HAL
open science

Séparation de source informée pour des mélanges stéréo instantanés utilisant un tatouage de l'index des sources localement prédominantes

Mathieu Parvaix, Laurent Girin

To cite this version:

Mathieu Parvaix, Laurent Girin. Séparation de source informée pour des mélanges stéréo instantanés utilisant un tatouage de l'index des sources localement prédominantes. CFA 2010 - 10ème Congrès Français d'Acoustique, Apr 2010, Lyon, France. pp.Cd-Rom. hal-00486818

HAL Id: hal-00486818

https://hal.science/hal-00486818v1

Submitted on 26 May 2010

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.

10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Séparation de source informée pour des mélanges stéréo instantanés utilisant un tatouage de l'index des sources localement prédominantes

Mathieu Parvaix, Laurent Girin

Grenoble Images Parole Signal Automatique (GIPSA-lab),CNRS UMR 5216, Grenoble-INP

{mathieu.parvaix,laurent.girin}@gipsa-lab.grenoble-inp.fr

Dans cette étude, nous traitons le problème de la séparation de sources musicales non-stationnaires dans le cas d'un mélange linéaire instantané stationnaire stéréo (deux canaux). Il s'agit de pouvoir isoler les signaux correspondants aux différents instruments et voix d'un mélange musical, pour permettre à un utilisateur de remixer librement la musique lors de sa restitution. Pour traiter ce problème difficile (du fait du nombre de sources généralement supérieur au nombre de capteurs, le mélange étant alors dit sous-déterminé), nous proposons un système original avec une séparation dite "informée", proche des approches de type "oracle", mais réaliste car reposant ici sur une configuration en deux niveaux : codeur pour la production du mélange, et décodeur pour la restitution. Au codeur, les signaux sources sont supposés disponibles avant leur mixage, comme c'est le cas pour une grande part des oeuvres musicales enregistrées en studio. Une analyse temps-fréquence comparant localement (c'est-à-dire sur un nombre restreint de bins temps-fréquence à la fois) le contenu des signaux sources permet de sélectionner la source ou les deux sources prédominantes composant le mélange (parmi $I \geq 2$ sources) en exploitant la parcimonie des signaux audio dans le plan temps-fréquence. Un code représentant l'index de cette (ces) source(s) est alors inséré dans le signal mélange par un procédé de tatouage (dont une version plus récente améliorée est présentée dans un autre papier dans ce même congrès). Au décodeur, où seul le signal mélange (tatoué) est disponible, l'extraction du tatouage et la sélection des sources correspondantes permet de réduire la configuration sous-déterminée à une configuration (sur-)déterminée et de procéder à la séparation en utilisant une inversion matricielle classique (la matrice de mélange globale étant supposée connue au décodeur ou transmise par tatouage). Des résultats de séparation très prometteurs obtenus pour la séparation de 4 signaux sources sont présentés.

1 Introduction

La séparation de sources a pour objectif de recouvrer I signaux sources $s_i[n], i \in [1, I]$, à partir de J observations de leur mélange $x_j[n], j \in [1, J]$. La configuration *sous-déterminée*, où $J < I$, est un cas de figure particulièrement complexe. Cette configuration ne peut pas être traitée par des techniques de Séparation Aveugle de Sources (SAS) / Analyse en Composantes Indépendantes (ACI) développées pour des mélanges (sur-)déterminés ($J \geq I$) [3] [7] [6] [5]. Cependant ce cas de figure présente un intérêt tout particulier dans le cas du traitement audio où, dans la configuration classique mono or stéréo, de nombreux instruments sont à séparer à partir de seulement une ou deux observations du mélange. Une telle séparation peut permettre, en aval, de manipuler séparément les différents éléments d'une scène audio, par exemple en modifiant le volume, le timbre ou la spatialisation d'un instrument, procédé connu sous le nom d'écoute active, ou remixage.

