenue par l'algorithme standardisé PEAQ (Perceptual Evaluation of Audio Quality) développé entre 1994 et 1998 par un groupe d'experts de l'Union Internationale des Télécommunications [Thiede et al., 2000].

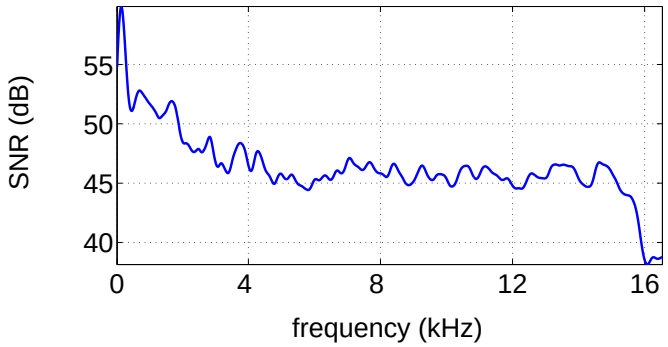


Figure 6.4 – Rapport Signal-sur-Bruit entre la décomposition MDCT du signal de mélange original et sa différence avec la décomposition MDCT du signal de mélange tatoué, dans la configuration SSI-I basique (débit de tatouage = 64kbits/s). Résultats moyennés sur 30s de signaux de mélange à 5 sources (différents styles musicaux).

ou molécule) et du débit de tatouage [7]. À ces trois configurations, nous avons décidé d'ajouter trois autres configurations dites "de référence", n'intégrant pas la totalité du processus de séparation par SSI-I présenté Figure 6.1, et qui servent donc de références aux mesures de performances.

La configuration Oracle_O est la configuration correspondant à l'estimateur Oracle idéal, comme introduit en [Vincent et al., 2007] : la combinaison optimale $\tilde{\mathcal{I}}_{ft}$ est utilisée pour effectuer l'estimation des sources, et le processus d'inversion est réalisé selon l'équation (6.5) et non l'équation (6.11), *i.e.* il n'y a pas de tatouage du signal de mélange. L'ensemble du traitement est réalisé à l'échelle d'un bin TF, sans phase de groupement moléculaire, et l'index des sources prédominantes à chaque bin (f, t) est supposé connu au décodeur. Cette configuration correspond à la séparation de sources optimale qu'il est possible d'obtenir avec le processus d'inversion locale du mélange. C'est à partir de cette configuration que seront évaluées les performances des autres algorithmes de séparation.

L'algorithme Oracle_M correspond à la même configuration que l'algorithme Oracle_O avec en plus une étape de groupement moléculaire des coefficients MDCT. La distinction entre cette configuration et la précédente vient du calcul de la combinaison de sources optimale, qui n'est plus calculée à l'échelle d'un atome (f, t), mais à l'échelle d'une molécule (p, q) de taille 1×2. Il y a dans cette configuration une seule valeur de $\tilde{\mathcal{I}}_{ft}$ pour estimer les deux coefficients consécutifs (f, t) et $(f, t + 1)$, tout comme dans la configuration light-watermark SSI-I. Il s'agit donc d'une sorte de moyennage de la combinaison de sources optimale sur l'ensemble des coefficients d'une molécule, et donc d'une diminution de la résolution d'un facteur $F \times T = 2$ par rapport à la configuration Oracle_O lors de la phase d'inversion du mélange. La séparation est donc effectuée à l'échelle moléculaire selon la formule (6.9). Il s'agit de mesurer l'influence du groupement moléculaire sur les performances de séparation par rapport aux perfor-

7. Ces caractéristiques sont fixés respectivement au bloc 3 de la Figure 6.1 pour l'échelle de traitement (par atome ou par molécule), et au bloc 6 pour le calcul de capacité et le tatouage.

mances optimales (comparaison Oracle_M/Oracle_O), et également de mesurer les effets de la watermark à ce débit d'insertion (comparaison Oracle_M/light-watermark SSI-I).

Enfin, nous avons également implémenté la technique de séparation semi-aveugle de sources en configuration sous-déterminée introduite dans [Bofill and Zibulevski, 2001], de manière à pouvoir juger de l'apport de la configuration *informée* sur les performances de séparation, en regard des performances obtenues avec une méthode similaire de séparation de sources par inversion, mais non informée. Rappelons que dans cette configuration (cf Section 2.5.2.2), par la suite dénommée BZ, les deux signaux prédominants (parmi les 4 ou 5 constituant le mélange $\mathbf{x}$), sont sélectionnés dans le plan TF par sélection de la combinaison linéaire des vecteurs de mélange (les colonnes de la matrice de mélange) qui conduit au chemin le plus court de l'origine à la donnée $\mathbf{x}$. Rappelons également (cf Figure 2.10) que la faiblesse de cette méthode géométrique est qu'elle ne peut pas considérer l'ensemble des combinaisons de sources possibles. L'insertion du tatouage dans la présente méthode de SSI-I résout ce problème, ainsi que celui de l'estimation de la matrice de mélange. Alors que l'estimation de la matrice de mélange constitue une difficulté supplémentaire pour les techniques de séparation aveugle de sources, la matrice de mélange $\mathbf{A}$ est ici transmise au décodeur. Dans le cas d'un mélange LISS, la matrice $\mathbf{A}$ est constituée de $2I$ coefficients, voire même, si chaque vecteur colonne normalisé est obtenu à partir de l'azimut chaque source, de I coefficients. Leur transmission au décodeur (une seule fois pour l'ensemble du morceau) constitue donc un coût négligeable par rapport à la transmission du tatouage de l'index des sources prédominantes.

Algorithme	Échelle	I_{ft}	Capacité d'insertion (kb/s)
Oracle_O	1 coeff. TF	≤ 2	-
Oracle_M	molécule	≤ 2	-
BZ	1 coeff. TF	2	-
SSI-I basique	1 coeff. TF	≤ 2	64
light-watermark SSI-I	molécule	≤ 2	32
full-watermark SSI-I	1 coeff. TF	≤ 2	250
free SSI-I	1 coeff. TF	$\leq I$	64 ($I = 4$) - 80 ($I = 5$)

TABLE 6.2 – Spécificités des différents algorithmes testés.

6.3.4.1 Performances de séparation

La qualité des sources séparées est évaluée de manière similaire à celle présentée à la Section 5.2.2, à la fois par des tests d'écoute informels au moyen d'un casque audio

de haute qualité, et par les mesures de performances introduites dans [Vincent et al., 2005]. De manière à mesurer la difficulté à séparer chaque signal source du mélange à laquelle elle appartient, le SNR_{in} de chacune des sources doit être pris en compte. Ainsi, pour les mélanges à 4 sources, le SNR_{in} moyen des sources s_1 à s_4 vaut respectivement -8.4, -7.1, -4.1, et -2.4dB. Pour les mélanges à 5 sources, le SNR_{in} moyen des sources s_1 à s_5 vaut respectivement -9.4, -8.3, -5.3, -3.7 et -7.8dB. La différence entre le SDR en sortie du système de SSI-I et le SNR_{in} fournit la véritable mesure de performance de séparation. Rappelons que les signaux sources s_1 à s_4 sont identiques dans les mélanges à 4 ou 5 sources, la matrice de mélange $\mathbf{A}$ pour un mélange 4 sources étant constituée des 4 première colonnes de la matrice $\mathbf{A}$ du mélange 5 sources correspondant.

Les résultats de séparation sont présentés à la Figure 6.5 pour les mélanges de 4 sources et à la Figure 6.6 pour les mélanges de 5 sources. Considérons tout d'abord les résultats de la configuration SSI-I basique. La première observation qui peut être faite, pour les mélanges à 4 sources comme pour les mélanges à 5 sources est que de très bonnes performances de séparation des sources sont obtenues avec des SDR-SNR_{in} supérieurs ou égaux à 20dB dans le cas d'un mélange 4 sources, et supérieurs ou égaux à 17dB dans le cas d'un mélange 5 sources. Bien que les performances de SDR et SIR indiquent que les sources estimées sont relativement différentes des sources initiales, leurs valeurs confirment l'efficacité de la méthode SSI-I de séparation par inversion en termes de reconstruction individuelle des sources, en particulier dans la configuration sous-déterminée complexe envisagée. La méthode de sélection des sources prédominantes est également validée par ces scores élevés. L'écart de performances entre les différentes sources s'explique en particulier par les différences de SNR_{in} . Les SDR les plus élevés sont obtenus pour la guitare basse (s_4) et les plus faibles pour la batterie (s_2), cependant, les SNR_{in} de la guitare basse étant plus élevés que ceux de la batterie, les performances globales sont plus équilibrées en terme de SDR-SNR_{in} , comme il peut être vu sur les figures 6.5c et 6.6c. Des SIR de sortie compris entre 35 et 42.5dB pour un mélange de 4 sources, et entre 29.5 et 34dB pour un mélange de 5 sources démontrent une très bonne réjection des interférences sur les signaux sources estimés, validant ainsi l'hypothèse de deux sources prédominantes à chaque région du plan TF : l'inversion locale du mélange avec la matrice de mélange constituée des vecteurs colonnes des deux sources prédominantes fournit une très bonne estimation de chacune des sources. La majorité de l'énergie de chaque source étant concentrée aux atomes TF où cette source est parmi les deux sources prédominantes, la plupart de son énergie est reconstruite par le processus d'inversion locale du sous-mélange *déterminé* constitué de ces deux sources. Les performances de SDR et SAR sont également très satisfaisantes. En raison des valeurs élevées de SIR en sortie, les mesures de SAR et de SDR sont quasiment identiques pour chacune des configurations. C'est pourquoi nous ne fournissons ici que les mesures de SDR. La qualité globale de séparation des signaux estimés dépend principalement des artefacts, en particulier du bruit musical. Des scores de SDR variant de 12.5 à 18dB pour un mélange de 4 sources, et de 10 à 13.5dB pour un mélange de 5 sources sont obtenus. Les performances de séparation décroissent donc avec le nombre

de sources composant le mélange, une source supplémentaire accroissant le risque de superposition dans le plan TF. Une baisse de 6.5dB pour le SIR et 3.5dB pour le SDR sont enregistrées en passant de 4 à 5 sources, alors que le SNR_{in} baisse seulement de 1.2dB. La Table 6.1 fournit une explication à cette baisse. En effet, pour les mélanges *new-wave* et *pop-rock*, de 13.3 à 15% de l'énergie de la source s_5 est concentrée aux bins TF où cette sources est la troisième source la plus énergétique (pourcentage nettement plus élevé que pour les sources s_1 à s_4). Or, nous avons pu noter que le classement énergétique d'une source au sein d'un mélange est souvent corrélé à sa sélection, ou non, parmi les sources prédominantes à un bin TF donné. Cela signifie qu'environ 15% de l'énergie de la source s_5 est concentrée à des bins TF où s_5 n'est pas sélectionnée parmi les deux sources prédominantes, ce qui signifie qu'à ces bins TF, son estimée est mise à zéro (cf (6.5)). A cette conséquence qui impacte seulement la source s_5, s'ajoute le fait que sur les bins TF où s_5 est d'énergie non négligeable par rapport au deux sources prédominantes (mais pas d'énergie suffisante pour être sélectionnée parmi les deux sources prédominantes), cette énergie influe directement sur l'estimation des deux sources prédominantes.

Les valeurs des performances de SIR, SDR et SDR-SNR_{in} démontrent la possibilité de manipuler individuellement les signaux séparés. La qualité des sources séparées est confirmée par des tests d'écoute qui montrent une très bonne réjection des interférences ainsi qu'une très haute qualité globale des sources séparées : chaque instrument est clairement isolé, et les artefacts très limités. Des exemples audio pour les différentes configurations peuvent être téléchargés à l'adresse `http ://www.gipsa-lab.inpg.fr /~mathieu.parvaix/IISS-demo.rar`. Le package inclut les signaux de mélange originaux et tatoués, ainsi que les signaux sources et les signaux estimés pour chacune des configurations présentées à la Table 6.2. Tous les signaux sont normalisés de telle sorte qu'il soit possible, pour le lecteur intéressé, de procéder à un remixage à partir des signaux séparés et du signal de mélange tatoué.

Attachons-nous maintenant à l'influence du tatouage sur les performances de séparation. Il apparaît que les résultats de la configuration SSI-I basique sont quasi identiques à ceux obtenus avec la configuration Oracle_{O}, qui constitue, rappelons-le, la borne supérieure des performances de séparation atteignables par inversion locale du processus de mélange à l'échelle d'un atome TF. La configuration SSI-I basique permet donc d'obtenir quasiment les meilleures estimées des signaux sources par inversion locale du mélange, sans que les dégradations subies par le signal de mélange au codeur, en particulier par l'insertion du tatouage (et dans une (bien) moindre mesure par la conversion PCM) ne dégrade le résultat de l'inversion. Cela signifie que l'insertion du tatouage à un débit "basique" (ici 64kbits/s/voie) a une influence extrêmement limitée sur le processus de séparation. Ceci est en accord avec les scores élevés de SNR présentés Figure 6.4. Le processus d'inversion à l'équation (6.11) donne des résultats quasiment identiques à ceux obtenus par l'équation (6.5). Cette observation est confirmée par les résultats similaires obtenus pour les configurations Oracle_{M} et light-watermark SSI-I (le débit de tatouage vaut alors 32kbits/s/voie). Cependant, si à l'inverse la capacité

d'insertion du tatouage devient trop élevée, comme c'est le cas dans la configuration full-watermark SSI-I[8] (avec un débit de tatouage d'environ 250kbits/s/voie), le tatouage influe directement sur les performances de séparation. Le mélange est alors trop fortement dégradé par l'insertion du tatouage. On note ainsi une baisse moyenne de 5dB et 3dB du SDR entre les configurations SSI-I basique et full-watermark SSI-I, pour des mélanges respectivement de 4 et 5 sources. Un utilisateur "client" du système de SSI-I devra donc être prudent s'il utilise un tatouage à haut débit pour la transmission d'informations autres que l'index des sources prédominantes (par exemple une information de codage de certaines sources dans le cas d'une approche hybride comme celle introduite Chapitre 7).

L'influence du groupement moléculaire à proprement parler est mesurée en comparant les configurations Oracle_O et Oracle_M d'une part (sans l'étape de tatouage), et les configurations SSI-I basique et light-watermark SSI-I d'autre part (avec l'étape de tatouage). Pour ces deux comparaisons, et aussi bien pour les mesures de SDR que de SIR, une baisse de 2dB pour un mélange de 4 sources et 3dB pour un mélange de 5 sources, est provoquée par le groupement de deux coefficients TF adjacents en une molécule (les molécules sont de dimension 1×2). Les dégradations des performances de séparation engendrées par le groupement moléculaire sont donc beaucoup plus significatives que celles produites par l'insertion du tatouage à débit limité. Par conséquent, dans le cas où le tatouage est de débit limité (*i.e.* typiquement 64kbits/s), la résolution de traitement sera privilégiée devant une réduction du débit d'insertion du tatouage. En effet, un tatouage de 64kbits/s ne modifie pas suffisamment le signal de mélange pour affecter les performances du processus d'inversion, alors que le moyennage de l'index $\mathcal{I}_{ft}$ sur deux coefficients adjacents a des conséquences notables sur la qualité des sources estimées. Pour conclure à la fois sur l'influence du tatouage et celle de la résolution de traitement, il apparaît que la configuration SSI-I basique offre finalement un bon compromis entre un tatouage à débit raisonnable (quasiment sans effet sur les performances de séparation) et une résolution optimale de traitement.

En ce qui concerne l'influence du paramètre I_{ft} sur la qualité de la SSI-I, une comparaison des performances de la configuration SSI-I basique avec $I_{ft} = 2$ et avec $I_{ft} \leq I$ montre une amélioration moyenne du SDR de 1.3dB pour un mélange de 4 sources et 1.6dB pour un mélange de 5 sources, lorsque le nombre de sources supposées simultanément actives est libre (et non plus limité au nombre d'observations J du mélange), en accord avec les résultats obtenus par Vincent et al. [Vincent et al., 2007]. Le gain en performances de séparation permis par un nombre de sources actives potentiellement supérieur au nombre d'observations du mélange est donc limité mais tout de même appréciable. Considérer $I_{ft} \leq I$ implique un code à I bits, donc en l'occurrence 5 bits dans le cas d'un mélange de 5 sources (le débit de tatouage passe alors de 64 à 80kbits/s/voie) alors que pour le même mélange un code de 4 bits est suffisant pour encoder $I_{ft} = 2$. Ce léger accroissement du débit de tatouage, qui demeure cependant

8. la watermark maximale en regard du seuil d'inaudibilité fournit par le MPA utilisé, est insérée sur le signal de mélange

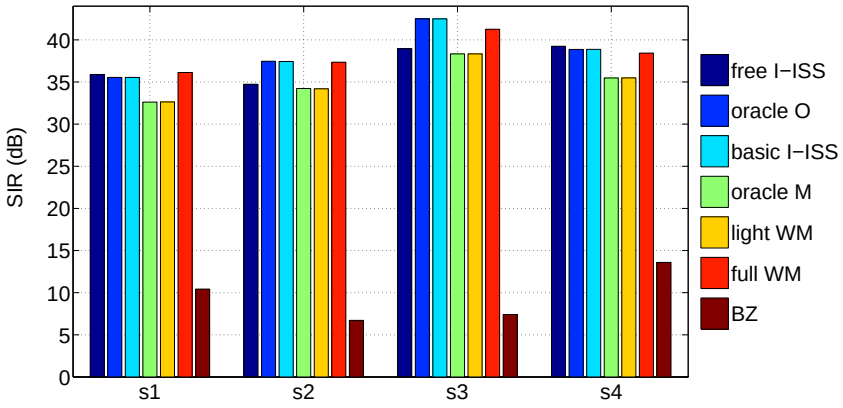
limité, n'a pas de conséquences notables sur les performances de séparation. Plus le nombre de sources composant le mélange (toujours supposé stéréophonique) est élevé, plus l'intérêt de considérer un nombre libre de sources simultanément actives se fait sentir. Notons que dans cette configuration, l'augmentation du nombre de combinaisons de sources à tester pour déterminer $\mathcal{I}_{ft}$ s'accompagne aussi d'un accroissement du temps de calcul au bloc 4 du codeur.

Enfin, la méthode de SSI-I est comparée avec la méthode de séparation semi-aveugle BZ [9]. Il apparaît clairement sur les Figures 6.5 et 6.6 que l'approche *informée* surpasse très largement les performances en mode semi-aveugle. Une amélioration de 10 à 13.5dB des scores de SDR est obtenue. Le critère de maximisation a posteriori choisi pour déterminer les sources prédominantes à chaque atome du plan TF permet à la fois une sélection optimale des sources, mais offre également l'ensemble des combinaisons de sources possibles, ce qui n'est pas le cas de la méthode géométrique d'estimation des sources proposée par Bofill et Zibulevski. Cette limitation de la méthode BZ transparaît dans le déséquilibre des performances de SDR entre les différentes sources d'un mélange. En effet, certaines combinaisons de sources n'étant pas sélectionnées par la méthode BZ, certaines sources sont très peu reconstruites, *i.e.* seule une faible proportion de leur pavage TF est non nul.