De nombreuses techniques de séparation de sources en configuration sous-déterminée se basent sur la parcimonie de signaux dans le plan temps-fréquence (TF) [15] [2]. Dans [10] [9] nous avons introduit le concept de Séparation de Sources Informée (SSI), avec une configuration spécifique codeur-décodeur correspondant aux deux grandes étapes que sont la production d'un signal audio (par exemple, dans le cadre de la musique, l'enregistrement/le mixage en studio) et sa restitution (par exemple lecture d'un CD-audio dans le cadre d'un usage privé). En plus des signaux de mélange disponibles lors de la séparation, au niveau de ce que l'on appelle ici le décodeur, les signaux sources sont supposés disponibles au niveau où est effectué le mélange, appelé ici le codeur. Au codeur, des paramètres sont extraits des signaux sources, et cette information supplémentaire est insérée dans le signal de mélange de manière inaudible par une technique de tatouage (*watermarking* en anglais). Extraire le tatouage au décodeur permet à un utilisateur qui n'a pas accès aux sources séparées originales (mais seulement au signal de mélange tatoué), de séparer ces sources à partir du mélange.

Comme pour la séparation aveugle de sources, différentes approches existent en SSI, en fonction des hypothèses faites sur les signaux sources (indépendance mutuelle, parcimonie) et sur le signal de mélange (instantané, anéchoïque, convolutif, sur/sous-déterminé). Il en résulte que l'information insérée dans le mélange, et la façon dont elle est utilisée lors de la séparation peut différer d'une configuration à l'autre. Dans [10] [9], un mélange linéaire instantané stationnaire (LIS) *monophonique* de signaux sources de musique ou de pa-