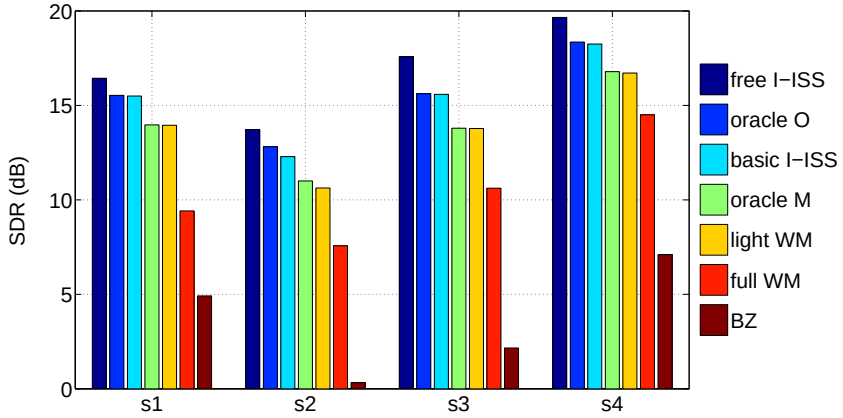
6.4 Conclusion sur la SSI-I

La technique de SSI-I décrite au cours de ce chapitre se base avant tout sur l'exploitation de la parcimonie des sources dans le domaine TF, pour réduire la dimension du mélange, ainsi que sur l'exploitation de la multi-dimensionnalité de ce mélange pour estimer les signaux sources par un processus d'inversion locale du mélange. La technique de séparation elle-même ainsi que l'information insérée par tatouage sur le signal de mélange sont toutes deux relativement simples. En comparaison des méthodes classiques aveugles ou semi-aveugles de séparation de sources basées sur la parcimonie et une inversion locale du processus de mélange, l'aspect informé de la SSI-I garantit une sélection optimale des sources prédominantes. Un autre avantage indéniable est bien sûr la transmission au décodeur de la matrice de mélange. La dégradation du signal de mélange par insertion d'un tatouage à un débit correspondant à la transmission de l'index des sources prédominantes est apparu comme négligeable sur le processus d'inversion du mélange, et a permis de réduire l'échelle de traitement à son niveau le plus élémentaire : l'ensemble du traitement de sélection des sources prédominantes et d'inversion du mélange peut ainsi être mené, en SSI-I, à l'échelle de chaque atome TF, au lieu de l'échelle moléculaire incontournable en SSI-C. La très faible dégradation du mélange par le tatouage et un traitement à l'échelle de chaque bin TF ont permis d'offrir des performances de séparation comparables aux performances de l'estimateur Oracle optimal décrit dans [Vincent et al., 2007].

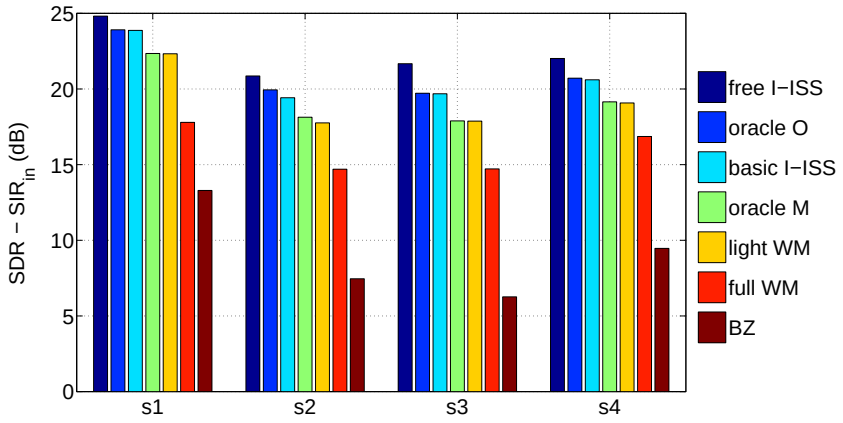
9. On rappelle que la matrice $\mathbf{A}$ est aussi supposée connue pour cette méthode de référence. La comparaison n'est faite que pour la phase d'estimation des sources.



(a) SIR.

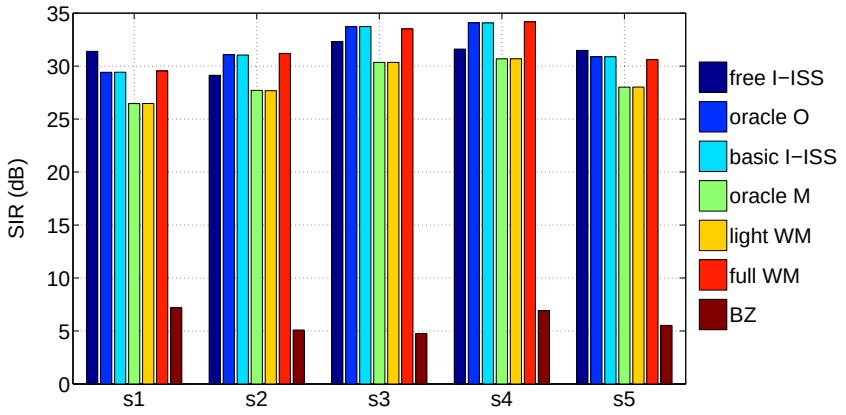


(b) SDR.

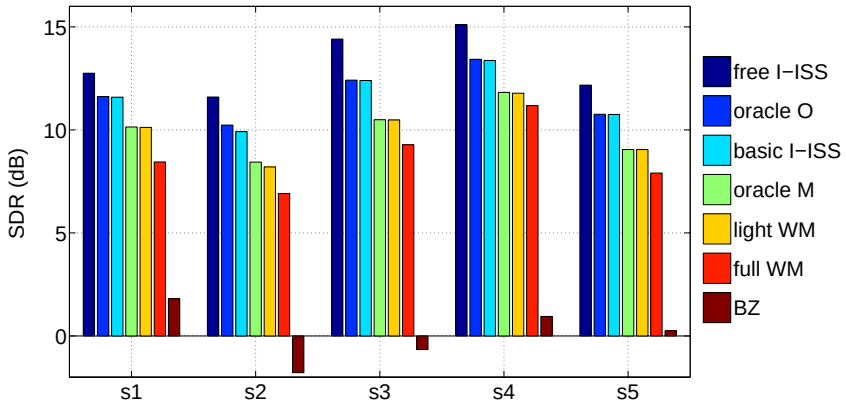


(c) SDR-SIR$_{\text{in}}$.

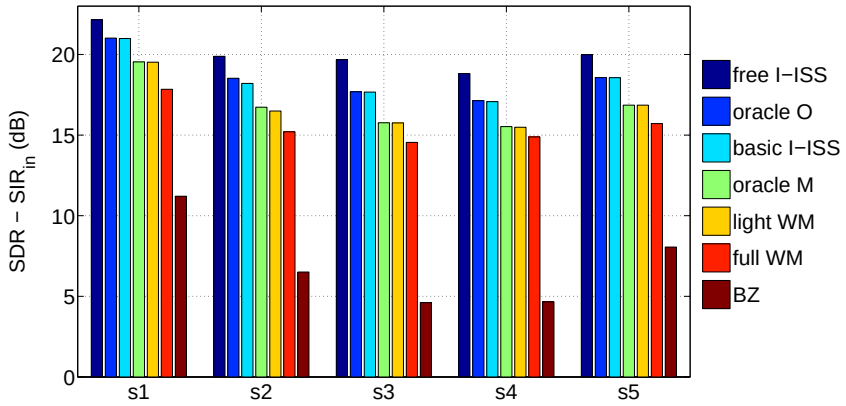
Figure 6.5 – Résultats de séparation pour l'ensemble des 7 configurations présentées Table 6.2. Performances moyennées sur 50 secondes de signal, provenant de 5 mélanges de 4 sources de différents styles musicaux. Les sources s1 à s4 sont un(e) guitare/piano, une batterie, une voix chantée, et une guitare basse.



(a) SIR.



(b) SDR.



(c) SDR-SIR$_{\text{in}}$.

FIGURE 6.6 – Résultats de séparation pour l'ensemble des 7 configurations présentées Table 6.2. Performances moyennées sur 50 secondes de signal, provenant de 5 mélanges de 5 sources de différents styles musicaux. Les sources s1 à s5 sont un(e) guitare/piano, une batterie, une voix chantée, une guitare basse, et l'une des trois sources trompette/choeurs/synthétiseur.

Une des limitations majeures de la technique de séparation de sources exploitant la parcimonie est précisément la superposition des signaux sources. Cette superposition peut s'avérer critique si l'on cherche à augmenter significativement le nombre d'instruments (au delà de 5). Autrement dit, la multiplication du nombre de sources ainsi que la similarité des signatures spectro-temporelles des sources composant un mélange implique une augmentation du risque de recouvrement des sources dans le plan TF. Dans le but de pallier à ce problème et de limiter la superposition des sources, une méthode envisagée, et développée au cours du chapitre suivant, consiste à combiner, dans approche hybride, la technique de SSI-C développée au Chapitre 5 avec la présente méthode de SSI-I. Dans un mélange stéréophonique composé par exemple de six sources, deux d'entre elles pourraient être extraites par codage, et les quatre autres sources estimées par SSI-I après que les deux premières sources décodées aient été soustraites du signal de mélange. C'est l'objet du Chapitre 7.

Chapitre 7

Une méthode hybride de séparation de sources informée couplant codage et indexation des sources

Nous introduisons dans ce chapitre une technique hybride de SSI, qui mêle l'approche par codage des signaux sources développée au Chapitre 5 (SSI-C) et l'approche basée sur la parcimonie des signaux sources introduite au Chapitre 6 (SSI-I). Dans cette nouvelle configuration, appelée SSI-CI pour Séparation de Sources Informée par Codage et Indexation des Sources, un sous-ensemble de sources est encodé par la technique de SSI-C, et son complémentaire est traité par inversion locale du sous-mélange privé des sources codées, selon la technique de SSI-I. Cette approche hybride permet de combiner les avantages des deux approches de SSI, par codage et par inversion du mélange. Nous nous plaçons, comme dans le chapitre précédent, dans la configuration d'un mélange linéaire instantané stationnaire *stéréophonique* (LISS) sous-déterminé.

Nous détaillons dans les sections suivantes les différentes étapes de la technique de SSI-CI, en particulier la phase de sélection des sources prédominantes et le procédé de séparation. Enfin des résultats obtenus pour des signaux de musique sont donnés à la Section 7.5.3 puis comparés à ceux obtenus par les techniques de SSI par codage et par indexation prises séparément. Des résultats de tests d'écoute sur la qualité des signaux audio estimés sont fournis à la Section 7.5.

7.1 Principes de la configuration hybride SSI-CI

L'objectif principal de la combinaison des approches de SSI par codage et par inversion est d'augmenter les performances de séparation obtenues par chacune des méthodes SSI-I et SSI-C considérées séparément. À terme, il s'agit également de fournir une méthode permettant de séparer des mélanges composés d'un plus grand nombre de sources que ceux traités jusqu'ici. En effet le risque de superposition des sources dans le plan TF augmente avec leur nombre, ce qui sous-entend une diminution des performances des méthodes de séparation basées sur la parcimonie. Or, un moyen

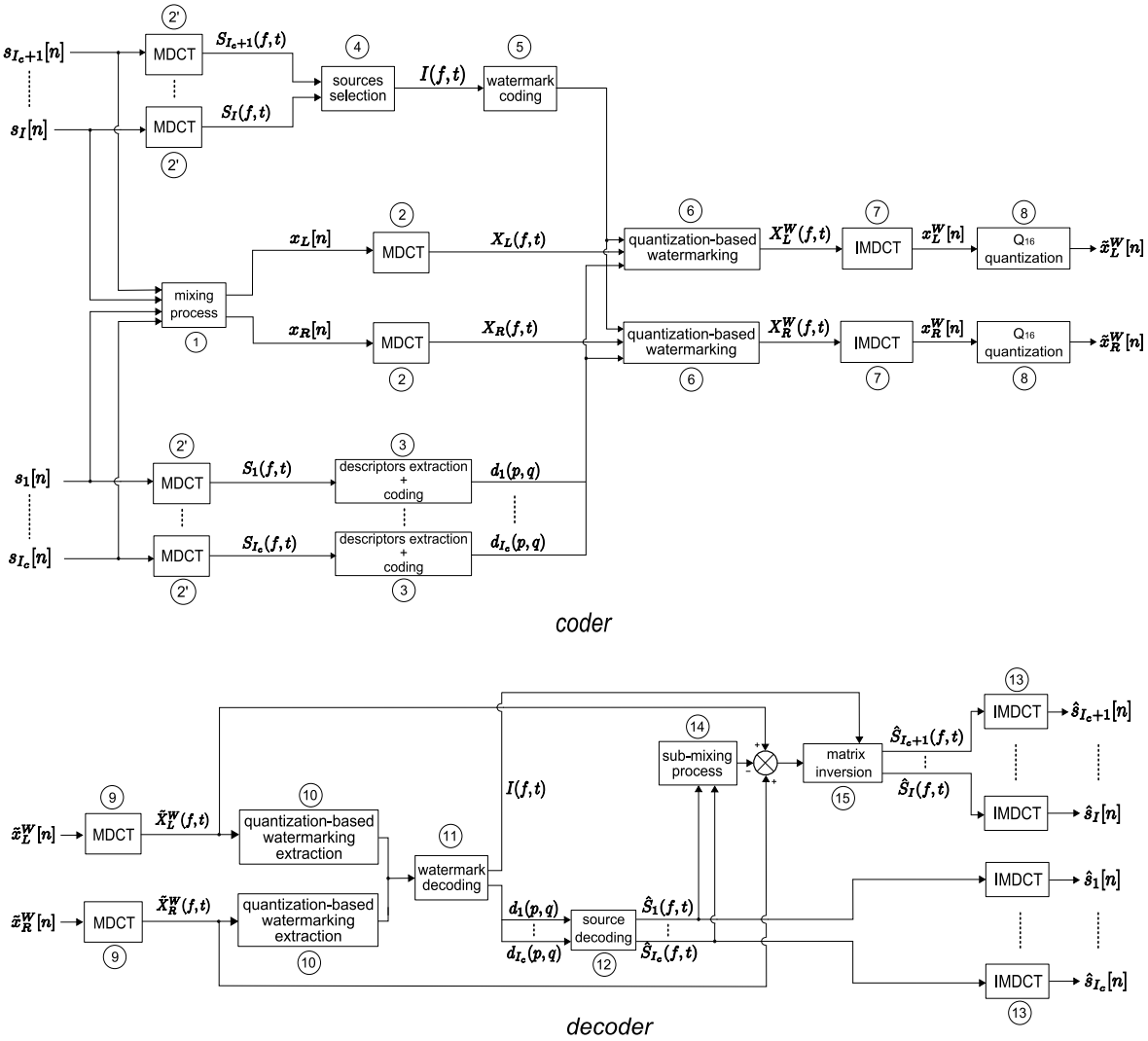


Figure 7.1 – Codeur/Décodeur du procédé hybride de SSI couplant indexation des sources actives et codage des signaux sources.

simple de réduire cette superposition des sources est d'utiliser la méthode de SSI-CI pour encoder un sous-ensemble d'entre elles et ainsi effectuer une première réduction de la dimension du mélange, avant même la réduction propre à la méthode SSI-I. Dans cette thèse, les tests menés se limitent à des mélanges à cinq sources pour des raisons de limitations de disponibilité de corpus, mais nous verrons que, même pour des mélanges à cinq sources, cette approche hybride permet de résoudre effectivement les problèmes de superposition résiduelle vus au Chapitre 6 (cf Section 6.3.2).

Au codeur, les signaux sources sont divisés en deux catégories : un sous-ensemble de I_c sources est encodé par la technique de codage présentée au Chapitre 5, alors que les index des sources prédominantes sont extraits du sous-mélange constitué des $I - I_c$ autres sources. Comme dans la méthode présentée au Chapitre 6, la combinaison de sources prédominantes est déterminée par un critère optimal a posteriori inspiré de l'estimateur Oracle introduit dans [Vincent et al., 2007]. Les deux informations correspondant au codage des I_c signaux sources encodés, et aux index des sources prédominantes parmi les $I - I_c$ autres sources sont toutes deux insérées dans le signal de

mélange par la technique de tatouage haute capacité introduite en 5.3. Au décodeur, la première étape du processus de séparation consiste à décoder les I_c signaux sources encodés par SSI-C, avant de soustraire ces sources décodées au signal de mélange tatoué. L'inversion du sous-mélange résultant, guidée par l'index des sources localement prédominantes, fournit les estimées des $I - I_c$ sources restantes. La réduction de la dimension du mélange par cette étape de codage permet de réduire significativement la superposition des signaux sources dans le plan TF, et de fait, permet d'accroître la qualité des sources estimées par SSI-I. De plus, la ressource nécessaire pour tatouer l'information des sources prédominantes étant relativement limitée en regard des capacités offertes par le MPA, une portion importante de la ressource peut être allouée à la SSI-C, permettant un codage performant des I_c signaux sources.

7.2 Description générale de la méthode hybride

Détaillons maintenant l'implémentation de la technique de SSI-CI. L'architecture duale codeur/décodeur propre à la séparation de sources informée en général, est bien évidemment conservée dans le cadre de la SSI-CI. Ce principe est ici illustré par le nouveau codeur/décodeur de la Figure 7.1. Dans cette section, nous présentons l'enchaînement des différents blocs de ce schéma, sans décrire leur contenu qui a déjà été détaillé dans les chapitres précédents. Nous apporterons des précisions sur les spécificités de cette combinaison hybride dans les sous-sections suivantes.

Les I_c signaux sources traités par SSI-C sont numérotés de 1 à I_c, et les $I-I_c$ signaux restants traités par SSI-I sont numérotés de $I_c + 1$ à I. Le processus de mélange LISS est réalisé au bloc 1 de la Figure 7.1. De façon similaire aux approches SSI-C et SSI-I, en SSI-CI, l'ensemble du traitement est réalisé dans le plan TF où la parcimonie naturelle des signaux sources audio est exploitée, tant pour réduire le coût de codage que pour limiter la superposition des signaux sources. La MDCT est à nouveau utilisée pour décomposer l'ensemble des signaux dans le plan TF (bloc 2 pour les signaux de mélange et bloc 2' pour les signaux sources). Les descripteurs des signaux sources s_1 à s_{I_c} sont extraits et codés au bloc 3. Parallèlement, le processus de sélection des sources prédominantes, propre à la SSI-I, est appliqué au bloc 4 aux sources s_{I_c+1} à s_I. La combinaison des index des sources actives à chaque bin TF est ensuite encodée au bloc 5. Les deux watermarks, celle de la SSI-C et celle de la SSI-I sont ensuite regroupées avant d'être insérées dans le signal de mélange au bloc 6, par la technique de tatouage par quantification. L'opération de synthèse du signal de mélange temporel par MDCT inverse est réalisée au bloc 7 avant conversion des échantillons temporels au format PCM 16-bits au bloc 8.

Au décodeur, seul le signal de mélange tatoué est disponible à l'utilisateur client. Au bloc 9, la décomposition MDCT du signal de mélange permet d'obtenir sa représentation TF. L'extraction de la watermark au bloc 10 est ensuite réalisée par le processus de quantification de décodage. Les deux informations, de codage et d'indexation des sources sont retrouvées au bloc 11. L'estimation des coefficients MDCT des signaux

sources $\hat{s}_1$ à $\hat{s}_{I_c}$ est réalisée au bloc 12 par décodage SSI-C, puis ces signaux sont soustraits au signal de mélange au bloc 14. Pour cela, on suppose, comme au Chapitre 6, que la matrice de mélange est connue au décodeur (on rappelle que son coût de transmission est supposé négligeable par rapport au coût de transmission des autres informations). L'index des sources prédominantes est ensuite utilisé pour estimer par inversion locale du mélange les coefficients MDCT des signaux $\hat{s}_{I_c+1}$ à $\hat{s}_I$ (bloc 15). Enfin, l'ensemble des I signaux estimés sont transformés par IMDCT dans le domaine temporel au bloc 13.

7.3 Détails d'implémentation

7.3.1 Une résolution de traitement différente pour le codage et l'inversion

La même résolution de MDCT est utilisée pour traiter l'ensemble des sources (à la fois celles codées par SSI-C et celles séparées par SSI-I) et le signal de mélange [1] : cela va de soi, il n'y a aucun intérêt à complexifier ce paramétrage. En revanche, une fois dans le plan TF, l'échelle de traitement des deux approches de séparation, par codage et par inversion, peut tout à fait différer. Nous avons vu au Chapitre 5 que l'information de codage des signaux sources utilisée en SSI-C est nettement plus volumineuse que celle encodant les index des sources prédominantes en SSI-I. En effet, de 4 à 26 bits étaient nécessaires, en fonction des descripteurs utilisés, pour encoder l'information de moyenne-gain-forme de chacun des signaux sources $s_i, i \in [1, \ldots, I_c]$ (cf Table 5.1). Un tel volume d'information n'est pas insérable à l'échelle d'un coefficient MDCT, et ce même si un MPA est utilisé pour déterminer la capacité disponible pour insérer l'information de tatouage : un traitement en SSI-C à l'échelle d'une molécule est donc à nouveau indispensable. Le choix de la taille de molécule doit là encore respecter un compromis : ne pas être trop faible pour que la quantification vectorielle soit efficace, et ne pas être trop grande pour limiter la superposition des sources. Étant donnée la forte capacité de tatouage permise par le MPA, une taille de 1×4 (un bin fréquentiel, et 4 bins temporels) offre un bon compromis et sera adoptée dans ce chapitre. En revanche, comme nous l'avons vu au Chapitre 6, l'information d'indexation des sources prédominantes est relativement compacte, et l'inversion du mélange est significativement affectée par le groupement moléculaire (cf Section 6.3.4), c'est pourquoi la SSI-I est effectuée, dans l'approche SSI-CI, à l'échelle de chaque bin TF.