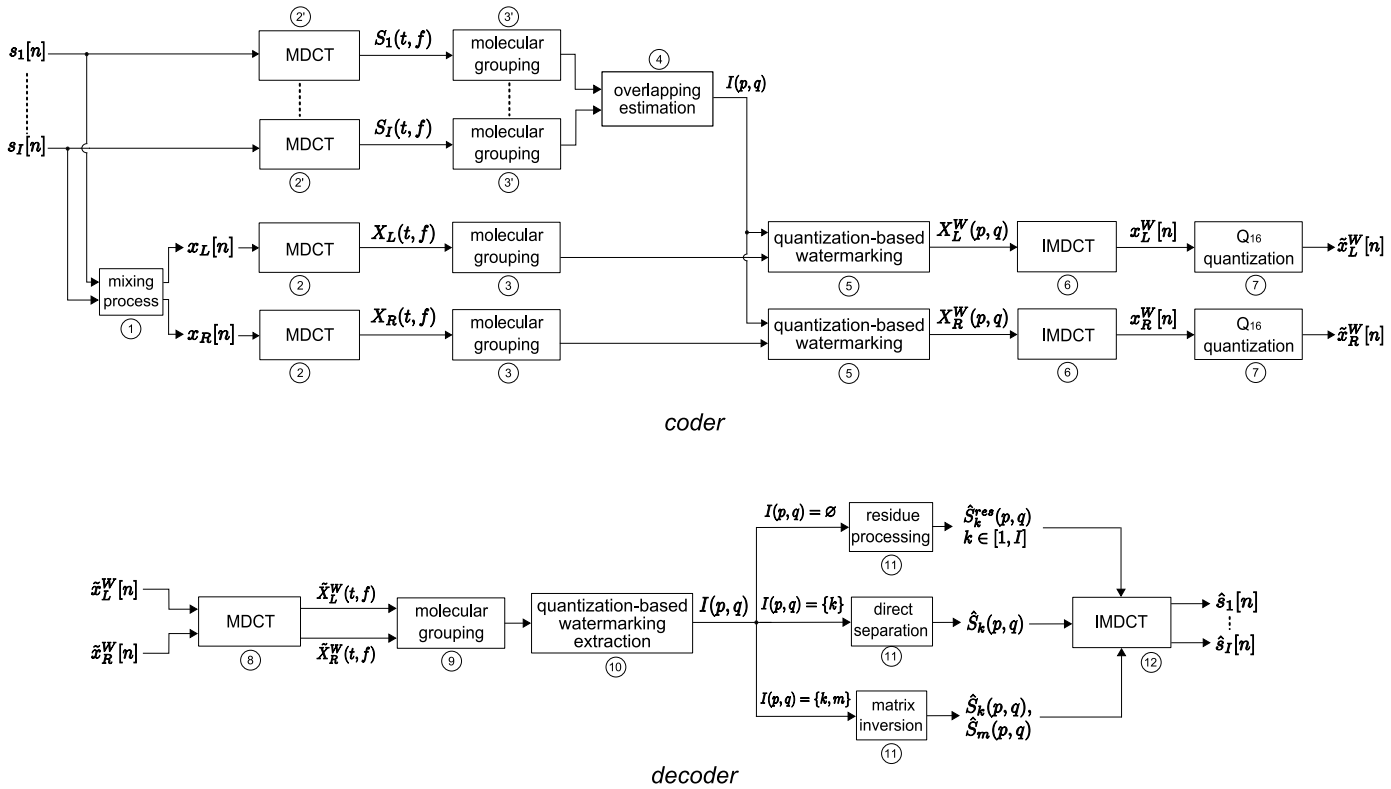

Figure 1 – Structure détaillée du système de SSI par indexation.

role était traité. L'information insérée consistait alors en des prototypes TF "moléculaires" des signaux sources issus de dictionnaires matriciels. Dans le cas présent, nous nous concentrons sur le cas d'un mélange LIS *stéréophonique* (2 voies) de signaux de musique. Dans ce cas de figure, l'information spatiale des sources est exploitée, et l'information supplémentaire insérée dans le mélange est réduite aux index des une à deux sources prédominantes dans chaque région du plan TF, cette information ayant été extraite par une analyse préalable des sources au codeur. Au décodeur, l'extraction de ces index permet de réduire localement le mélange, initialement sous-déterminé, à un mélange (sur-)déterminé. Le processus de séparation peut alors être réalisé par des techniques d'inversion matricielle classiques. Cette approche informée se rapproche ainsi des techniques de séparation dites *oracle* développées dans [14] [8], le résultat de l'oracle obtenu ici au codeur étant transmis au décodeur par tatouage.

Cet article s'organise de la façon suivante. La méthode proposée est décrite à la Section 2. Les résultats obtenus pour des signaux de musique sont donnés à la Section 3. Enfin, une conclusion sur cette étude et des perspectives de travail sont fournies à la Section 4.

2 La méthode de SSI par indexation

La Figure 1 présente le schéma de la technique de SSI par indexation introduite dans cet article. Certains des blocs fonctionnels de ce schéma, identiques à ceux décrits dans [10] ne seront pas détaillés. Le présent article se concentre plutôt sur les nouvelles techniques d'analyse des sources et de séparation (blocs 4 et 11 de la Figure 1). Dans cette étude, le processus de mélange (bloc 1) est une simple multiplication d'un vecteur de I signaux sources par une matrice constante $2 \times I$, correspondant à un mélange LIS stéréophonique.

2.1 Décomposition MDCT et groupement moléculaire

Les signaux sources considérés ici sont des instruments/voix-chantée jouant un même morceau de musique (mais enregistrés séparément). Ces différentes sources, non-stationnaires, présentent une grande variabilité temporelle et fréquentielle, et une très forte superposition dans le domaine temporel. Il a été montré dans [15] [2] [10] que les signaux de parole et de musique présentent une parcimonie naturelle dans le domaine TF, impliquant une superposition beaucoup plus faible dans le domaine TF que dans le domaine temporel.

Comme dans [10], la transformée en cosinus discrète modifiée (MDCT) est utilisée pour décomposer les signaux dans le plan TF. Cette transformée est choisie en raison de ses propriétés de concentration de l'énergie du signal, limitant efficacement la superposition des sources. De plus, elle présente l'avantage d'être à coefficients réels et à reconstruction parfaite [12]. Cette transformée est utilisé aux blocs 2, 2' et 8 de la Figure 1 alors que la transformée inverse correspondante (IMDCT) est utilisée aux blocs 6 et 12 pour regénérer les signaux temporels à partir de leur décomposition MDCT. La MDCT étant une transformée linéaire, le problème de séparation de sources demeure linéaire/instantané dans le domaine TF pour chaque bin TF. La MDCT est appliquée sur des fenêtres de taille W=2048 échantillons (46.5ms pour une fréquence d'échantillonnage f_e =

44.1kHz), avec une superposition de 50% entre deux fenêtres successives. La matrice MDCT obtenue à la suite de la décomposition d'un signal x est alors $\mathcal{M}_x = \{X(f,t)\}$ de taille 1024 canaux fréquentiels (notés f) par $L/1024$ canaux temporels (notés t), où L est le nombre d'échantillons total de x. La taille de fenêtre W est choisie pour suivre la dynamique de signaux audio, tout en respectant une résolution fréquentielle adaptée à la séparation.