1. On utilise une fenêtre d'analyse de 2048 échantillons, soit 46.5ms à la fréquence d'échantillonnage $f_e = 44.1kHz$. Il a été vu au Chapitre 6, et étudié dans [Nesbit and Plumbley, 2008] que la parcimonie des signaux de musique est maximale pour cette résolution.

7.3.2 Une SSI-C modifiée

Contrairement aux choix faits au Chapitre 5, en SSI-CI, il est choisi d'utiliser des dictionnaires de prototypes scalaires pour encoder les descripteurs de gain et de moyenne des molécules. Ce choix permet de s'affranchir de la transmission des facteurs d'échelles des quantificateurs scalaires uniformes utilisés en SSI-C pour quantifier moyenne et gain (cf Section 5.1.4.1) et simplifier ainsi globalement la procédure de codage. Par conséquent, pour une molécule $M_{pq}^{s_i}$ du signal source s_i au canal fréquentiel p et temporel q, encodée par le triplet de descripteurs moyenne-gain-forme (cf Section 5.1.4), l'information de gain/moyenne transmise au décodeur en SSI-CI consiste ici en l'indice (dans le dictionnaire correspondant) du coefficient scalaire le plus proche du gain/moyenne de $M_{pq}^{s_i}$ au sens de la distance Euclidienne. En contrepartie du gain de ressources permis par l'utilisation de dictionnaires de gain et de moyenne, il est nécessaire que ces dictionnaires soient connus au codeur comme au décodeur.

7.3.3 Détails de la combinaison codage et inversion

Dans l'approche hybride SSI-CI, le procédé de SSI-I ne concerne qu'un sous ensemble de $I - I_c$ sources, s_{I_c+1} à s_I. Autrement dit, le signal de mélange utilisé dans le processus d'inversion (cf équation (6.5)) est un sous-mélange constitué des $I-I_c$ sources s_{I_c+1} à s_I, et la combinaison $\mathcal{I}_{ft}$ de J sources prédominantes au bin TF (f,t) est déterminée parmi l'ensemble d'indices $i \in [I_c+1,\ldots,I]$. En accord avec les notations introduites au Chapitre 6, on note $\mathbf{S}_c$ le vecteur des sources estimées par SSI-C, et $\mathbf{A}_c$ la sous-matrice de la matrice de mélange $\mathbf{A}$ constituée des colonnes correspondantes. La matrice $\mathbf{A}_{\bar{c}}$ désigne la sous-matrice complémentaire de $\mathbf{A}_c$ dans $\mathbf{A}$, c'est-à-dire la sous-matrice de $\mathbf{A}$ composée des colonnes correspondant aux sources séparées par SSI-I, notées $\mathbf{S}_{\bar{c}}$. Toujours en accord avec les notations du Chapitre 6, la matrice $\mathbf{A}_{\mathcal{I}_{ft}}$ est donc ici une sous-matrice de la sous-matrice de mélange $\mathbf{A}_{\bar{c}}$, et $\hat{\mathbf{S}}_{\mathcal{I}_{ft}}$ est le vecteur des sources prédominantes estimées parmi $\mathbf{S}_{\bar{c}}$.

Au codeur, le sous-mélange des sources à estimer par SSI-I est donc $\mathbf{X}_{\bar{c}} = \mathbf{A}_{\bar{c}}\mathbf{S}_{\bar{c}}$. Ce sous-mélange peut aussi s'écrire en soustrayant au mélange initial les sources à encoder par SSI-C, soit $\mathbf{X}_{\bar{c}} = \mathbf{X} - \mathbf{A}_c\mathbf{S}_c$. La recherche de la combinaison de sources optimale $\mathcal{I}_{ft}$ définie à la Section 6.2.2 est ici appliquée au processus d'inversion suivant

$$\left\{ \begin{array}{rcl} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) & = & \mathbf{A}_{\mathcal{I}_{ft}}^{\dagger}(\mathbf{X}(f,t) - \mathbf{A}_c\ \mathbf{S}_c(f,t)) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}_{ft}}}(f,t) & = & \mathbf{0} \end{array} \right. \tag{7.1}$$

Au décodeur, les sources $[\hat{S}_{I_c+1},\ldots,\hat{S}_I]$ sont estimées au bloc 15 de la Figure 7.1 après que les signaux $[\hat{S}_1,\ldots,\hat{S}_{I_c}]$, estimés par SSI-C au bloc 12 aient été soustraits du mélange tatoué au bloc 14 (avec les gains de la matrice $\mathbf{A}_c$ correspondant). En d'autres termes, le processus d'inversion de l'équation (6.11) devient au bloc 15

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) &= \mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}(\tilde{\mathbf{X}}^{W}(f,t) - \mathbf{A}_c\ \hat{\mathbf{S}}_c(f,t)) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}_{ft}}}(f,t) &= \mathbf{0} \end{cases} \tag{7.2}$$

Illustrons ces équations sur un exemple d'un mélange LISS de 5 sources, qui sera la configuration choisie dans les expérimentations. Le vecteur source est ici : $\mathbf{S}(f,t) = [S_1(f,t), S_2(f,t), S_3(f,t), S_4(f,t), S_5(f,t)]^{\mathrm{T}}$. Si s_2 et s_4 sont les deux signaux sources encodés par SSI-C, et si $\mathcal{I}_{ft} = \{3,5\}$, *i.e.* si s_3 et s_5 sont les deux sources prédominantes au bin TF (f,t), alors $\mathbf{A}_{\mathcal{I}_{ft}} = [\mathbf{A}_3, \mathbf{A}_5] = \mathbf{A}_{35}$, $\mathbf{S}_{\mathcal{I}_{ft}}(f,t) = [S_3(f,t), S_5(f,t)]^{\mathrm{T}}$, et l'ensemble des sources non prédominantes parmi $[s_1,\ s_3,\ s_5]$ est réduit à la source s_1, d'où $\mathbf{S}_{\overline{\mathcal{I}}_{ft}}(f,t) = \{S_1(f,t)\}$. Les coefficients MDCT $\hat{S}_2(f,t), \hat{S}_4(f,t)$ sont estimés par SSI-C, et $\hat{S}_1(f,t), \hat{S}_3(f,t)$ et $\hat{S}_5(f,t)$ sont estimées selon l'équation

$$\begin{cases} \hat{\mathbf{S}}_{35}(f,t) &= \mathbf{A}^{\dagger}_{35}(\tilde{\mathbf{X}}^{W}(f,t) - \mathbf{A}_{24}\ \hat{\mathbf{S}}_{24}(f,t)) \\ \hat{S}_1(f,t) &= 0 \end{cases} \tag{7.3}$$

En SSI-CI, le processus d'inversion du mélange de la SSI-I peut être doublement affecté par les dégradations subies par le signal de sous-mélange entre le codeur et le décodeur, comme le montrent les équations (7.1) et (7.2). Une première dégradation concerne directement le signal de mélange initial. Elle intervient au bloc 6 du codeur, par l'insertion de la watermark globale composée des informations de codage et des index des sources prédominantes qui transforme le signal de mélange $\mathbf{X}(f,t)$ en $\mathbf{X}^{W}(f,t)$ [2]. En fonction de la taille de la watermark insérée, la détérioration $\mathbf{X}(f,t) - \mathbf{X}^{W}(f,t) \simeq \mathbf{X}(f,t) - \tilde{\mathbf{X}}^{W}(f,t)$, peut plus ou moins significativement affecter le processus d'inversion.

Une seconde source de dégradation affecte les signaux sources, et intervient au décodeur lors de la phase d'estimation des signaux $\hat{S}_1$ à $\hat{S}_{I_c}$ au bloc 12. Le bruit de codage $\mathbf{S}_c(f,t) - \hat{\mathbf{S}}_c(f,t)$ des sources s_1 à s_{I_c} impacte le signal de sous-mélange lors de la phase de soustraction des sources décodées $\hat{\mathbf{S}}_c$ au bloc 14, et par conséquent agit directement sur le processus d'estimation des sources par SSI-I, comme le montre l'équation (7.2). Dans l'approche SSI-CI, un compromis doit donc être trouvé entre la qualité de codage de $\hat{\mathbf{S}}_c$ et la dégradation de $\mathbf{X}^{W}$, par un réglage adéquat de la taille de l'information de codage (puisque l'information d'indexation est de taille fixe). Plus le débit d'information insérée par tatouage augmente, *i.e.* plus le débit de codage augmente, meilleur est le codage des signaux sources s_1 à s_{I_c}, et de fait, plus le bruit de codage est réduit. Cependant, un débit de codage élevé implique également une plus grande dégradation du mélange par le tatouage. Rappelons que le MPA utilisé permet de régler la capacité disponible en fonction des besoins (dans la limite de l'inaudibilité), et donc d'étudier les performances globales de séparation selon plusieurs réglages du

2. Rappelons que $\mathbf{X}^{W}$ subit aussi la conversion au format PCM 16 bits, transformant le signal en $\tilde{\mathbf{X}}^{W}$, mais les effets de cette conversion sont supposés négligeables devant ceux du tatouage (cf Section 6.2.3).

débit de codage. Nous discutons ce principe du réglage du débit de codage dans la section suivante.

7.4 Allocation de la ressource de tatouage : principes généraux et exemple

En SSI-CI, l'allocation de bits est un problème multi-contraintes résultant de la conjonction des contraintes propres à chacune des méthodes de SSI-C et de SSI-I. En SSI-C, c'est la capacité d'insertion du tatouage, elle-même déterminée par le MPA, qui détermine l'allocation de la ressource à la fois entre les différents signaux sources à coder et entre les différents descripteurs utilisés pour encoder chaque signal source. C'est cette allocation de la ressource de codage entre les différentes sources et les différents descripteurs qui conditionne directement les performances de séparation. Dans l'approche SSI-I, l'information correspondant aux index des sources prédominantes est de taille fixe, dépendant uniquement du nombre de sources à estimer par inversion du mélange. Par conséquent, la capacité nécessaire pour insérer l'information de SSI-I est fixe. C'est donc en fonction de chacune de ces caractéristiques que doit être faite l'allocation de la ressource en SSI-CI. Concernant la distribution de l'information de tatouage sur le signal de mélange, elle est, contrairement à l'allocation de la ressource de tatouage, complètement arbitraire. Dans le but d'encoder les portions basses fréquences des signaux avec le plus de précision possible, le choix est fait d'insérer l'information (limitée) utile à la SSI-I dans les hautes fréquences. En d'autres termes, on réserve la ressource basses fréquences pour coder cette portion des signaux sources (rappelons que l'information de codage d'une zone du plan TF d'une source est insérée sur cette même zone du plan TF du mélange), et on code l'information d'index des sources prédominantes dans les hautes fréquences.

Reprenons l'exemple d'un mélange de 5 signaux sources introduit à la section précédente et qui sera utilisé dans les expérimentations. Nous fixons ici une première configuration d'allocation de la ressource nécessaire pour encoder les informations utiles aussi bien à la SSI-I qu'à la SSI-C. Il est choisi d'encoder les deux signaux sources s_2 et s_4 par SSI-C, et par conséquent, le sous-mélange des sources à séparer par SSI-I est constitué des trois signaux sources s_1, s_3 et s_5. Il existe donc un total de $\binom{I-I_c}{J} = \binom{3}{2} = 3$ combinaisons de 2 sources prédominantes parmi les 3 sources composant le sous-mélange. Un message de 2 bits est donc suffisant pour encoder l'information de SSI-I à chaque bin TF ce qui correspond à un débit de 32kbits/s/voie pour encoder l'information de SSI-I relative à la bande fréquentielle [0-16kHz] [3] (voir Section 6.2.4). Le seuil de masquage du MPA est réglé de manière à fournir une capacité de 150kbits/s/voie. Ce taux fixé empiriquement est un compromis entre une capacité confortable de codage nécessaire pour encoder avec précision les deux signaux sources, et une dégradation limitée du signal de mélange lors de l'insertion du tatouage, cette dernière condition étant fixée par

3. Rappelons que la SSI n'est pas effectuée au delà de 16kHz, la faible sensibilité de l'oreille humaine à des fréquences supérieures permettant de "négliger" ces portions de signal.

le processus d'inversion du mélange. Rappelons en effet qu'au Chapitre 6, on obtenait une dégradation négligeable des performances de séparation pour un signal de mélange tatoué avec un débit de 64 à 80kbits/s, et on obtenait une dégradation significative avec un débit de l'ordre de 250kbits/s. De plus, un débit de codage de 80-32=48kbits/s n'apparaît pas suffisant pour encoder correctement une source. On choisit donc ici un débit intermédiaire de 150kbits/s. La capacité totale étant de 150kbits/s/voie et 32kbits/s/voie étant dédiés au tatouage de la watermark de SSI-I, 118kbits/s/voie (*i.e.* ici 118kbits/s/source) sont disponibles pour encoder les descripteurs SSI-C des sources s_1 et s_4.

7.5 Expérimentations

Les résultats de séparation par SSI-CI présentés dans cette section sont effectués sur les mêmes mélanges linéaires instantanés de 5 signaux de musique (voix+instruments) échantillonnés à 44100Hz que ceux présentés à la Section 6.3. Rappelons que les signaux sources sont, selon le morceau considéré : s1 = une guitare ou un piano, s2 = une batterie (une piste pour l'ensemble de la batterie), s3 = une voie chantée (homme ou femme), s4 = une guitare basse, s5 = une trompette, des choeurs, ou un synthétiseur. Les tests sont effectués sur 5 extraits de 10 secondes chacun, de styles musicaux différents.

7.5.1 Limitation de la superposition des signaux sources par l'étape de codage

Rappelons que l'objectif principal de la combinaison des méthodes SSI-C et SSI-I est de réduire la superposition des signaux source dans la plan TF de manière à augmenter les performances de séparation qui seraient obtenues par SSI-C ou SSI-I seules. Une première mesure consiste à vérifier la validité de l'hypothèse faite en SSI-CI, en effectuant une mesure croisée de la superposition des sources en fonction de leur énergie, comme nous l'avons vu à la Section 6.3.2. Les résultats sont présentés à la Table 7.1 pour les cinq mélanges de test. Le pourcentage moyen [4] de l'énergie totale de chaque source en fonction de son rang énergétique au sein du mélange est fourni pour un mélange initial de cinq sources, Table 7.1a, et un de ses sous-mélanges composé de trois sources, Table 7.1b. Dans le cas où le mélange est consitué des 5 signaux sources s_1 à s_5, la Table 7.1a fait apparaître que de 89.7 (pour s_5) à 98.8% (pour s_3) de l'énergie totale de chaque signal source est concentrée aux bins TF où cette source est parmi les deux sources les plus énergétiques du mélange. Cependant, il apparaît que 9.5, 1.2 et 10.3% de l'énergie des sources s_1, s_3 et s_5 respectivement ne sont pas reconstruits par SSI-I. En effet, rappelons, en accord avec l'équation (6.11), que les estimées des sources considérées comme non-prédominantes à chaque bin TF sont mises à zéro lors de la phase d'estimation des signaux sources. Un pourcentage de l'ordre de 10% n'est

4. Moyenné sur les 5 mélanges de la Table 6.1.

pas négligeable, et cette mise à zéro d'une si large portion de signal peut entraîner des artefacts, notamment du type bruit musical. La Table 7.1b présente les résultats correspondants pour un sous-mélange des trois sources s_1, s_3 et s_5, traités par SSI-I, les signaux s_2 et s_4 étant encodés par SSI-C. La soustraction des sources s_2 et s_4 au mélange a pour conséquence de réduire à 3.4, 0.3 et 3.6% respectivement le pourcentage d'énergie totale de s_1, s_3 et s_5 qui est mis à zéro dans le processus de SSI-I. Il apparaît donc clairement que réduire la dimension du mélange en soustrayant les sources destinées à être codées diminue la superposition des signaux sources, et par conséquent entraîne le traitement par SSI-I d'une plus grande portion des signaux non-codés. Notons qu'en plus d'augmenter le nombre de bins TF où chaque source est sélectionnée, la réduction du mélange influe également sur les performances de séparation aux bins TF où un signal était déjà considéré comme prédominant dans le mélange à $I = 5$ sources. En effet, soustraire I_c sources, ici s_2 et s_4, permet de supprimer le "bruit" causé par ces sources aux bins TF où les $I - I_c$ sources non encodées, ici s_1, s_3 et s_5, figuraient parmi les sources prédominantes (dans le mélange initial à 5 sources). En résumé, diminuer la dimension du mélange permet de reconstruire plus de bins TF (parmi les bins mis à zéro avec le mélange initial, certains sont désormais reconstruits avec le sous-mélange), et avec une meilleure qualité (aux bins déjà reconstruits avec le mélange original, mais où les sources ensuite encodées par SSI-C dégradaient le signal de mélange).

7.5.2 Comparaison des trois configurations de SSI

Différentes configurations de SSI sont testées pour évaluer les performances de séparation de la méthode hybride SSI-CI en comparaison des performances obtenues par la méthode par codage SSI-C du Chapitre 5 et par la méthode SSI-I par inversion du mélange du Chapitre 6. Les différentes configurations des algorithmes testés sont présentées à la Table 7.2.

L'algorithme SSI-C fait référence à la technique de SSI basée sur le (seul) codage des signaux sources. Le taux d'insertion moyen est similaire sur chacune des deux voies et de l'ordre de 290kbits/s/voie. On choisit ici un débit d'insertion proche du maximum permis par les derniers développements du système de tatouage de [Pinel et al., 2009]. La dimension de molécule choisie est 1×4. L'information de codage est distribuée sur les deux voies du mélange. Les mélanges considérés étant composés de cinq signaux sources, il est choisi d'encoder les deux signaux sources ayant le spectre le plus étalé, soit s_1 et s_2, sur la première voie, et les trois autres sources, soit s_3, s_4 et s_5, sur la deuxième voie. Le débit de codage est donc de $290/2 = 145$kbits/s/source sur la première voie, et $290/3 \approx 97$kbits/s/source sur la deuxième voie[5]. Ce choix est arbitraire (même s'il semble le plus logique), et une allocation à la fois différente pour

5. Ces débits sont importants par rapport aux débits typiques des algorithmes de compression courants de type MPEG, pour des sources monophoniques. Cependant, nous n'avons pas la prétention d'égaler ces algorithmes avec le codage vectoriel relativement simple que nous utilisons. Dans le futur, un codage non-propriétaire pourra être utilisé de manière à encoder plus efficacement les sources ou à encoder un plus grand nombre de sources. Ce point sera rediscuté dans à la section 7.7

Rang	s_1	s_2	s_3	s_4	s_5
1	69.1	86.6	92.1	87.6	65.4
2	21.3	10.7	6.7	10.7	24.3
3	7.3	2.0	1.0	1.5	8.5
4	1.9	0.5	0,2	0.2	1.7
5	0.3	0.1	3.10^{-2}	3.10^{-2}	0.2

(a) Mélange original.

Rang	s_1	s_3	s_5
1	80.7	93.5	79.8
2	15.9	6.2	16.7
3	3.4	0.3	3.6

(b) Sous mélange composé de s_1, s_3 et s_5.

Table 7.1 – Pourcentage de l'énergie totale de chaque signal source en fonction de son rang énergétique dans le mélange (moyenné sur 5 mélanges de 10 secondes de musiques de différents styles). Étude à l'échelle de l'atome temps-fréquence avec un mélange composé (a) des 5 signaux sources originaux, (b) des 3 sources s_1, s_3 et s_5, les sources s_2 et s_4 étant encodés par SSI-C.

Algorithme	Code SSI-I (bits)	débit SSI-I (kbits/s/voie)	débit SSI-C (kbits/s)	Tatouage total (kbits/s/voie)
SSI-C	-	-	$2 \times 145 + 3 \times 97$	290
SSI-I	4	64	-	64
SSI-CI	2	32	$118 + 118$	150
SSI-CI opt.	2	32	$150 + 86$	150

Table 7.2 – Configurations des algorithmes testés.

chaque source et n'impliquant pas un nombre entier de sources codées par voie pourait également être appliquée. Une étude systématique de l'allocation du débit de codage entre les différentes sources composant des mélanges musicaux, en fonction de la nature des sources et de leur combinaison dans le mélange est un travail très conséquent qui reste à mener et qui dépasse le cadre de cette thèse.