Comme dans [10], l'information extraite des signaux sources est insérée dans le signal de mélange par une technique de tatouage des coefficients MDCT (voir Section 2.2). La capacité d'insertion de l'information d'un seul coefficient MDCT étant trop faible pour contenir l'ensemble du message à tatouer, des coefficients voisins dans le plan TF sont groupés en *molécules*. Une molécule M^x_{pq} est une $F \times T$ sous-matrice de coefficients MDCT localisée aux coordonnées *moléculaires* (p,q) dans le plan TF (voir Figure 2). Dans le présent article, la dimension d'une molécule est 1×4. Les étapes d'analyse de la prédominance des signaux, du tatouage et de la séparation sont toutes réalisées à l'échelle d'une molécule de coefficients MDCT.

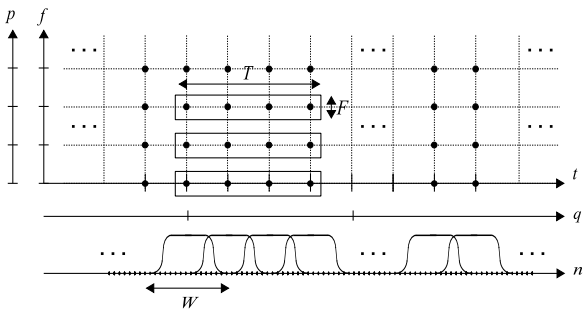

Figure 2 – Représentation schématique de la décomposition temps-fréquence et du groupement moléculaire.

2.2 Le procédé de tatouage

La technique de tatouage utilisée dans ces travaux est identique à celle introduite dans [10] [9]. Elle est inspirée de la Quantization Index Modulation (QIM) de [4], adaptée aux coefficients MDCT. Succinctement, le principe est le suivant : le message inséré est porté par une quantification des coefficients MDCT du mélange à l'aide d'un quantificateur spécifique dont les sous-niveaux sont associés avec les valeurs de la watermark. La capacité d'insertion de l'information résulte d'un traitement conjoint : la maximisation d'un pas de quantification de référence sous contrainte d'inaudibilité, et la minimisation d'un sous-pas de quantification sous contrainte de robustesse à un bruit de quantification. Comme nous ciblons l'application CD-audio pour notre système, le bruit de quantification considéré résulte de la quantification linéaire 16 bits appliquée au signal de mélange tatoué lors de la conversion au format CD-audio (bloc 7 de la Figure 1). Le système proposé est conçu de manière à ce que la quantification des coefficients MDCT au codeur (bloc 5) et au décodeur (bloc 10) fournisse le même résultat. Cette technique de quantification a montré qu'elle assurait une capacité d'insertion de l'information conséquente, allant jusqu'à 150kb/s pour des signaux de musique [10]. Une telle capacité est largement suffisante pour insérer l'information utilisée dans la présente étude (voir Sections 2.3 et 4). Une version améliorée de ce système, incluant un modèle psychoacoustique et permettant des débits moyens allant jusqu'à 250 kbits/s, est présentée dans le même congrès [11].

2.3 Analyse de la contribution des sources

Contrairement aux descripteurs utilisés dans [10] [9] 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 par indexation est beaucoup plus simple. Elle n'en reste pas moins fondamentale pour permettre la séparation. Une analyse des signaux sources, réalisée au bloc 4 de la Figure 1, permet de déterminer la participation locale de chaque source au signal de mélange. L'information extraite, pour chaque molécule du plan TF, de cette analyse est le nombre de sources prédominantes sur cette molécule (*i.e.* les sources dont l'énergie est très supérieure à celle des autres sources), et l'index de ces sources parmi l'ensemble des I sources. Étant donné la parcimonie des signaux audio dans le plan TF, seuls trois cas de figure sont considérés : 0) aucune source n'est présente, 1) seulement une source est présente, comme par exemple dans [15], ou 2) deux sources sont présentes, comme par exemple dans [2]. Le mélange possédant deux voies, les cas 1) et 2) correspondent à des configurations (sur-)déterminées et peuvent être traités comme décrit à la Section 2.4.