L'algorithme SSI-I fait référence à la technique de SSI basée uniquement sur la parcimonie des signaux sources et l'inversion du signal de mélange. Le processus est

réalisé à l'échelle d'un seul bin TF. Le mélange étant stéréophonique et composé de 5 signaux sources, le nombre card($\mathcal{P}$) de combinaisons de 2 signaux parmi les 5 signaux sources est de $\binom{5}{2} = 10$, et de fait, un message de 4 bits est suffisant pour encoder la combinaison $\mathcal{I}_{ft}$ de sources prédominantes à chaque bin TF, ce qui correspond à un débit de 64kbits/s/voie pour traiter la plage fréquentielle [0-16kHz] (cf Section 6.2.4.1).

Deux configurations de la nouvelle approche hybride SSI-CI sont également testées. La distinction entre ces deux configurations tient à l'allocation de la ressource de codage entre les deux sources s_2 et s_4. La première configuration, notée SSI-CI, est celle de l'exemple de la Section 7.4 : codage de s_2 et s_4 à 118kbits/s, et séparation par inversion de s_1, s_3 et s_5 avec une information d'index à 32kbits/s/voie (soit un débit total de tatouage de 150kbits/s/voie). Dans cette configuration une ressource identique est allouée aux deux signaux source à encoder, alors que dans la deuxième configuration, notée SSI-CIopt., l'allocation de la ressource de codage tient compte du contenu spectral de ces deux sources. Dans les tests que nous avons menés, s_2 est la batterie et s_4 est la guitare basse. Le spectre de la guitare basse est particulièrement parcimonieux et concentré dans les basses fréquences (BF), typiquement dans la bande [0-5kHz]. Au contraire, le spectre de la batterie est lui beaucoup plus étalé dans toute la bande de fréquence [0-16kHz] [6]. Une capacité de codage supérieure est donc allouée au signal s_2 de manière à encoder les descripteurs (moyenne, gain et forme) du signal de batterie dans toute la bande de fréquences [0-16kHz]. Un débit moyen de 150kbits/s est alloué au codage de la source s_2, alors que seulement 86kbits/s sont alloués au codage de s_4 (on garde donc le même débit total de tatouage de 150kbits/s/voie). On notera que dans ces deux configurations, le codage de la guitare basse ne concerne que la bande [0-5kHz], qui représente 99.98% de l'énergie totale du signal [7] (ce n'est pas le cas pour la configuration SSI-C). Une amélioration des performances de séparation est attendue d'une telle allocation, plus adaptée à chaque signal.

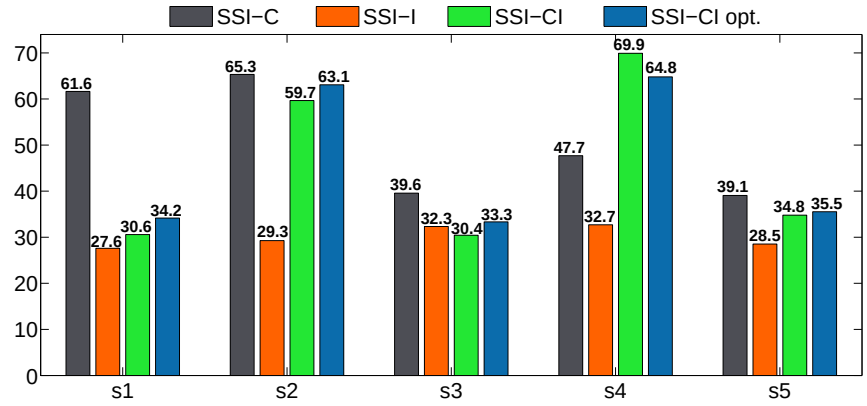
7.5.3 Résultats de séparation

La qualité des sources séparées a été jugée par des tests d'écoute informels avec un casque de qualité, ainsi que des mesures de performances de séparation de sources introduites dans [Vincent et al., 2005]. Tout comme au Chapitre 6, une mesure SNR_{in} permettant de juger de la difficulté à séparer chaque signal source en fonction de sa contribution dans le mélange est également fournie. On rappelle que le SNR_{in} moyen des sources s_1 à s_5 vaut respectivement -9.4, -8.3, -5.3, -3.7 et -7.8dB, mettant en évidence la plus grande difficulté à estimer les sources s_1, s_2 et s_5 par rapport à s_3 et s_4.

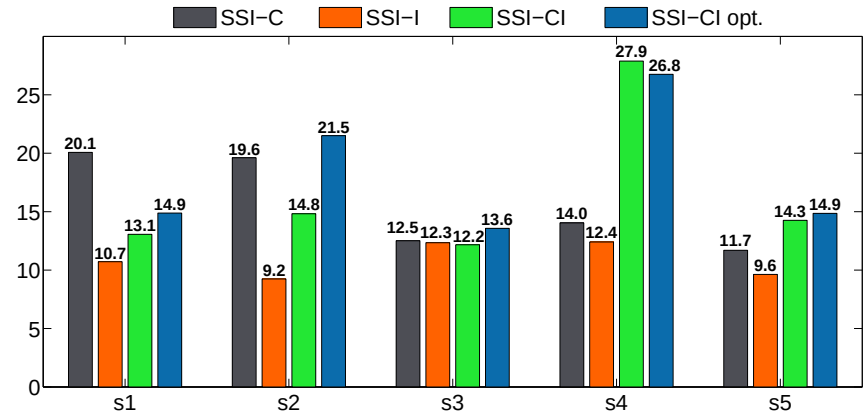
Les performances de séparation moyennées sur les 5 mélanges de test sont présentées Figure 7.2. Il apparaît tout d'abord que de très bonnes performances de séparation sont obtenues pour chacune des quatre configurations de SSI testées, et ce pour l'ensemble des mesures utilisées. Une très bonne réjection des interférences produites par

6. La piste de batterie est en fait composée de plusieurs instruments de spectres très différents comme une grosse caisse (basses fréquences), une caisse claire, et des cymbales.

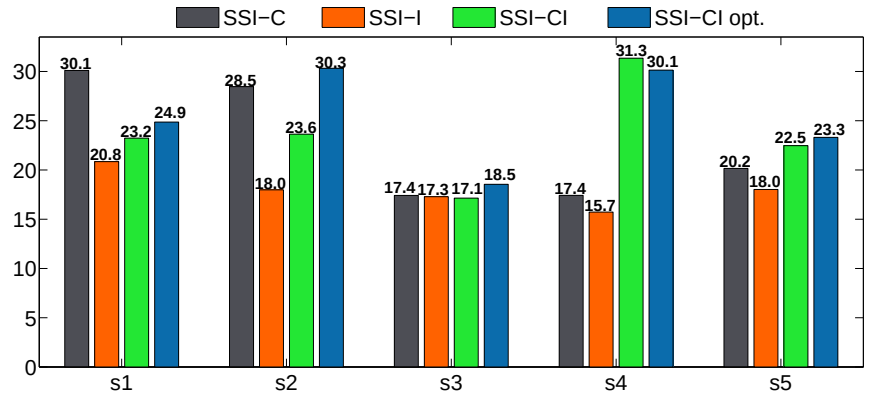
7. Moyenne calculée sur les 5 signaux de guitare basse de 10 secondes utilisés dans nos 5 mélanges.



(a) SIR (dB)



(b) SDR (dB)



(c) SDR-SNR_{in} (dB)

Figure 7.2 – Résultats de séparation pour l'ensemble des 4 configurations présentées Table 7.2. Performances moyennées sur 50 secondes de signal, provenant de 5 mélanges de différents styles musicaux, de 5 sources. Les sources s1 à s5 sont un(e) guitare/piano, une batterie, une voix chantée, une guitare basse, et l'une des trois sources trompette/choeurs/synthétiseur.

les autres sources est obtenue comme en témoignent des valeurs de SIR comprises entre 27.6 et 69.9dB. Chacune des sources est clairement isolée, ce qui est confirmé par les

tests d'écoute. Des valeurs élevées de SDR confirment une très bonne estimation de chacun des signaux sources, et une qualité globale de séparation en accord avec les performances obtenues par codage (cf Chapitre 5, Section 5.3) et par inversion du processus de mélange (cf Chapitre 6). La bonne isolation de chaque source, les artefacts réduits et plus généralement la bonne qualité audio des signaux estimés permet tout à fait de procéder à un rehaussement ou au contraire une réduction d'un signal au sein du mélange jusqu'à complète "extinction" de la source (application remix/karaoké par simple addition ou soustraction que ce soit dans le domaine MDCT ou dans le domaine temporel sur les signaux resynthétisés). Dans le cas de la soustraction d'un signal estimé au mélange, les possibles artefacts de cette source qui apparaissent lorsqu'elle est considérée individuellement sont masqués, au sein du mélange, par les autres sources. Les méthodes de SSI hybrides permettent même une écoute de qualité de chaque source estimée individuellement, comme nous le verrons plus tard. Les exemples audio correspondant aux différentes configurations de la Table 7.2 peuvent être téléchargés à l'adresse `http ://www.gipsa-lab.inpg.fr/~mathieu.parvaix/SSI-CI-demo.rar`. Le package inclut les signaux de mélange originaux et tatoués, ainsi que les signaux sources et les signaux estimés. Tous les signaux sont normalisés de telle sorte qu'il soit possible, pour le lecteur intéressé, de procéder à un remixage à partir des signaux séparés et du signal de mélange tatoué.

Considérons maintenant plus en détails les performances obtenues pour chacune des quatre configurations testées. Des mesures de SDR comprises entre 11.7 et 20.1dB attestent des bonnes performances de l'approche par codage SSI-C en terme de reconstruction de chaque signal source individuellement, en particulier étant donné les faibles $\mathrm{SNR}_{\mathrm{in}}$ pour l'ensemble des sources. Les différences de performances entre les sources s_1/s_2 et les sources $s_3/s_4/s_5$ s'expliquent par la plus grande ressource de codage allouée aux sources s_1 et s_2 avec, comme nous l'avons vu, 145kbits/s/source, contre 97kbits/s/source pour $s_1/s_2/s_3$. Une ressource 1.5 fois supérieure pour s_1 et s_2 permet un meilleur codage des signaux sources, à la fois par une quantification plus précise du descripteur de gain, mais également un recours plus systématique au descripteur de forme dont il a été vu qu'il permet un meilleur codage des signaux sources que lorsque seul le descripteur de gain d'une molécule est utilisé (cf Chapitre 5). L'utilisation de la combinaison de descripteurs moyenne-gain-forme permet de plus, en utilisant des molécules prototypes de dictionnaires de forme, de ne pas avoir recours au mélange lors du processus de reconstruction. Il est par conséquent possible de s'affranchir de la superposition des sources dans le signal de mélange, ce qui conduit à une meilleure reconstruction des signaux comme l'attestent des scores de SDR supérieurs à 19.5dB pour s_1 et s_2 et variant de 11.7 à 14.0dB pour s_3 à s_5, et à une forte réjection des interférences démontrée par des SIR supérieurs à 60dB pour s_1 et s_2, et variant de 39.1 à 47.7dB pour s_3 à s_5. Notons que les sources s_1 et s_2 présentant des $\mathrm{SNR}_{\mathrm{in}}$ légèrement plus faibles que les autres sources, leurs performances de codage sont accentuées dans les mesures de SDR-$\mathrm{SNR}_{\mathrm{in}}$, comme l'illustre la Figure 7.2c.

Les performances de l'approche SSI-I par inversion du mélange, sont celles présen-

tées au Chapitre 6, Section 6.3.4 dans la configuration SSI-I basique pour un mélange de 5 sources (les mêmes mélanges sont utilisés dans cette section). Rappelons que les résultats obtenus par la méthode SSI-I sont similaires aux performances de l'estimateur Oracle en supposant $J = 2$ sources simultanément actives, *i.e.* les meilleures performances atteignables par inversion locale du mélange. Ces performances, relativement élevées, sont sans commune mesure avec les performances atteignables par une technique classique de séparation aveugle de sources. Les signaux estimés par SSI-I présentent, pour certains, des artefacts ne permettant pas une écoute transparente [8], mais ils sont d'une qualité tout à fait suffisante pour une application de type remix/karaoké. Nous remarquons cependant que les performances obtenues par SSI-I sont ici inférieures aux performances de codage SSI-C. Ceci s'explique par la forte augmentation du débit de codage permise par les derniers développements du MPA.

L'intérêt de l'approche hybride SSI-CI apparaît clairement sur la Figure 7.2. Rappelons que dans cette configuration les sources s_2 et s_4 sont encodées par SSI-C et les sources s_1, s_3 et s_5 sont traitées par SSI-I. Des scores très élevés de SDR et SIR sont obtenus pour les sources s_2 (batterie) et s_4 (guitare basse). La source s_4 qui était allouée 97kbits/s pour encoder l'ensemble de la bande [0-16kHz] dans la configuration SSI-C seule, est désormais, dans la configuration SSI-CI, allouée 118kbits/s pour encoder la bande [0-5kHz], ce qui se traduit par un accroissement (spectaculaire) du SDR de 14dB, et du SIR de 20dB. Au contraire, pour la source s_2 à large bande, le débit de ressource allouée entre la configuration SSI-C seule et la configuration SSI-CI passe de 145 à 118kbits/s pour encoder la bande de fréquence [0-16kHz], ce qui se traduit par une baisse de 4.8dB du SDR et 5.6dB du SIR entre ces deux configurations. Le codage de la source s_4 dans la configuration SSI-C est clairement sous-optimal, et un codage limité à la bande [0-5kHz], *i.e.* semblable à la stratégie adoptée dans les configurations SSI-CI amènerait vraissemblablement les performances de codage de la source s_4 au niveau de celles-ci. L'élément essentiel de l'approche SSI-CI est que, le sous-mélange traité par inversion en SSI-CI étant composé de seulement trois signaux sources au lieu de cinq, les scores de SDR et SIR obtenus pour s_1 et s_5 sont nettement supérieurs à ceux obtenus par la configuration SSI-I seule, en accord avec la modification de la répartition énergétique entre un mélange à 5 ou 3 sources vue à la Table 7.1 [9].

Enfin, l'intérêt d'une allocation différente de la ressource de codage est illustrée par la comparaison entre les configurations SSI-CI et SSI-CI opt. Le signal s_2 (batterie) ayant un contenu spectral beaucoup moins parcimonieux que celui du signal s_4 (guitare basse) se voit allouer une capacité de codage de 150kbits/s (au lieu de 118kbits/s) et par conséquent, la ressource de codage de l'estimée de la source s_4 passe elle de 118kbits à 86kbits. Cette meilleure allocation de la ressource de codage en fonction du contenu spectral de chaque signal se traduit par une hausse de 6.7dB du SDR et 3.4dB du SIR pour s_2, alors que les performances pour la source s_4 n'ont que faiblement diminuées. Ce meilleur équilibrage entre s_2 et s_4 profite même aux autres sources estimées par

8. *i.e.* de qualité audio similaire à celle des signaux sources originaux.
9. Pour la source s_3, l'amélioration est plus faible, mais ce point reste difficile à expliquer.

inversion, en raison d'une meilleure estimation du sous-mélange après soustraction des sources codées (cf Section 7.3.3) : entre SSI-CI et SSI-CI opt, s_1, s_3 et s_5 gagnent ainsi entre 0.6 et 1.8dB SDR et entre 0.7 et 3.6dB SIR. Même si cette seconde configuration SSI-CI opt. n'est pas encore pleinement optimale, elle fait apparaître l'importance d'une allocation adéquate de la ressource de codage.

Finalement, la configuration SSI-CI opt. apparaît comme la meilleure en moyenne : avec des performances de codage de s_2 et s_4 similaires à celles obtenues par SSI-C, et pour s_1, s_3 et s_5, de meilleures performances que par SSI-I [10]. Mentionnons enfin autre avantage essentiel de l'approche hybride : dans cette expérience, le débit total de tatouage en SSI-CI est quasiment divisé par deux par rapport à celui de SSI-C !

7.6 Évaluation perceptive

Afin de valider la méthode de séparation de sources informée, des tests audio formels sont présentés au cours de cette section. Étant donnée la lourdeur de la mise en place de tels tests, seule la configuration de SSI présentant les meilleures performances de séparation est testée, *i.e.* la SSI-CI présentée dans ce chapitre. Ces tests ont pour objectif de mesurer la qualité

- des signaux sources estimés individuellement,
- d'un sous-mélange du mélange initial obtenu en retranchant un des signaux sources estimés par SSI-CI au mélange disponible en entrée du décodeur (*i.e.* une application de type karaoké généralisé à chaque instrument).

7.6.1 Définition des tests perceptifs

Nous avons considéré deux types de tests, dénommés AX et ABX. Dans le test AX, le sujet doit juger de la qualité du signal dégradé A par rapport au signal de référence X en attribuant une note mesurant la dégradation de A sur une échelle à 5 niveaux présentée à la Table 7.3. Cette grille de mesure de la dégradation d'un signal audio par rapport à un signal de référence a initialement été établie par les experts de "l'International Telecommunication Union, Radiocommunication Bureau" ou ITU-R pour évaluer perceptuellement les performances de codeurs audio [ITU-R BS.1116, 1997] [ITU-R BS.562-3, 1990]. La mesure la plus couramment utilisée, notée SDG pour *Subjective Difference Grade* consiste en la différence entre la note (sur l'échelle à 5 niveaux de la Table 7.3) attribuée par le sujet au signal de référence et la note attribuée au signal dégradé.

$$\text{SDG} = \text{Note}_{\text{signal codé}} - \text{Note}_{\text{signal de référence}}$$

Une note de SDG de zéro indique que le codage est transparent, *i.e.* qu'il est impossible au sujet de faire la distinction entre le signal original et le signal dégradé. Une note de -4 indique au contraire une très forte dégradation.

10. La meilleure performance obtenue pour s_1 par SSI-C s'explique avant tout par la capacité de codage très élevée allouée à cette source.

DÉGRADATION ABSOLUE			SDG
	5,0	Imperceptible	0,0
	4,9 ⋮ 4,0	Perceptible mais pas gênante	-0,1 ⋮ -1,0
	3,9 ⋮ 3,0	Légèrement gênante	-1,1 ⋮ -2,0
	2,9 ⋮ 2,0	Gênante	-2,1 ⋮ -3,0
	1,9 ⋮ 1,0	Très gênante	-3,1 ⋮ -4,0

Table 7.3 – Échelle ITU-R de mesure subjective de la dégradation de la qualité audio.

Le test ABX correspond à un "test en double-aveugle avec triple stimulus et référence cachée" comme décrit dans [Bosi and Goldberg, 2003]. Ce test est communément admis comme étant un moyen particulièrement fiable et précis de mesurer de faibles dégradations dans les systèmes audio. Ce test permet de vérifier si le sujet est capable de distinguer un signal dégradé d'un signal original, de manière qualitative, et non quantitative. X correspond au signal de référence. Parmi la paire (A,B), l'un des signaux correspond au signal X. Le sujet écoute le signal X, puis les signaux A et B, présentés dans un ordre aléatoire. Il doit identifier lequel de A ou B correspond au signal de référence X. Trois réponses sont possibles :

- Le signal A est le signal de référence (*i.e.* le sujet pense que A=X)
- Le signal B est le signal de référence (*i.e.* le sujet pense que B=X)
- Le signal A ou B est le signal de référence (*i.e.* le sujet est incapable de distinguer une différence entre le signal dégradé et le signal X).

7.6.2 Tests comparatifs

Une des questions que l'on est en droit de se poser lorsque l'on considère pour la première fois le principe de la séparation de sources informée est : pourquoi ne pas transmettre par tatouage les pistes des signaux sources sous un format compressé au lieu de tenter d'estimer ces signaux à partir du signal de mélange[11] (soit par SSI-C, soit par SSI-I, soit par SSI-CI) ? De manière à comparer cette approche codage multi-pistes et la SSI, nous avons procédé à des tests d'écoute comparatifs entre les sources estimées par SSI-CI et une version compressée des sources originales correspondantes avec l'algorithme AAC [AAC, 2004], pour un débit d'information à transmettre au décodeur équivalent. En d'autres termes, la somme des débits des sources compressées

11. Notons qu'il ne s'agit alors plus de séparation de sources mais d'un codage multi-pistes.

en AAC correspond au débit de tatouage utilisé en SSI-CI.

Trois tests sont proposés pour chaque morceau (*i.e.* chaque mélange) :

- **test1 :** test de type AX, où X représente un signal source de référence, *i.e.* un signal source original, et A représente le signal dégradé correspondant, *i.e.* soit le signal compressé par AAC, soit le signal estimé par SSI-CI. Il s'agit donc d'un test d'écoute individuelle des sources.
- **test2 :** autre test AX, où cette fois X représente un sous-mélange original obtenu en soustrayant une source originale au mélange complet, et A représente le mélange correspondant dégradé obtenu en soustrayant au mélange complet un signal source dégradé, ce dernier correspondant soit au signal compressé AAC, soit au signal estimé par SSI-CI. Il s'agit d'un test d'écoute de type karaoké.
- **test3 :** test de type ABX où X représente un sous-mélange original, comme au test2. A et B sont deux sous-mélanges correspondant. Parmi A et B, l'un correspond à X, et l'autre est une version dégradée de X, obtenue de la même manière qu'au test2. Il s'agit donc d'un test d'écoute de type karaoké.

Par abus de langage, et par volonté de simplicité, la dénomination "sous-mélange ACC" désignera dans toute la suite le sous-mélange dégradé obtenu en soustrayant une source compressée par AAC au mélange complet ; la dénomination "sous-mélange SSI-CI" désignera elle le sous-mélange dégradé obtenu en soustrayant une source estimée par SSI-CI au mélange complet.

7.6.3 Conditions expérimentales

Les signaux utilisés pour les tests perceptifs sont extraits des signaux utilisés pour les tests informels menés aux Chapitres 6 et 7. Les signaux tests sont 3 mélanges de 5 sources d'une durée de 10 secondes chacun, de styles *new wave*, *rock* et *pop-rock*. Les signaux estimés par la méthode SSI-CI sont ceux correspondant aux résultats de la configuration SSI-CI opt. présentés à la Section 7.5.3, avec un débit total de codage de 150kbits/s/voie, soit un total de 300kbits/s pour le mélange stéréo. Les signaux sources mono sont compressés au format AAC avec les débits suivant :

- s_1=guitare/piano : 64kbit/s
- s_2=batterie : 96kbits/s
- s_3=voix : 64kbits/s
- s_4=guitare basse : 32kbits/s
- s_5=choeurs/synthétiseur : 48kbits/s

soit un total de 304kbits/s pour l'ensemble des 5 sources composant le mélange, similaire à celui utilisé dans la configuration SSI-CI opt.

Les tests ont été menés en chambre anéchoïque sur un PC disposant d'une carte son M-AUDIO Delta 44, et d'un casque HI-FI Sennheiser eH 250. L'interface graphique a été implémentée sous Matlab. Le test a été proposé à 15 sujets, musiciens et non musiciens, certains profanes en terme de tests d'évaluation de qualité audio, et d'autres plus familiers avec ce type de tests. Pour chaque chanson, chaque sujet est amené à

écouter un total de 20 extraits pour les tests 1 et 2, et 30 extraits pour le test 3, soit un total de 70 extraits de 10 secondes. La durée de ces trois tests étant relativement élevée, il a été choisi de ne faire faire le test à chacun des 15 sujets que sur 2 des 3 chansons (avec une répartition de 10 sujets par chanson).

7.6.4 Résultats et interprétations

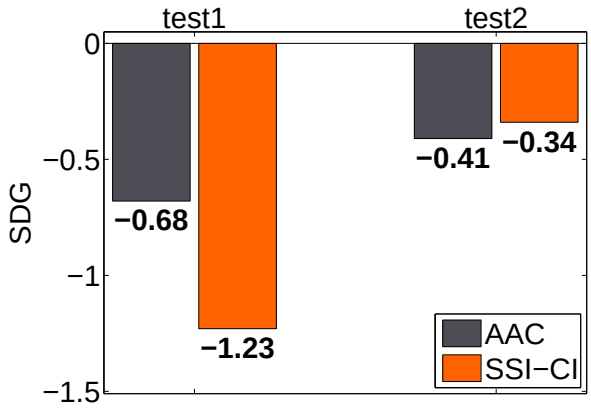
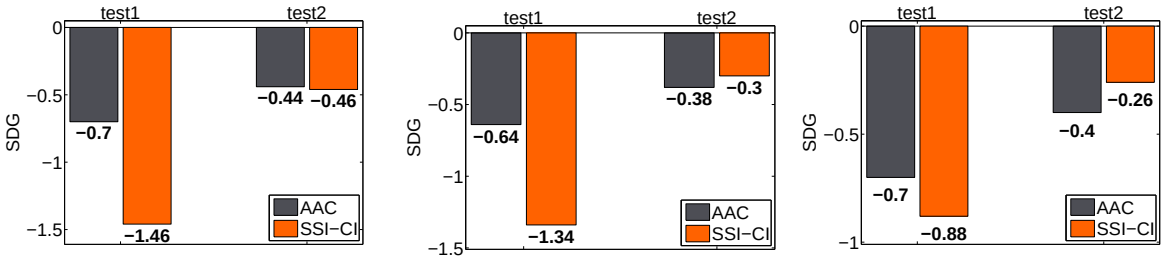


Figure 7.3 – Notes moyennes de SDG obtenues pour les deux méthodes AAC et SSI-CI, pour les tests 1 et 2.



(a) Chanson 1 - new wave (b) Chanson 2 - rock (c) Chanson 3 - pop rock

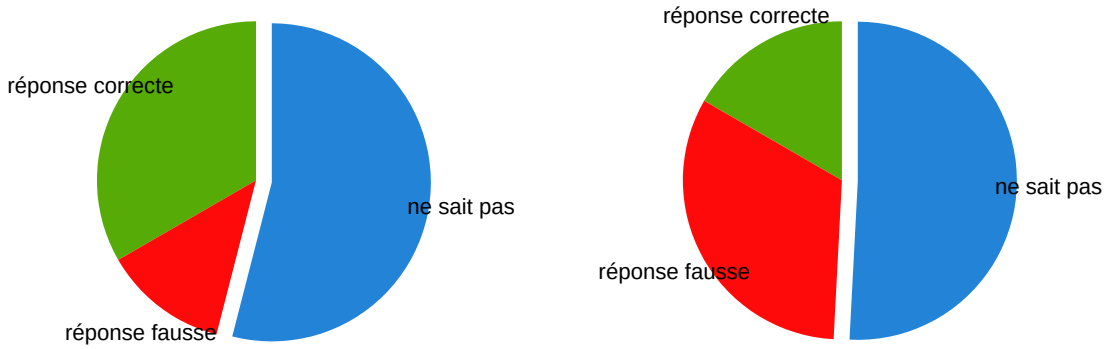
Figure 7.4 – Détails par chanson des notes de SDG obtenues pour les deux méthodes AAC et SSI-CI, pour les tests 1 et 2.

Les résultats des tests 1 et 2 moyennés sur les 3 chansons sont présentés Figure 7.3, et les résultats par chanson sont présentés Figure 7.4. Concernant le test1, *i.e.* l'écoute des pistes séparées, les résultats de compression apparaissent (naturellement) supérieurs à ceux des signaux estimés par SSI-CI. Le SDG moyen de -0.68 obtenu pour les signaux compressés AAC traduit une dégradation perceptible mais pas gênante par rapport aux signaux originaux. Un SDG moyen de -1.23 pour les signaux estimés par SSI-CI indique une dégradation perceptible, et très légèrement gênante. Notons que les dégradations des signaux estimés par SSI-CI ont particulièrement été perçues par les sujets dans les portions de silence des signaux source, où le bruit, soit de codage, soit musical en fonction que la source ait été estimée par codage (SSI-C) ou par

inversion du mélange (SSI-I), n'est pas masqué par le signal. Des tendances équivalentes transparaissent pour chacune des chansons testées, comme le montre la Figure 7.4. Les écarts de performances SSI-CI/AAC du test1 sont principalement dus aux portions de silence, de durée variable d'un morceau à l'autre, durant lesquelles les artefacts de la méthode SSI-CI sont particulièrement mis en évidence.

Concernant le test2, la méthode SSI-CI apparaît au moins aussi performante que la compression AAC, avec un SDG moyen de -0.34. La dégradation du sous-mélange SSI-CI est donc quasiment imperceptible, et du même ordre de grandeur que celle obtenue pour la compression AAC. Les imperfections des signaux sources individuels estimés par SSI-CI qui ont été mis en évidence au test1 sont masquées au sein du sous-mélange SSI-CI. Ces résultats confirment les tests informels de la Section 7.5.3, et ils confirment l'intérêt de la méthode SSI-CI pour des applications de type remixage/karaoké.

Les résultats du test3 sont présentés à la Figure 7.5, pour les méthodes d'estimation des sources, AAC et SSI-CI. Il apparaît Figure 7.5 que les sujets n'ont pas été capables de discerner le sous-mélange dégradé du sous-mélange de référence, dans 54% des cas pour le sous-mélange AAC, et 51% pour le sous-mélange SSI-CI ce qui atteste de la qualité du remix obtenu en soustrayant au mélange complet une source estimée par SSI-CI. Les scores équivalents entre les deux méthodes et proches de 50% des sujets ne discernant pas le sous-mélange de référence démontrent la qualité des sous-mélanges dégradés. Parmi les sujets pensant distinguer la référence X parmi le couple A/B, 33% ont répondu correctement pour le sous-mélange AAC et 17% pour le sous-mélange SSI-CI, alors que 13% se sont trompés pour le sous-mélange AAC et 33% pour le sous-mélange SSI-CI. Cette distinction entre les scores des sujets pensant identifier correctement le mélange de référence laisse à penser qu'un artefact permet de plus facilement identifier le sous-mélange AAC. Les résultats du test3 confortent ceux obtenus au test2 et confirment que la qualité du sous-mélange SSI-CI est indiscernable de celle du sous-mélange AAC et surtout du sous-mélange original (seulement 17% d'identification correcte de la référence lors du test référence/SSI-CI).



(a) AAC (b) SSI-CI

FIGURE 7.5 – Détails par technique du test 3 de type ABX.

7.7 Conclusion

La méthode hybride de séparation de sources informée présentée dans ce chapitre combine à la fois les avantages d'une approche de SSI basée sur la parcimonie des sources, et ceux d'une approche par codage des sources à l'aide de descripteurs TF adaptés. Le processus de séparation par SSI-I, relativement simple, est ici amélioré par l'utilisation conjointe de l'approche par codage SSI-C, dont l'efficacité a déjà été observée au Chapitre 5. La réduction de la dimension du mélange original en codant un sous-ensemble des signaux sources a pour effet de limiter la superposition des sources dans le plan TF. Des performances de séparation globales significativement supérieures à celles obtenues par les méthodes SSI-C ou SSI-I considérées séparément sont alors obtenues, même pour un mélange ayant un nombre relativement limité de sources [12] (ici 5). Ce système apparaît comme un compromis satisfaisant entre l'approche par codage, efficace pour un nombre raisonnable de sources, mais relativement lourde computationnellement, et l'approche basée sur la parcimonie des sources, beaucoup plus légère computationnellement parlant, mais limitée en cas de forte superposition des signaux sources au sein du plan TF. Le caractère *informé* de cette technique qui combine deux types d'informations complémentaires sur les sources permet donc d'atteindre des performances de séparation très largement supérieures à celles atteignables par des méthodes de séparation aveugles. La qualité audio des signaux obtenus avec l'approche SSI-CI semble tout à fait adaptée à l'application de type remix/karaoké, mais ne se limite pas à celle-ci. En effet, des tests perceptifs formels ont montré que la qualité des sources séparées est globalement proche de celle obtenue avec une compression AAC de chaque source, avec des débits certes faibles, mais fort convenables pour une technique de compression aussi performante que l'AAC et pour des sources doublement monophoniques (une source sonore et une voie). Le débit de codage AAC reste ici convenable en raison du nombre limité de sources dans nos expérimentations. On peut penser que pour un plus grand nombre de sources, les performances de la méthode SSI-CI seraient affectées, proportionnellement, moins significativement que celles de la méthode par compression AAC. Pour un mélange constitué de 7 signaux sources par exemple, 2 à 3 pourraient être encodés par SSI-C et les 4 ou 5 autres estimés par SSI-I avec des résultats proches de ceux obtenus au Chapitre 6. En revanche, deux sources supplémentaires à encoder en AAC impliqueraient une réduction significative du débit de codage pour les autres sources, et la probable apparition d'artefacts.

12. Rappelons que l'approche hybride est avant tout destinée aux mélanges constitués d'un large nombre de sources.

Conclusion

Rappel de la problématique

Ce travail de thèse s'inscrit, de façon originale, dans la thématique de la séparation de sources. La principale application visée par cette approche est l'écoute active de la musique. L'objectif de cette thèse était la réalisation d'un système complet [13] de séparation de sources audio (principalement musicales) permettant à un utilisateur "fournisseur" de mettre à disposition d'un utilisateur "client" un mélange musical (typiquement stéréophonique) au contenu augmenté par tatouage numérique, afin que l'utilisateur client puisse séparer les différents instruments/voix composant le mélange, lors de sa restitution. L'écoute active de la musique ici ciblée, peut inclure tous types de post traitements sur les signaux sources séparés (application de type karaoké étendue aux instruments, modification de la spatialisation des sources, etc...). Le problème posé est donc celui de la séparation de haute qualité d'un grand nombre de sources musicales au sein d'un mélange de faible dimension (une ou deux observations du mélange selon que le mélange considéré est monophonique ou stéréophonique), *i.e.* en configuration sous-déterminée. Pour mener à bien les objectifs de séparation particulièrement contraignant que nous nous sommes fixés, une configration originale est adoptée, dans laquelle les signaux sources sont disponibles avant mélange, *i.e.* durant l'étape de production musicale. Cette approche est basée sur la combinaison de deux domaines du traitement du signal jusqu'alors disjoints, la séparation de source, et le tatouage audio-numérique.

Contributions principales et résultats obtenus

Dans toute la thèse, nous nous sommes placés dans la configuration de mélanges linéaires instantanés stationnaires sous-déterminés (monophoniques et stéréophoniques). Différentes approches ont été considérées au cours de cette thèse :

- une première approche dédiée aux mélanges monophoniques orientée vers un codage vectoriel des signaux sources,
- une deuxième approche exploitant la parcimonie des sources dans le plan TF et la dimension stéréophonique du mélange,
- enfin une approche hybride permettant une exploitation conjointe des atouts de

13. C'est à dire aussi bien la partie codeur correspondant à l'utilisateur "fournisseur" que la partie décodeur correspondant à l'utilisateur "client".

chacune des précédentes méthodes.

La première approche considérée au Chapitre 5 de cette thèse, nommée SSI-C, a initialement été mise au point pour une configuration sous-déterminée extrême (une seule observation du mélange). Le choix a été fait de coder la contribution des sources dans le mélange. Des descripteurs des signaux sources ont alors été définis, de manière à discriminer le plus possible les différentes sources relativement à leur participation au sein du mélange (descripteur de gain), voire de décrire chaque signal source indépendamment du mélange (descripteurs moyenne-gain-forme). Parallèlement, un premier système de tatouage basé sur la technique QIM a été implémenté, permettant un codage de qualité de deux à quatre sources composant un mélange LISM. Cette première implémentation a avant tout permis de mettre en évidence la faisabilité de la technique de SSI basée sur le codage des signaux sources, en validant le choix des descripteurs utilisés, la contrainte sur la précision de leur codage, mais aussi l'efficacité du système de tatouage utilisé (inaudibilité + qualité du décodage). Il a aussi pu être mis en évidence l'apport de la quantification vectorielle, et la nécessité d'avoir recours à une description des signaux sources individuellement, indépendamment du mélange (*i.e.* autre que le descripteur de gain). Cette première série de travaux a donné lieu aux publications suivantes :

- **M. Parvaix**, L. Girin, and J.-M. Brossier, "A watermarking-based method for single-channel audio source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, 2009, pp.101-104,
- **M. Parvaix**, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," in *IEEE Trans. Audio., Speech, and Language Process.*, 18 (6), 1464-1475, 2010.

Le codage de l'information de forme d'une molécule source a prouvé qu'il permet de s'affranchir de l'influence parasite des autres sources sur le mélange. Néanmoins, des artefacts persistaient sur les signaux sources estimés, principalement dus à la superposition des signaux sources dans les portions du plan TF où l'estimation des signaux sources est faite grâce au signal de mélange. Deux pistes de recherche ont alors été menées de front. Tout d'abord le développement d'une technique de tatouage plus performante utilisant un MPA (réalisée par Jonathan Pinel au cours de son stage de Master puis de sa thèse au GIPSA-lab) a permis d'augmenter significativement la capacité de tatouage et d'améliorer ainsi les performances de séparation par une utilisation plus systématique du descripteur de forme. Ensuite, nous nous sommes orientés vers une approche moins coûteuse en ressources de tatouage que l'approche de SSI par codage.

Introduite au Chapitre 5 sous l'appellation de SSI-I, une nouvelle approche de SSI a ainsi été développée, plus classique du point de vue de la séparation de sources. L'hypothèse de Window Disjoint Orthogonality [14] communément faite en séparation de sources dans la configuration sous-déterminée [Yilmaz and Rickard, 2004], étant ap-

14. *i.e.* l'hypothèse qu'une seule source a une énergie significative dans le mélange à un instant et une fréquence donnée,

paru insatisfaisante pour des mélanges de signaux de musique, une nouvelle approche informée a été développée, permettant de considérer autant de sources prédominantes à un bin TF donné qu'il y a d'observations du mélange, suivant en cela la logique de [Bofill and Zibulevski, 2001]. Le caractère informé de cette nouvelle approche basée sur l'indexation de sources prédominantes a permis de lever les problèmes critiques en séparation aveugle de sources de détermination des sources prédominantes, et d'estimation de la matrice de mélange. Cette nouvelle approche, beaucoup moins coûteuse en meta-information à transmettre au décodeur, couplée à l'augmentation de la ressource de tatouage, a également permis de réduire la résolution de traitement, qui est apparu comme un facteur essentiel en SSI-I [15]. L'exploitation combinée de l'hypothèse de parcimonie des sources, et de la multi-dimensionnalité des mélanges considérés (stéréophoniques dans le cadre de l'application musicale visée), a permis le développement d'une méthode de séparation de sources efficace et peu coûteuse en ressources de tatouage en comparaison de l'approche SSI-C. Des performances de séparation quasi identiques à celles de l'estimateur Oracle introduit par Vincent et al. [Vincent et al., 2007] attestent de la qualité des résultats obtenus. Cette deuxième série de travaux a donné lieu aux publications suivantes (la troisième est seulement dans une première phase de review) :

- **M. Parvaix**, and L. Girin, "Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'10)*, Dallas, Texas, 2010,
- **M. Parvaix**, and L. Girin, "Séparation de sources informée pour des mélanges stéréo instantanés utilisant un tatouage de l'index des sources localement prédominantes," in *Actes du 10ème Congrès Français d'Acoustique*, Lyon, France, 2010,
- **M. Parvaix**, and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," in *IEEE Trans. Audio., Speech, and Language Process.*, *soumis*.

Plus encore que pour l'approche par codage des signaux sources, il est apparu en SSI-I que la superposition dans le plan TF de sources ayant des signatures spectro-temporelles similaires est un facteur limitant majeur. C'est pourquoi il a été introduit au Chapitre 7 une nouvelle variante de la SSI, nommée SSI-CI, qui mêle les approches de SSI par codage et de SSI par indexation des sources prédominantes. Cette approche est actuellement la plus aboutie et la plus performante tant au niveau du processus de séparation à proprement parlé qu'au niveau de la technique de tatouage. La SSI-CIa pour but d'offrir des performances de séparation supérieures à celle des méthodes SSI-C et SSI-I prises séparément en combinant les atouts de chacune de ces deux méthodes. De la méthode par codage SSI-C, la SSI-CI utilise la qualité du codage vectoriel (permis par une capacité d'insertion de l'information importante), et de la

15. Il s'agit d'une autre piste à exploiter dans les travaux futurs, par l'utilisation d'une transformée TF multi-résolution, à échelle fréquentielle logarithmique par exemple.

SSI-I la SSI-CI reprend la simplicité de mise en oeuvre et l'information de tatouage limitée. Cette configuration a permis d'améliorer les performances de séparation sur des mélanges de test à 5 sources. Cette méthode hybride a le potentiel pour traiter un nombre de sources plus important même si cela reste à confirmer par des tests supplémentaires.

En ce qui concerne les publications :

- L'approche hybride sera soumise à publication prochainement.
- Même si elle est relativement restreinte, ma participation au développement du système de tatouage de J. Pinel a donné lieu à la publication suivante :
  J. Pinel, L. Girin, C. Baras, and **M. Parvaix**, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Int. Congress on Acoustics (ICA'10)*, Sydney, Australia, 2010.
- Pour finir l'aspect publication, je souhaite mentionner la publication de mon travail de Master en liaison avec la problématique du Molecular Matching Pursuit décrite notamment dans l'Annexe A :
  M. Parvaix, S. Krishnan, and C. Ioana, "An audio watermarking method based on Molecular Matching Pursuit," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, Nevada, 2008.

Une première série de valorisations : demande de dépôt de brevet et implémentation d'un premier démonstrateur

Ce travail de thèse, qui présente à la fois une démarche méthodologique nouvelle basée sur des outils théoriques et un caractère appliqué, possède un fort potentiel de valorisation industrielle à plus ou moins court-terme. Ainsi une demande de dépôt de brevet a été effectuée auprès de l'INPI en liaison avec le service valorisation de l'établissement :

M. Parvaix, L. Girin, J.-M. Brossier, et S. Marchand, "Procédé et dispositif de formation d'un signal mixé, procédé et dispositif de séparation de signaux, et signal correspondant", Brevet, Numéro de dépôt 09 52397, 2007, France.

Par ailleurs, un premier démonstrateur temps-réel a été réalisé pour démontrer la faisabilité des applications cibles de karaoké et remixage. Ce démonstrateur a été réalisé en collaboration avec le LaBRI sur la base de la méthode SSI-I (telle qu'elle était développée à GIPSA-lab fin 2009) [16]. Il prend la forme d'un logiciel, fonctionnant sur Mac, permettant de charger un fichier wav stéréo contenant le mix actif [17], et

16. Le LaBRI a réalisé la traduction du code Matlab de l'algorithme du décodeur en langage C/C++, la gestion du flux audio, et la réalisation de l'interface graphique.

17. Ce fichier de mix est généré au préalable avec l'algorithme de l'encodeur sous Matlab (il n'existe pas encore de démonstrateur autonome d'encodeur).

de l'écouter en manipulant les sources par l'intermédiaire d'une interface graphique étonnante de simplicité (voir Figure 7.6) : les sources peuvent être individuellement déplacées autour de l'auditeur par cliquer/glisser de l'icône correspondante, et on peut ainsi contrôler séparément le volume et le panning gauche/droit de chaque source. Des mix actifs de chansons connues (ou moins connues) ont été réalisés à partir de pistes studio pour illustrer l'intérêt de la technologie sur des morceaux de musique s'approchant des mix professionnels. A titre d'exemple, des configurations chant + basse + batterie + guitare + choeurs ou chant + basse + batterie + synthétiseur + cuivres ont été testées.

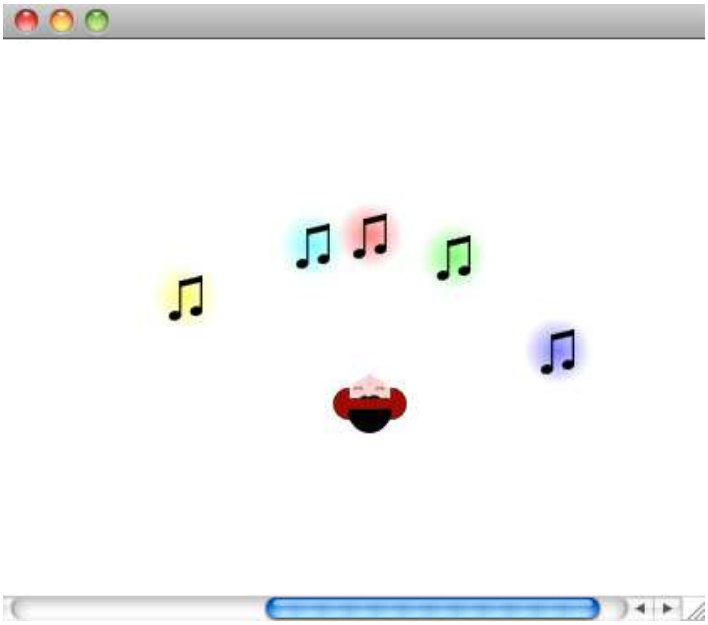


FIGURE 7.6 – Interface graphique utilisateur du démonstrateur de SSI par indexation des sources (approche introduite au Chapitre 6). Version 1.0.

Perspectives

Les perspectives offertes par ce travail sont très nombreuses. Nous donnons ci-dessous quelques exemples de pistes scientifiques à développer à court et moyen terme.

Optimisation de la parcimonie

La superposition des signaux sources dans le domaine transformé (dans notre cas le domaine TF) est apparu comme le principal facteur limitant en séparation de sources, principalement pour les méthodes basées sur la parcimonie. Aussi une étude poussée sur différents types de transformées, autre que la MDCT pourrait être envisagée, et la superposition des sources évaluée expérimentalement pour des signaux de musique, ainsi que la robustesse des ces transformées au processus de tatouage utilisé en SSI. La Cosine Packet Transform étudiée par Nesbit et al. dans [Nesbit et al., 2006] et [Nesbit

et al., 2007] dans le cadre de la séparation de sources par masquage binaire dans le plan TF pourrait être envisagée, de même que la Constant Q Transform [Brown, 1991] qui offre l'avantage d'une échelle fréquentielle logarithmique permettant une bonne résolution fréquentielle en basses fréquence, et une meilleure résolution temporelle en hautes fréquences. Une comparaison de plusieurs transformées TF est fournie dans [Tan and Févotte, 2005] dans le cadre de la séparation aveugle de sources en configuration sous-déterminée par des méthodes d'estimation Bayesienne s'appuyant sur la parcimonie des sources dans le domaine transformé (hypothèse de distribution de Student-t des signaux sources). Les transformées MDCT et DWT (Discrete Wavelet Transform) y sont plus précisément étudiées. Une étude récente [Suresh and Sreenivas, 2009] sur la propriété de convolution/multiplication de la MDCT lors du passage du domaine temporel au domaine fréquentiel pourra également servir de base à une extension de la SSI à des mélanges plus complexes.

Développement de méthodes de séparation informée prenant mieux en compte les procédés de l'industrie musicale

Cette phase concerne à la fois des aspects très théoriques et des aspects très appliqués. Le mélange linéaire instantané stationnaire (i.e. par matrices à coefficients réels constants) traité jusqu'ici (avec un certain succès) est un procédé de mélange relativement simpliste, qui, même s'il fournit des mix de qualité très acceptable pour le profane, ne correspond que partiellement à la complexité des techniques employées dans la production musicale professionnelle [18]. Des phénomènes plus ou moins complexes de réverbérations entrent généralement en jeu (soit par réverbération naturelle de la chambre d'enregistrement, soit par simulation numérique post-enregistrement), avec des caractéristiques généralement différentes sur les deux voies : on rentre alors dans le cadre de mélanges convolutifs (voir la Section 2.2.3), plus difficiles à traiter [Pedersen et al., 2007] [Parra and Spence, 2000] [Melia and Rickard, 2007]. Le cas de mélanges convolutifs dans la configuration sous-déterminée a été beaucoup moins étudié que dans la configuration sur-déterminée, mais on peut néanmoins se référer à [Aïssa-El-Bey et al., 2007] [Blin et al., 2004] [Bofill and Monte, 2006] [Olsson and Hansen, 2006] [Peterson and Kadambe, 2003] [Winter et al., 2007] pour des exemples de traitements de ce cas de figure. On peut avoir aussi à traiter des sources dont le mixage ne peut pas être modélisé par de tels procédés [19], et qui doivent être considérées comme des sources "true stereo". Enfin, la phase de post-production du mix (appelée masterisation dans le langage spécialisé) implique des traitements non-linéaires tels que

18. La source mixée par un tel procédé est supposée être une source ponctuelle qui arrive sur les deux voies (gauche et droite) sans retard, ni réverbération. Un effet stéréophonique significatif est toutefois réalisé en fixant des coefficients différents pour les deux voies. Ce procédé peut suffire à représenter correctement des sources qui traditionnellement ne nécessitent pas un effet stéréo très important pour bien sonner (typiquement la voix du chant principal, généralement située au centre de la scène sonore, et la guitare basse / contrebasse, peu valorisée par des effets stéréo complexes).

19. Cela peut être le cas d'instruments de grande dimension, comme le piano ou la batterie, qui ne peuvent pas être considérés comme une seule source ponctuelle.

la compression dynamique susceptible de perturber significativement la séparation de sources en remettant en cause l'hypothèse de linéarité des mélanges. Ce projet nécessite donc une phase de développements techniques abordant ces points dont l'objectif est de rendre la technologie proposée la plus compatible possible avec les techniques de production musicale professionnelles, en vue d'améliorer la qualité des mix actifs, et de satisfaire la communauté (très exigeante) des artistes et des professionnels en charge du mixage. Comme dans la séparation de sources classique, le traitement convolutif sera vraisemblablement abordé dans le domaine temps-fréquence (par transformée de Fourier à court terme) où le mélange convolutif s'apparente alors à un mélange instantané (pour chaque canal fréquentiel). Il s'agira de généraliser le problème de l'inversion guidée par l'information de tatouage à ce cas de figure, en intégrant les nouvelles difficultés liées à la complexification du problème [20]. Le cas des sources "true stereo" pourra être abordé avec l'approche par codage (SSI-C) mentionnée en Section 4. L'utilisation d'un codage universel (type AAC par exemple) sera implémenté et testé pour remplacer le codage vectoriel, propriétaire, utilisé dans cette thèse, pour s'affranchir des difficultés de l'apprentissage des dictionnaires de prototypes de descripteurs, et bénéficier des raffinements existants dans ce domaine (notamment le codage très efficace de la stéréo dans AAC). Dans l'idéal, cette approche codage universel des sources pourrait être appliquée à tout type de source du moment qu'on est capable de contrôler sa contribution au mélange (*i.e.* si on somme des sources préalablement "stéréophonisées" par des procédés instantanés, convolutifs, ou "true stereo".) Cependant la capacité de tatouage du mix reste limitée et ne permet pas de tatouer le résultat de l'encodage de toutes les sources, au-delà d'un certain nombre raisonnable. La stratégie sera alors le développement d'un système hybride, similaire à celui développé au Chapitre 7, dans lequel les sources les plus complexes sont encodées, et les sources instantanées et convolutives sont séparées par des procédés basés sur la parcimonie et l'inversion, après soustraction des sources encodées, combinant ainsi les avantages de chaque approche. Le problème des traitements non-linéaires reste quand à lui ouvert.

Concernant les méthodes de séparation de sources visant plus spécifiquement les mélanges convolutifs de sources musicales, on peut citer par exemple les travaux de Viste et al. qui introduisent dans [Viste and Evangelista, 2006] une méthode de séparation de sources dédiée aux applications musicales. En groupant par instrument les zones du plan TF situées autour des partiels (un ou plusieurs partiels de chaque note de chaque instrument ayant pu être isolé(s)) cette technique offre une solution dans le cas où plusieurs instruments jouent en harmonie et où des partiels (des notes) de ces différents instruments se superposent dans le plan TF [21], rendant les méthodes statistiques de séparation de sources inefficaces.

20. Par exemple : compatibilité de la taille des filtres de mélange et de démélange en regard de la taille de la transformée, diminution de la parcimonie des sources par la réverbération, augmentation de la taille de tatouage dédiée au codage du procédé de mélange, etc.

21. Si les signaux sont relativement stationnaires, les partiels des différents instruments sont alors corrélés

Développement d'un système de SSI pour mix au format compressé

Le format compressé occupe une place de plus en plus importante sur le marché de la musique (plates-formes de téléchargement, lecteurs multimédia portables, stockage massif de musique sur disque dur, etc.) Il apparaît donc nécessaire d'envisager l'extension de la technologie vers des signaux de mélange compressés (ou du moins une première extension montrant la faisabilité du problème). Ici la tâche se complique car d'une part, par définition, le format compressé ne permet pas une insertion d'information de séparation importante (puisqu'on a déjà réduit la représentation du signal à une forme très compacte), et d'autre part la compression peut modifier des caractéristiques du signal mixé d'une façon susceptible de perturber le processus de séparation (effectué à partir de ce mélange). Ce problème de compression influe donc à la fois sur le tatouage et sur le procédé de séparation. Toutefois, des études très préliminaires réalisées à GIPSA-lab ont montré la possibilité d'étendre la séparation de sources aux signaux compressés, au moins pour des mélanges linéaires instantanés (ceux traités dans cette thèse), avec une détérioration quasiment négligeable des performances de séparation lorsque les débits de compression ne sont pas trop élevés (ce qui est de moins en moins une contrainte compte tenu de la bande passante des réseaux actuels). Ces études doivent être confirmées, et l'implantation d'un système complet intégrant le flux de tatouage dans le format compressé [22] doit être réalisé. Au-delà, la combinaison de l'extension des configurations de mélanges (comme vue à la section précédente) avec la compression du signal mixé reste un problème ouvert, à la fois difficile et passionnant.

22. Cela pourra se faire soit par insertion directe sur le signal de mélange comme il est réalisé dans cette thèse, ou bien, tout du moins dans un premier temps, en exploitant les zones prévues pour l'intégration de metadata dans les flux compressés.

Annexes

Annexe A

Molecular Matching Pursuit

A.1 Principe et Algorithme

Dans le cadre de la configuration SSI-C, une possibilité autre qu'un découpage régulier du plan TF en molécules adjacentes introduit à la Section 5.1.2 existe. Il s'agit d'utiliser un algorithme de décomposition itératif inspiré des algorithmes de type Matching Pursuit (MP) [Mallat and Zhang, 1993] et adapté au cadre moléculaire. Le Matching Pursuit repose sur une idée simple : étant donné un dictionnaire $\mathcal{D} = \{c_\lambda\}_{\lambda \in \Lambda}$ de signaux de référence et un signal x, cet algorithme consiste à reconstruire le signal x de façon itérative à partir de sa projection sur $\mathcal{D}$. Dans notre cas, il s'agirait de la base de coefficients MDCT. Lors de la première itération de l'algorithme classique de MP, on recherche la projection maximale de s sur $\mathcal{D}$. Cette projection est soustraite au signal initial pour obtenir le premier résidu. Lors de la seconde itération, l'algorithme est appliqué au résidu, et ainsi de suite, jusqu'à un critère d'arrêt qui peut être, par exemple, un seuil sur l'amplitude de la projection maximale. Cet algorithme peut être vu comme un cas particulier de la technique de Projection Pursuit issue de l'analyse statistique [Friedman and Tukey, 1974].

L'algorithme du Molecular Matching Pursuit (MMP) est introduit par Daudet [Daudet, 2006] puis utilisé sous une version modifiée dans [Parvaix et al., 2008]. La différence majeure entre l'algorithme classique du MP et le MMP est que ce dernier prend en compte la structure même du signal en retranchant au résidu à chaque itération tout un groupe de projections ou *atomes* ayant des propriétés semblables. Dans notre cas, il s'agira à nouveau d'une *molécule* composée d'atomes localisés dans le même voisinage temps-fréquence. La MDCT étant une transformation parcimonieuse, l'essentiel de l'énergie du signal est concentrée dans un petit nombre de coefficients MDCT. En ce qui nous concerne, ceci permet de dire que le signal x peut être approximé par un faible nombre de molécules en regard des dimensions du plan TF entier. Ces molécules sont bien localisées aux endroits de forte énergie du signal. On peut donc écrire l'approximation suivante en reprenant les notations du paragraphe précédent :

$$x = \sum_{(p,q) \in \{P \times Q\}} M_{pq}^x + R \tag{A.1}$$

où P et Q sont deux ensembles d'indices tels que $card(P)$ est petit devant $W/2F$ et $card(Q)$ est petit devant $\left(\frac{2N}{W}+1\right)/T$, et R est le résidu de la décomposition, faible devant la projection.

L'algorithme de MMP en tant que tel peut alors se mettre sous la forme suivante :

1. Calculer la décomposition temps-fréquence en coefficients MDCT de x
2. Sélectionner la molécule $M_{pq}^{x} = \left\{m_{ij}^{x}\right\}_{(i,j)\in\{I\times J\}}$ de coefficients MDCT centrée autour du coefficient MDCT m_{i_1,j_1} de plus grande énergie dont les coordonnées dans le plan temps-fréquence sont (i_1, j_1). Alors les intervalles I et J sont définis par $I = \{i, |i - i_1| \leq E[T/2]\}$ et $J = \{j, |j - j_1| \leq E[F/2]\}$, T et F étant les dimensions en temps et fréquence de la molécule.
3. Soustraire cette molécule au plan temps-fréquence
4. Retourner à l'étape 2 tant que $|m_{i_1,j_1}| > \varepsilon_{seuil}$

L'algorithme de MMP peut offrir un avantage certain par rapport à un simple découpage régulier du pavage temps-fréquence dans la mesure où il permet de "classer" les molécules en fonction de leur énergie. Dans une région spectrale donnée, cette énergie est généralement corrélée à la contribution de la molécule à la qualité audio des signaux. Même si des molécules de faible énergie ont leur importance, notamment en hautes fréquences, pour affiner la qualité du signal, en général, plus l'énergie d'une molécule est conséquente, plus son apport sur le rendu auditif du signal est grand. D'une manière générale, il n'est donc pas nécessaire pour reconstruire un signal de bonne qualité d'écoute d'utiliser l'ensemble des coefficients MDCT de son pavage temps fréquence. Les coefficients de très faible amplitude (et a fortiori les coefficients nuls) ne sont pas indispensables, et sont naturellement éliminés par le MMP. Le nombre de molécules traité si l'on utilise l'algorithme du MMP est donc en général significativement inférieur à celui d'un pavage régulier temps-fréquence. Les molécules de trop faible amplitude, généralement localisées dans les hautes fréquences, sont ainsi jugées inutiles (sur un critère à fixer) pour la séparation de sources. En d'autres termes, le caractère parcimonieux de la décomposition par MMP est donc supposé être cohérent avec notre problème de séparation.

A.2 Quantification et MMP

En plus de fournir un niveau de référence pour le tatouage par le quantificateur Q_2, le quantificateur Q_1 introduit à la Section 5.1.3.1 a un autre rôle fondamental dans le cas de l'utilisation du Matching Pursuit : il permet alors d'assurer que les molécules obtenues au décodeur par le MMP sont les mêmes que celles obtenues au codeur. Ceci est très important : si les molécules constituant le support élémentaire du tatouage ne sont pas exactement similaires au codeur et au décodeur, il est alors impossible de recouvrer correctement le tatouage. Le choix des résolutions R_1 et R_2 doit être tel que

la valeur d'un coefficient MDCT sur la grille Q_1 ne soit pas modifiée, ni par la phase de tatouage (qui place le coefficient en question sur la grille R_2) ni par aucune autre contrainte que puisse être amené à subir le signal tatoué. Par conséquent, le groupement de coefficients en molécules par MMP fournit le même résultat après les blocs 3 et 10 de la Figure 5.1 respectivement au codeur et au décodeur. Ceci ne serait pas assuré par l'application directe du MMP sur les coefficients MDCT quantifiés par Q_2 après tatouage (l'ordre relatif des amplitudes des coefficients MDCT peut être modifié par le tatouage). Notons que c'est cette condition qui explique l'ordre des blocs 2 et 3 dans le traitement. Encore une fois, cette correspondance exacte entre les molécules au codeur et au décodeur est indispensable à la bonne lecture de la watermark dont le support élémentaire est une molécule.

A.3 Descripteur de gain et groupement moléculaire par MMP

Dans ce paragraphe, nous donnons quelques remarques sur la séparation à partir du descripteur de gain (cf Section 5.1.4.1) dans le cas du groupement moléculaire par MMP (au lieu d'un découpage régulier du plan temps-fréquence), même si cette configuration n'a pas été implémentée. Une étape supplémentaire au codeur est dans ce cas nécessaire par rapport au cas du pavage régulier. En effet, les molécules obtenues après décomposition, que ce soit pour les sources ou pour le mélange, ne sont pas localisées de façon régulière dans le plan temps-fréquence et n'ont pas forcément le même nombre de coefficients suivant le critère de génération d'une molécule par MMP. Une molécule de mélange ne correspond donc pas obligatoirement à une molécule de source exactement au même endroit du plan TF (voir figure 4.4). Il faudrait donc dans un premier temps estimer le "degré de superposition" des molécules sources par rapport à celle du mélange. On peut déterminer par une mesure de corrélation si les molécules des signaux sources et celles du mélange sont ou non superposées dans le plan TF. Un classement est effectué pour chaque molécule de chaque source en deux catégories, superposée ou disjointe avec une molécule de mélange, en fonction de la valeur de la corrélation entre cette molécule et la molécule de mélange située dans le même voisinage TF à laquelle elle est comparée[1]. Dans le cas où des molécules de plusieurs sources sont superposées à la même molécule de mélange, un calcul de la contribution énergétique relative locale de chaque source par rapport au mélange peut être réalisé selon le même principe qu'à la formule (5.15) avec des molécules non-rectangulaires. Cette information peut être par la suite exploitée pour la séparation de façon analogue à celle décrite à l'équation (5.19). La Figure A.1 donne un aperçu, toujours schématique, mais plus précis que celui de la Figure 4.4, de la reconstruction de l'estimée $\hat{S}_i$ de S_i dans le cas d'un mélange de deux sources.

1. Une double sélection est effectuée, parmi les molécules de mélange proches de la molécule source traitée dans le plan TF, est sélectionnée celle qui lui est le plus corrélée (celle qui lui ressemble le plus)

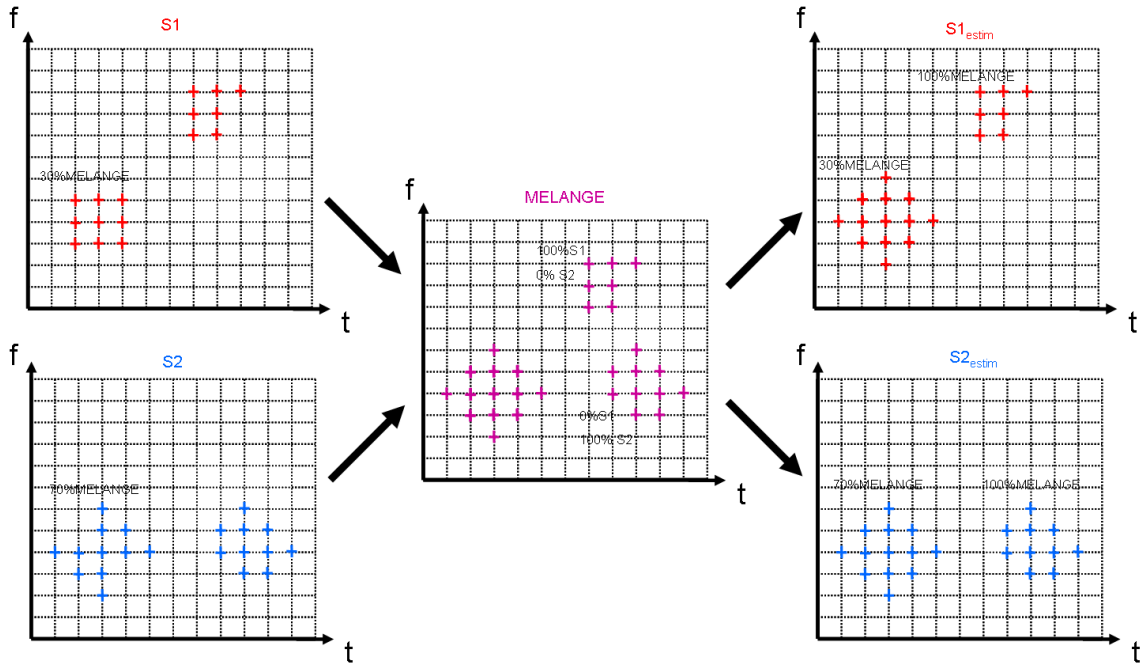


Figure A.1 – Reconstruction de deux sources à partir de l'information de gain dans le cas de l'utilisation du MMP.

Annexe B

Sélection des sources prédominantes en SSI-I

B.1 Cas de deux sources prédominantes

Nous illustrons ici l'implémentation de la sélection des sources prédominantes présentée à la Section 6.2.2 lorsque $I_{ft} = \text{card}(\mathcal{P}) = J$ dans le cas de figure d'un mélange LISS de quatre signaux sources.

La sélection des J sources prépondérantes permet de ramener localement le mélange à une configuration déterminée, et permet donc une estimation des sources actives par inversion exacte du mélange (matrice de mélange carrée de rang plein). L'équation de mélange s'écrit dans le domaine temporel

$$\begin{bmatrix} x_L(t) \\ x_R(t) \end{bmatrix} = \begin{bmatrix} a_{L1} & a_{L2} & a_{L3} & a_{L4} \\ a_{R1} & a_{R2} & a_{R3} & a_{R4} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \\ s_4(t) \end{bmatrix} \tag{B.1}$$

et dans le domaine temps-fréquence

$$\begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} = \begin{bmatrix} a_{L1} & a_{L2} & a_{L3} & a_{L4} \\ a_{R1} & a_{R2} & a_{R3} & a_{R4} \end{bmatrix} \cdot \begin{bmatrix} S_1(f,t) \\ S_2(f,t) \\ S_3(f,t) \\ S_4(f,t) \end{bmatrix} \tag{B.2}$$

Considérer par exemple que les sources s_1 et s_3 sont prépondérantes au bin temps-fréquence (f,t) revient à faire l'approximation suivante sur le signal de mélange

$$\begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} \approx \begin{bmatrix} a_{L1} & a_{L3} \\ a_{R1} & a_{R3} \end{bmatrix} \cdot \begin{bmatrix} S_1(f,t) \\ S_3(f,t) \end{bmatrix} \tag{B.3}$$

La différence entre le vecteur sources estimé $\hat{\mathbf{S}}$ et le vecteur source original $\mathbf{S}$ en (f,t) est, dans cette hypothèse où s_1 et s_3 sont prépondérantes au bin temps-fréquence (f,t) :

$$\begin{bmatrix} \hat{S}_1(f,t) - S_1(f,t) \\ \hat{S}_2(f,t) - S_2(f,t) \\ \hat{S}_3(f,t) - S_3(f,t) \\ \hat{S}_4(f,t) - S_4(f,t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{13}^{-1}\mathbf{A}_{24} \\ -\mathbf{I}_2 \end{bmatrix} \cdot \mathbf{S}_{24}(f,t) \tag{B.4}$$

avec

$$\mathbf{A}_{13} = \begin{bmatrix} a_{L1} & a_{L3} \\ a_{R1} & a_{R3} \end{bmatrix} \quad \text{et} \quad \mathbf{S}_{24}(f,t) = \begin{bmatrix} S_2(f,t) \\ S_4(f,t) \end{bmatrix} \tag{B.5}$$

Une expression analogue de la différence entre $\hat{\mathbf{S}}$ et $\mathbf{S}$ peut être obtenue pour chaque couple de sources supposées prédominantes $\{s_l, s_k\}$, $k, l \in [1, I]$. On en déduit que $\tilde{\mathcal{I}}_{ft}$ est déterminé en identifiant la combinaison $\mathcal{I}_{ft}$ minimisant

$$\| \mathbf{B}_{\mathcal{I}_{ft}} \mathbf{S}_{\overline{\mathcal{I}_{ft}}}(f,t) \|^2 \tag{B.6}$$

parmi les $C_2^4 = 6$ combinaisons possibles de $\mathcal{P} = [\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}]$, et avec

$$\mathbf{B}_{\mathcal{I}_{ft}} = \begin{bmatrix} \mathbf{A}_{\mathcal{I}_{ft}}^{\dagger} \mathbf{A}_{\overline{\mathcal{I}_{ft}}} \\ -\mathbf{I}_{I-I_{ft}} \end{bmatrix} \tag{B.7}$$

Si par exemple $\tilde{\mathcal{I}}_{ft} = \{1,2\}$ les signaux sources s_1 et s_2 sont estimés par l'équation

$$\begin{bmatrix} \hat{S}_1(f,t) \\ \hat{S}_2(f,t) \end{bmatrix} = \begin{bmatrix} a_{L1} & a_{L2} \\ a_{R1} & a_{R2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} \tag{B.8}$$

alors que $\hat{S}_3 = \hat{S}_4 = 0$.

B.2 Cas d'une seule source prédominante

Nous illustrons ici l'implémentation de la sélection d'une unique source prédominante présentée à la Section 6.2.2 dans le cas de figure d'un mélange LISS de quatre signaux sources.

La sélection de la source prépondérante permet de ramener localement le mélange à une configuration sur-déterminée, et permet donc une estimation de la source active par pseudo-inversion du mélange (la matrice de mélange est alors réduite à un vecteur colonne).

Considérer par exemple que la sources s_2 est prépondérante au bin temps-fréquence (f,t) revient à faire l'approximation suivante sur l'équation de mélange dans le domaine temps-fréquence :

$$\begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} \approx \begin{bmatrix} a_{L2} \\ a_{R2} \end{bmatrix} \cdot S_2(f,t) \tag{B.9}$$

L'hypothèse est faite que seule la source s_2 est active au bin TF (f,t), donc $\hat{S}_1(f,t) = \hat{S}_3(f,t) = \hat{S}_4(f,t) = 0$. De plus la pseudo-inverse d'un vecteur colonne $\mathbf{A}_k$ est définie par

$$\mathbf{A}_k^{\dagger} = \mathbf{A}_k^{\mathrm{T}} / \parallel \mathbf{A}_k \parallel_2^2 \tag{B.10}$$

c'est à dire dans le présent exemple

$$\mathbf{A}_2^{\dagger} = \frac{1}{a_{L2}^2 + a_{R2}^2} \left[a_{L2} \; a_{R2} \right] \tag{B.11}$$

La différence entre le vecteur sources estimé $\hat{\mathbf{S}}$ et le vecteur source original $\mathbf{S}$ en (f,t), dans l'hypothèse où s_2 est prépondérante au bin temps-fréquence (f,t) est donnée par

$$\begin{bmatrix} \hat{S}_1(f,t) - S_1(f,t) \\ \hat{S}_2(f,t) - S_2(f,t) \\ \hat{S}_3(f,t) - S_3(f,t) \\ \hat{S}_4(f,t) - S_4(f,t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_2^{\dagger} \mathbf{A}_{134} \\ -\mathbf{I}_3 \end{bmatrix} \cdot \mathbf{S}_{134}(f,t) \tag{B.12}$$

avec

$$\mathbf{A}_2 = \begin{bmatrix} a_{L2} \\ a_{R2} \end{bmatrix} \quad \text{et} \quad \mathbf{A}_{134} = \begin{bmatrix} a_{L1} & a_{L3} & a_{L4} \\ a_{R1} & a_{R3} & a_{R4} \end{bmatrix} \quad \text{et} \quad \mathbf{S}_{134}(f,t) = \begin{bmatrix} S_1(f,t) \\ S_3(f,t) \\ S_4(f,t) \end{bmatrix} \tag{B.13}$$

Une expression analogue de la différence entre $\hat{\mathbf{S}}$ et $\mathbf{S}$ peut être obtenue pour chaque source s_k supposée prédominante $\{s_k\}$, $k \in [1, I]$. De façon similaire à l'équation (B.6), $\tilde{\mathcal{I}}_{ft}$ est déterminé en identifiant l'indice $\mathcal{I}_{ft}$ minimisant

$$\parallel \mathbf{B}_{\mathcal{I}_{ft}} \mathbf{S}_{\overline{\mathcal{I}_{ft}}}(f,t) \parallel^2$$

parmi les $C_1^4 = 4$ singletons possibles de $\mathcal{P} = [\{1\}, \{2\}, \{3\}, \{4\}]$. Si par exemple $\tilde{\mathcal{I}}_{ft} = \{3\}$ le signal source s_3 est estimé par

$$\hat{S}_3(f,t) \;=\; \begin{bmatrix} a_{L3} \\ a_{R3} \end{bmatrix}^{\dagger} \cdot \begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} \;=\; \frac{1}{a_{L3}^2 + a_{R3}^2} \left[a_{L3} \; a_{R3} \right] \cdot \begin{bmatrix} X_L(f,t) \\ X_R(f,t) \end{bmatrix} \tag{B.14}$$

les autres sources étant fixées à zéro par hypothèse : $\hat{S}_1 = \hat{S}_2 = \hat{S}_4 = 0$.

Annexe C

La technique de tatouage QIM avec MPA

Nous donnons dans cette Annexe les grandes lignes du MPA et de la technique de tatouage introduits à la Section 5.3.2, et utilisés dans [Pinel et al., 2009].

Le MPA

Le modèle psycho-acoustique utilisé est directement inspiré du MPA du standard MPEG-AAC [AAC, 2004] [Derrien et al., 2000]. Comme nous l'avons déjà mentionné, le MPA fournit, pour chaque sous-bande fréquentielle de tatouage, un seuil de masquage, $M(f)$ qui représente la puissance maximum de la distortion induite par le tatouage qui peut être insérée dans le signal de mélange sous contrainte d'inaudibilité.

Le MPA est calculé dans le domaine TF, mais en utilisant la DFT, et non la MDCT. En résumé, une première courbe de masquage est calculée à partir de la convolution du spectre de puissance de la DFT du signal de mélange avec une fonction d'étalement qui modélise le phénomène de masquage fréquentiel. Cette courbe est ensuite ajustée en fonction de la tonalité du signal (mieux estimée par l'information de phase de la transformée de Fourier), puis combinée au seuil d'audition absolu pour fournir le seuil de masquage global. Enfin, un rapport signal à masque (Signal-to-Mask Ratio ou SMR en anglais) est calculé comme la différence entre le seuil de masquage global et le spectre du signal. Le seuil de masquage en MDCT $M(f)$ est obtenu en calculant le rapport entre le spectre de puissance en MDCT et le SMR. Un ajustement de la courbe de masquage avant addition au seuil de masquage absolu permet de controler la capacité disponible pour le tatouage, et ainsi de l'adapter aux besoins.

La QIM

Rappelons que la technique de Quantization Index Modulation (QIM) utilisée est directement appliquée sur les coefficients MDCT. Le MPA utilisé fournit, à chaque canal fréquentiel f, une courbe de masquage $M(f)$ qui détermine la capacité maximale

$C(f)$ d'information de tatouage qui peut être insérée dans le signal de mélange sous contrainte d'inaudibilité. Pour chaque coefficient MDCT, à chaque canal fréquentiel, un jeu de $2^{C(f)}$ quantificateurs entrelacés est alors défini. Les quantificateurs sont uniformes et leur entrelacement régulier de manière à minimiser la probabilité d'erreur de décodage. Chaque quantificateur représente un code de $C(f)$ bits. Tatouer un code de $C(f)$ bits sur un coefficient MDCT revient à quantifier ce coefficient MDCT avec le quantificateur indexé par ce code. La Figure C.1 illustre ce principe avec une capacité d'insertion $C(f) = 2$, *i.e.* 4 quantificateurs entrelacés. Le code de 2 bits à tatouer est ici '01'.

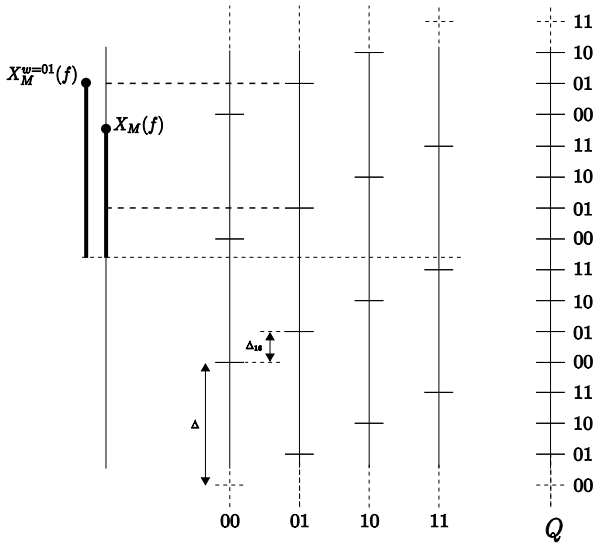


Figure C.1 – Exemple d'un jeu de 4 quantificateurs pour la quantification de type QIM, et le quantificateur Q résultant.

Un coefficient MDCT est donc remplacé par sa valeur quantifiée sur le quantificateur correspondant au code à tatouer. Au décodeur, ce code est retrouvé en comparant le coefficient MDCT quantifié transmis (possiblement dégradé par la conversion PCM 16 bits) avec chacun des $2^{C(f)}$ quantificateurs, supposés connus au décodeur. Le quantificateur possédant un niveau de quantification le plus proche du MDCT quantifié transmis est sélectionné, fournissant ainsi le code par lequel il est indexé.

Rappelons qu'en SSI le tatouage doit vérifier deux contraintes majeures : d'une part être inaudible, ce qui est assuré par l'utilisation d'un MPA dont les caractéristiques sont décrites plus tard, et être robuste à la conversion PCM 16 bits intervenant en fin de codeur. Intéressons nous d'abord à cette dernière contrainte, et aux effets de la PCM sur les coefficients MDCT. Il a été démontré dans [Pinel et al., 2009] que le bruit $B(f)$ introduit sur les coefficients MDCT par la PCM possède une variance $\sigma^2_{MDCT}(f)$ identique à celle, notée σ^2_{16}, du bruit uniforme introduit sur les échantillons temporels par la PCM. Par conséquent, $\sigma^2_{MDCT}(f)$ est indépendant de la fréquence, et l'on a :

$$\sigma^2_{MDCT}(f) = \sigma^2_{16} = \frac{(2^{-15})^2}{12}, \forall f \in \left[0, \frac{N}{2}\right] \tag{C.1}$$

où N représente la taille de la fenêtre d'analyse des MDCT. En accord avec le théorème

central limite, comme vu dans [Pinel et al., 2009], le bruit introduit sur les coefficients MDCT par la PCM suit une loi normale :

$$B(f) \sim \mathcal{N}\left(0, \sigma_{16}^2\right), \forall f \in \left[0, \frac{N}{2}\right]$$

Connaissant la distribution du bruit blanc gaussien additif introduit sur les MDCT par la PCM, il est possible de définir le pas de quantification minimal Δ_{16} entre deux quantificateurs entrelacés en fonction d'une probabilité p_e d'erreur de décodage prédéfinie (p_e est ici fixée à 10^{-6}). Il a été établi dans [Pinel et al., 2009] que

$$\Delta_{16} = 2\sqrt{2}\sigma_{16}\text{erf}^{-1}(1 - p_e) \tag{C.2}$$

avec erf la fonction d'erreur définie par $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t$.

Enfin, le pas de quantification $\Delta(f)$ des quantificateurs entrelacés utilisés au canal fréquentiel f vaut

$$\Delta(f) = \Delta_{16} 2^{C(f)} \tag{C.3}$$

La contrainte de robustesse à la conversion PCM étant respectée, il reste à établir la condition sur $\Delta(f)$ qui assure l'inaudibilité du tatouage.

La contrainte d'inaudibilité est directement reliée au seuil de masquage $M(f)$ fournit par le MPA. La puissance de l'erreur induite par le tatouage doit demeurer sous ce seuil de masquage. La dégradation maximale induite par le tatouage par quantification tel que considéré en SSI est $\Delta(f)/2$, par conséquent, la contrainte d'inaudibilité peut s'écrire

$$(\Delta(f))^2 < M(f), \forall f \in \left[0, \frac{N}{2}\right] \tag{C.4}$$

Par combinaison des formules (C.3) et (C.4), la capacité maximale d'insertion de l'information de tatouage au canal fréquentiel f sous contrainte d'inaudibilité vaut finallement :

$$C(f) = \lfloor \frac{1}{2} \log_2 \left(\frac{M(f)}{\Delta_{16}^2} \right) + 1 \rfloor \tag{C.5}$$

Le processus de séparation en SSI étant guidé par le tatouage, il est primordial que la watermark décodée soit identique à celle insérée au codeur (à la probabilité d'erreur préétablie près). Contrairement à la première implémentation du tatouage présentée à la Section 5.1.3 dans laquelle les quantificateurs utilisés étaient construits de manière à pouvoir être retrouvés au décodeur, lors de l'utilisation d'un MPA, la dégradation subie par le signal de mélange lors de la phase de tatouage empêche de retrouver la même courbe de masquage $M(f)$ au décodeur qu'au codeur. Par conséquent, comme il apparaît dans l'équation (C.5), il est impossible de retrouver les mêmes capacités d'insertion qu'au codeur, d'où l'impossibilité de décoder correctement le tatouage. Une des solutions retenue est de diviser le spectre en deux zones, une zone hautes fréquences

(HF) composée des derniers canaux fréquentiels et une zone basse fréquence (BF) constituée de la plus grande partie du spectre (qui correspond surtout à la zone la plus sensible pour l'oreille humaine). Les capacités des zones de BF sont alors tatouées dans la partie HF, et ce avec une résolution fixe connue au décodeur. Le tatouage des capacités BF $C_{BF}(f)$ dans la portion HF est basée sur l'observation que la puissance des signaux audio est généralement nettement en dessous du seuil d'audition absolue dans ces portions du spectre.

Au décodeur, les capacités $C_{BF}(f)$ sont dans un premier temps décodées (en utilisant les QSU de capacités $C_{HF}(f)$ connues), puis le tatouage utile à la séparation proprement dite est dans un deuxième temps décodé en utilisant les QSU de capacités $C_{BF}(f)$. Il a été observé dans [Pinel et al., 2009] que les valeurs de $C_{BF}(f)$ obtenues expérimentalement sont toujours inférieures strictement à 15, d'où un codage possible de cette valeur de capacité sur 4 bits. Cependant les capacités disponibles en HF (pour une portion HF de taille raisonable) ne sont pas suffisantes pour tatouer les capacités $C_{BF}(f)$ à chaque bin TF. Par conséquent, il est choisit de travailler par sous-bandes fréquentielles, dites sous-bandes de tatouage : l'ensemble des coeffcients MDCT d'une sous-bande auront la même capacité d'insertion. Plusieurs découpages en sous-bandes de tatouage sont étudiés dans [Pinel et al., 2009]. Dans la présente thèse, nous n'avons considéré que le découpage en 32 sous-bandes également distribuées. Pour une fenêtre d'analyse de $N = 2048$ échantillons, chaque sous-bande contient donc $\frac{N/2}{32} = 32$ bins fréquentiels. Les deux dernières sous-bandes sont utilisées pour définir la zone HF, et les capacités hautes fréquences $C_{HF}(f)$ sont alors fixées à 1 et 2 bits sur la 31 et 32^e sous-bande respectivement.

Bibliographie

(2004). Iso/iec 13818-7 : Information technology - generic coding of moving pictures and associated audio information - part 7 : Advanced audio coding (AAC).

Aoki, M., M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda (2001). Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science and Technology 22*(2), 149–157.

Araki, S., H. Sawada, and S. Makino (2007). *Blind Speech Separation*, Chapter K-means based Underdetermined blind speech separation. Springer.

Araki, S., H. Sawada, R. Mukai, and S. Makino (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing 87*(8), 1833–1847.

Aïssa-El-Bey, A., K. Abed-Meraim, A., and Y. Grenier (2007). Blind separation of underdetermined convolutive mixtures using their time-frequency representation. *IEEE Trans. on Audio, Speech and Language Processing 15*(5), 1540–1550.

Aïssa-El-Bey, A., N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, A., and Y. Grenier (2007). Underdetermined blind separation of nondisjoint sources in the time-frequency domain. *IEEE Transactions on Signal Processing 55*(3), 897–907.

Baker, H. L. (1984). *Vector Quantization of Digital Images*. Ph. D. thesis, Department of Electrical Engineering, Stanford University.

Balan, R., J. Rosca, and S. Rickard (2003a). Non-square blind source separation under coherent noise by beamforming and time-frequency masking. In *Int. Conf. on Independant Component Analysis and Signal Separation (ICA'03)*, Nara, Japon, pp. 313–318.

Balan, R., J. Rosca, and S. Rickard (2003b). Scalable non-square blind source separation in the presence of noise. In *Int. Conf on Acoustics, Speech and Signal Processing (ICASSP'03)*, Hong-Kong, China, pp. 293–296.

Baumgarte, F. and C. Faller (2003). Binaural cue coding - part i : Psychoacoustic fundamentals and design principles. *IEEE Trans. on Speech and Audio Processing 11*(6), 509–519.

Belouchrani, A. and K. Abed-Meraim (1993). Séparation aveugle au second ordre de sources corrélées. In *GRETSI*, Juan-Les-Pins, France, pp. 309–312.

Benaroya, L., F. Bimbot, and R. Gribonval (2006). Audio sources separation with a single sensor. *IEEE Trans. Audio, Speech, et Language Proc. 14*(1), 191–199.

Benaroya, L., R. Gribonval, and F. Bimbot (2001). Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Proceedings of the 18th Symposium GRETSI'01 on Signal and Image Processing*, Toulouse, France.

Blin, A., S. Araki, and S. Makino (2004). A sparseness-mixing matrix estimation (smme) solving the underdetermined bss for convolutive mixtures. In *Int. Conf on Acoustics, Speech and Signal Processing (ICASSP'04)*, Montréal, Canada, pp. 85–88.

Bofill, P. (2002). Underdetermined blind separation of delayed sound sources in the frequency domain. Technical report, Universitat Politecnica de Catalunya. UPC-DAC-2001-14.

Bofill, P. and E. Monte (2006). Underdetermined convoluted source reconstruction using lp and socp, and a neural approximator of the optimizer. In *Intl. Conf on Independent Component Analysis and Signal Separation (ICA'06)*, Charleston, SC, USA, pp. 569–576.

Bofill, P. and M. Zibulevski (2001). Underdetermined blind source separation using sparse representations. *Signal Processing 81*(11), 2353–2362.

Boney, L., T. Ahmed, and H.Khaled (1996). Digital watermarks for audio signals. In *IEEE Intl. Conf. on Multimedia Computing and Systems (ICMCS'96)*, Hiroshima, Japan, pp. 473–480.

Bosi, M. and R. Goldberg (2003). *Introduction to odigital audio coding*. Kluwer Academic Publishers.

Bourcet, P., D. Masse, and B. Jahan (1995). Système de diffusion de données. Brevet d'Invention 95 06727, Télédiffusion de France.

Brandenburg, K. and M. Bosi (1997). Overview of mpeg audio : Current and future standards for low bit-rate audio coding. *Journal of the Audio Engineering Society 45*(1), 4–21.

Bronkhorst, W. (2000). The cocktail party phenomenon : A review on speech intelligibility in multiple-talker conditions. *Acta Acustica 86*, 117–128.

Brown, J. (1991). Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America 89*(1), 425–434.

Casey, M. and A. Westner (2000). Separation of mixed audio sources by independent subspace analysis. In *Int. Computer Music Conf.*, Berlin, Germany.

Chen, B. and C.-E. Sundberg (2000). Digital audio broadcasting in the fm band by means of contiguous band insertion and precanceling techniques. *IEEE Trans. on Communications 48*(10), 1634–1637.

Chen, B. and G. Wornell (2001). Quantization index modulation : a class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Information Theory 47*, 1423–1443.

Cho, N., Y. Shiu, and C.-C. J. Kuo (2007). Audio source separation with matching pursuit and content-adaptative dictionaries. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, New Paltz, NY.

Comon, P. (1991). Independent component analysis. In *Proc. Int. Sig. Proc. Workshop on Higher-Order Statistics*, Chamrousse, France, pp. 111–120.

Comon, P. and C. Jutten (2007a). *Séparation de sources - Au-delà de l'aveugle et applications*, Volume 2. Hermès-Lavoisier.

Comon, P. and C. Jutten (2007b). *Séparation de sources - Concepts de base et analyse en composantes indépendantes*, Volume 1. Hermès-Lavoisier.

Comon, P. and C. Jutten (2010). *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press.

Costa, M. (1983). Writing on dirty paper. *IEEE Trans. Information Theory 29*, 439–441.

Cox, I., F. Leighton, and T. Shamoon (1997). Secure spread spectrum watermarking for multimedia. *IEEE Trans. on Image Processing 6*(12), 1673–1687.

Cox, I. J., M. L. Miller, and A. L. McKellips (1999). Watermarking as communications with side information. *IEEE Proceedings 87*(7), 1127–1141.

Cvejic, N. and T. Seppanen (2002a). Increasing the capacity of LSB-based audio steganography. In *IEEE Workshop on Multimedia Signal Processing*, St. Thomas, Virgin Islands, USA, pp. 336–338.

Cvejic, N. and T. Seppanen (2002b). A wavelet domain LSB insertion algorithm for high capacity audio steganography. In *IEEE Digital Signal Processing Workshop*, Number 10, Georgia, USA, pp. 53–55.

Cvejic, N. and T. Seppanen (2004). Increasing robustness of lsb audio steganography using a novel embedding method. In *Proc. of the Intl. Conf. on Information Technology : Coding and Computing (ITCC'04)*, Volume 2, Las Vegas, NV, USA, pp. 533–537.

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Stat. 21*, 2–8.

Daudet, L. (2006). Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Trans. Audio, Speech and Language Proc. 14*(5).

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

Derrien, O., S. Larbi, M. P. Guimares, and N. Moreau (2000). Le codeur mpeg-2 aac expliqué aux traiteurs de signaux. *Ann. Télécommun. 55*(9-10), 442–461.

Ellis, D. (1996, June). *Prediction-driven computational auditory scene analysis*. Ph. D. thesis, Massachusetts Institute of Technology (MIT), Cambridge, MA.

Ellis, D. (1999). Using knowledge to organize sound : The prediction-driven approach to computational auditory scene analysis, and its application to speech/non-speech mixture. *Speech Communication 27*(3), 281–298.

Faller, C. (2004). Parametric coding of spatial audio. In *Int. Conf on Digital Audio Effects*, Naples, Italy.

Faller, C. (2006). Parametric multichannel audio coding ; synthesis of coherence cues. *IEEE Trans. on Audio, Speech and Language Processing 14*(1), 299–310.

Faller, C. and F. Baumgarte (2003). Binaural cue coding - part ii : Schemes and applications. *IEEE Trans. on Speech and Audio Processing 11*(6), 520–531.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. J. Wiley & Sons.

Fisher, W. M., G. R. Doddington, and K. M. Goudie-Marshall (1986). The darpa speech recognition research database : Specifications and status. *Proc. of DARPA Workshop on Speech Recognition*, 93–99.

Friedman, J. and J. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. In *IEEE Trans. on Computers*, Volume 23, pp. 881–890.

Garcia, R. (1999). Digital watermarking of audio signals using psychoacoustic auditory model and spread spectrum theory. In *107th Convention of Audio Engineering Society (AES)*, New York, NY.

Gil-Je, L., Y. Eun-Jun, and Y. Kee-Young (2008). A new lsb based digital watermarking scheme with random mapping function. In *Intl. Symposium on Ubiquitous Multimedia Computing (UMC'08)*, Volume 2, Hobart, Australia, pp. 130–134.

Godsmark, D. and G. Brown (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication 27*(3), 351–366.

Gray, R. M. and A. Gersho (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Pub.

Gribonval, R. and E. Bacry (2003). Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Proc. 51*(1), 101–112.

Gribonval, R. and S. Lesage (2006). A survey of blind component analysis for blind source separation : principles, perspectives and new challenges. In *Europ. Symp. on Artificial Neural Networks (ESANN'06)*, Bruges, Belgique, pp. 323–330.

Herre, J., C. Faller, S. Disch, C. Ertel, J. Hilpert, K. Linzmeier, C. Spenger, and P. Kroon (2004). Spatial audio coding : Next-generation efficient and compatible coding of multi-channel audio. In *117th Convention of Audio Engineering Society (AES)*, San Francisco, CA, USA.

Herre, J., H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and F. Myburg (2005). The reference model architecture for mpeg spatial audio coding. In *118th Convention of Audio Engineering Society (AES)*, Barcelona, Spain.

Hulle, M. V. (1999). Clustering approach to square and non-square blind source separation. In *IEEE Workshop on Neural Networks for Signal Processing*, Madison, WI, USA, pp. 315–323.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent Component Analysis*. Wiley & Sons.

Iliev, A. and M. Scordilis (2004). Multi level high capacity data hiding technique for stereo audio. In *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, pp. 1793–1797.

ITU-R BS.1116, International Telecommunications Union, R. S. B. r. I. (1997). Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.

ITU-R BS.562-3, International Telecommunications Union, R. S. B.-. (1990). Subjective assessment of sound quality.

Jourjine, A., S. Rickard, and O. Yilmaz (2000). Blind separation of disjoint orthogonal signals : Demixing n sources from 2 mixtures. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'00)*, Volume 5, Istanbul, Turkey, pp. 2985–2988.

Karvanen, J. and A. Cichocki (2003). Measuring sparseness of noisy signals. In *Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, pp. 125–130.

Kinoshita, T., S. Sakai, and H. Tanaka (1999). Musical sound source identification based on frequency component adaptation. In *IJCAI Workshop on Computational Auditory Scene Analysis*, Stockholm, Sweden, pp. 18–24.

Lin, J., D. Grier, and J. Cowan (1997). Feature extraction approach to blind source separation. In *IEEE workshop on Neural Networks for Signal Processing (NNSP)*, Amelia Island Plantation, Florida, pp. 398–405.

Linde, Y., A. Buzo, and R. M. Gray (1980). Algorithm for vector quantizer design. *Trans. IEEE Commun. 28*(1), 84–95.

Linh-Trung, N., A. Belouchrani, K. Abed-Meraim, and B. Boashash (2005). Separating more sources than sensors using time-frequency distributions. *Journal of Applied Signal Processing 2005*(17), 2828–2847.

Liu, Y.-W. (2007). Sound source segregation assisted by audio watermarking. In *IEEE Int. Conf. Multimedia and Expo*, Beijin, China, pp. 200–203.

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

Melia, T. and S. Rickard (2005). Extending the duet blind source separation technique. In *Structure et Parcimonie pour la Représentation Adaptative de Signaux (SPARS'05)*, Rennes, France.

Melia, T. and S. Rickard (2007). Underdetermined blind source separation in echoic environments using desprit. *Journal on Advances in Signal Processing (EURASIP)*, 19 pages.

Molla, K. and K. Hirose (2003). Single-mixture audio source separation by subspace decomposition of hilbert spectrum. *IEEE Trans. Audio, Speech, et Language Proc. 15*(3), 893–900.

Murata, N., S. Ikeda, and A. Ziehe (2001). An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing 41*, 1–24.

Nakamura, T., R. Tachibana, and S. Kobayashi (2002). Automatic music monitoring and boundary detection for broadcast using audio watermarking. In *SPIE Electronic Imaging : Security and Watermarking of Multimedia Content IV*, Volume 4675, pp. 170–180.

Nesbit, A., M. Davies, M. Plumbley, and M. Sandler (2006). Source extraction from two-channel mixtures by joint cosine packet analysis. In *European Signal Processing Conference (EUSIPCO)*, Florence, Italy.

Nesbit, A. and M. Plumbley (2008). Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'08)*, Las Vegas, NV, USA, pp. 41–44.

Nesbit, A., M. D. Plumbley, and M. E. Davies (2007). Audio source separation with a signal-adaptive local cosine transform. *Signal Processing 87*(8), 1848 – 1858.

Oehler, K. L. and R. M. Gray (1993). Mean-gain-shape vector quantization. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'93)*, Minneapolis, Minnesota, USA, pp. 241–244.

O'Grady, P. and B. Pearlmutter (2004). Hard-lost : modified k-means for oriented lines. In *Irish Signals and Systems conference*, Belfast, Ireland, pp. 247–252.

O'Grady, P., B. A. Pearlmutter, and S. Rickard (2005). Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology 15*(1), 18–33.

Olsson, R. K. and L. K. Hansen (2006). Blind separation of more sources than sensors in convolutive mixtures. In *Int. Conf on Acoustics, Speech and Signal Processing (ICASSP'06)*, Volume 5, Toulouse, France, pp. 657–660.

Parra, L. and C. Spence (2000). Convolutive blind separation of non-stationnary sources. *IEEE Trans. on Speech and Audio Processing 8*(3), 320–327.

Parvaix, M., S. Krishnan, and C. Ioana (2008). An audio watermarking method based on molecular matching pursuit. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'08)*.

Pedersen, M., J. Larsen, U. Kjems, and L. Parra (2007). A survey of convolutive blind source separation methods. *Springer Handbook on Speech Processing and Speech Communication*, 1–34.

Peterson, J. M. and S. Kadambe (2003). A probabilistic approach for blind source separation of underdetermined convolutive mixtures. In *Int. Conf on Acoustics, Speech and Signal Processing (ICASSP'03)*, Volume 6, Hong Kong, China, pp. 581–584.

Pham, D. and J.-F. Cardoso (2001). Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. on Signal Processing 49*(9), 1837–1848.

Pinel, J., C. Baras, and L. Girin (2009). Développement d'un système de tatouage audionumérique haute capacité. Master's thesis, GIPSA-lab, Grenoble, France.

Pinel, J., L. Girin, and C. Baras (2010a). A high-rate data hiding technique for uncompressed audio signals. *IEEE Trans. on Audio, Speech and Language Processing*. submitted.

Pinel, J., L. Girin, and C. Baras (2010b). Une technique de tatouage "haute-capacité" pour signaux musicaux au format cd-audio. In *Actes du Congrès Français d'Acoustique*, Lyon, France.

Pinel, J., L. Girin, C. Baras, and M. Parvaix (2010). A high-capacity watermarking technique for audio signals based on MDCT-domain quantization. In *Int. Congress on Acoustics (ICA)*, Sydney, Australia.

Princen, J. P. and A. B. Bradley (1986). Analysis/synthesis filter bank design based on time domain aliasing cancellation. In *IEEE Trans. Acoust. Speech Sig. Proc.*, Volume 64, pp. 1153–1161.

Princen, J. P., A. W. Johnson, and A. B. Bradley (1987). Subband/transform coding using filter bank designs based on time domain aliasing cancellation. In *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'87)*, Volume 12, Dallas, Texas, USA, pp. 2161–2164.

Rickard, S. (2006). Sparse sources are separated sources. In *European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy.

Rickard, S. and O. Yilmaz (2002). On the approximate w-disjoint orthogonality of speech. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'02)*, Orlando, FL, USA, pp. 529–532.

Rulon, B., M. Shaw, and K. Donohue (1999). A comparison of audio compression transforms. In *IEEE Southeastcon'99*, Lexington, KY, USA, pp. 253–257.

Sawada, H., R. Mukai, S. Araki, and S. Makino (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. on Speech and Audio Processing 12*(5), 530–538.

Sinha, D. and J. Johnston (1996). Audio compression at low bit rate using a signal adaptive switched filterbank. In *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Volume 2, Atlanta, Georgia, USA, pp. 1053–1056.

Suresh, K. and T. Sreenivas (2009). Linear filtering in DCT IV/DST IV and MDCT/MDST domain. *Signal Processing 89*(6), 1081–1089.

Tachibana, R. (2003). Audio watermarking for live performance. *SPIE Electronic Imaging : Security and Watermarking of Multimedia Content V 5020*, 32–43.

Tan, V. Y. F. and C. Févotte (2005). A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS'05)*, Rennes, France.

Thiede, T., W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, and C. Colomes (2000). PEAQ - the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society 48*(1), 3–29.

Trefethen, L. and D. Bau (1997). *Numerical Linear Algebra*, pp. 77–85. Society for Industrial and Applied Mathematics.

Vincent, E., R. Gribonval, and C. Fevotte (2005). Performance measurement in blind audio source separation. *IEEE Trans. Speech Audio Process. 14*(4), 1462–1469.

Vincent, E., R. Gribonval, and M. Plumbleys (2007). Oracle estimators for the benchùarking of source separation algorithms. *Signal Processing 87*(2007), 1933–1950.

Viste, H. and G. Evangelista (2006). A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures. *IEEE Trans. on Audio, Speech and Language Processing 14*(3), 1051–1061.

Winter, S., W. Kellermann, H. Sawada, and S. Makino (2007). Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization. *Journal on Advances in Signal Processing EURASIP 2007*, 12 pages.

Wolfe, P. and S. Godsill (2003). A gabor regression scheme for audio signal analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 103–106.

Woodruff, J. and B. Prado (2007). Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo recordings. *Journal on Advances in Signal Processing (EURASIP) 2007*, 10 pages.

Yilmaz, O. and S. Rickard (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Proc. 52*(2004), 1830–1847.

Zibulevsky, M., P. Kisilev, Y. Zeevi, and B. A. Pearlmutter (2002). Blind source separation via multinode sparse representation. *Advances in Neural Information Processing Systems 14*, 1049–1056.

Zibulevsky, M., B. A. Pearlmutter, P. Bofill, and P. Kisilev (2001). *Independent Component Analysis : Principles and Practice*. Cambridge University Press.