Le cas 0 s'explique par la parcimonie des signaux de musique : la plupart des coefficients de la décomposition MDCT des signaux sont en effet d'amplitude proche de zéro. La faible énergie du mélange dans ces portions du plan TF a deux conséquences majeures. Tout d'abord, ces zones du plan TF sont de faible importance pour la qualité audio globale du signal, c'est pourquoi il n'y apparaît pas nécessaire de procéder à la séparation. De plus, l'amplitude de la décomposition MDCT du mélange dans ces portions du plan TF étant très faible, il en va de même de la capacité d'insertion de l'information, trop faible pour permettre de procéder au tatouage (cf [10]). Pour cette raison, un prétraitement de la présente méthode consiste en un seuillage des signaux de mélange : seules les molécules d'énergie suffisante sont considérées comme pertinentes pour un plus ample traitement. Les molécules inférieures à ce seuil énergétique sont considérées comme un résidu traité séparément.

Pour traiter les cas 1 et 2, le ratio énergétique suivant est défini et calculé pour chaque molécule $M^{s_i}_{pq}$ de chaque signal source s_i :

$$R_i(p,q) = \frac{\sum\limits_{(f,t)\in\{P\times Q\}} |S_i(f,t)|^2}{\sum_{j\neq i} \sum\limits_{(f,t)\in\{P\times Q\}} |S_j(f,t)|^2} \qquad (1)$$

où $P \times Q = [(p-1)F, pF-1] \times [(q-1)T, qT-1]$. Pour chaque molécule, les deux sources de plus forts ratios, par exemple s_k et s_m, sont sélectionnées. Enfin, le cas 2 est ramené au cas 1 si $R_m(p,q) < \varepsilon R_k(p,q)$ avec ε un

scalaire petit devant 1 (typiquement 0.05). Dans ce cas de figure, seule la source s_k est sélectionnée. Le résultat de cette analyse est encodé dans le tatouage $I(p,q)$ à l'aide d'un nombre limité de bits (en comparaison de l'information utilisée dans [10]), typiquement moins de 4 bits pour encoder les 11 combinaisons possibles dans le cas d'un mélange de 4 signaux sources. Cela représente un débit d'information d'environ 20kb/s pour chaque voie, pour des molécules de taille 1×4. Notons que, si le mélange est localement sous-déterminé, *i.e.* si plus de deux sources sont en fait présentes dans le mélange à la suite de l'étape préliminaire de seuillage, alors seules les deux sources prédominantes sont estimées au décodeur, les autres sources étant considérées comme du bruit.

2.4 Le processus de séparation

La séparation des signaux sources est réalisée à l'échelle des molécules du signal de mélange (bloc 11 de la Figure 1), après que le tatouage $I(p,q)$ ait été extrait et décodé (bloc 10). La matrice $\mathbf{A} = \{a_{ji}\}$, de taille $2 \times I$, est identique pour chaque molécule, et supposée connue au décodeur car de faible dimension et pouvant aisément être transmise au décodeur par tatouage. Pour chaque molécule, trois cas de figure sont à envisager, en fonction de la combinaison de sources $I(p,q)$, correspondant aux trois cas introduits à la Section 2.3. Si aucune source n'est active (cas 0), la molécule de mélange est considérée comme un résidu (dont l'énergie peut être partagée entre les signaux sources estimés, ou qui peut tout simplement être non traitée). Si une seule source est active (cas 1), disons s_k, le mélange stéréo est réduit à[1] :

$$\begin{bmatrix} X_L(p,q) \\ X_R(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} \\ a_{2k} \end{bmatrix} \left[S_k(p,q) \right] \tag{2}$$

Une estimation de $S_k(p,q)$ peut alors être facilement obtenue par :

$$\hat{S}_k(p,q) = \frac{1}{a_{1k}} \tilde{X}_L^W(p,q) \;\text{ ou }\; \hat{S}_k(p,q) = \frac{1}{a_{2k}} \tilde{X}_R^W(p,q) \tag{3}$$

ou une combinaison de ces deux expressions. La notation $\tilde{X}^W$ désigne la version tatouée et convertie au format CD-audio de X telle qu'utilisée au décodeur. Si les deux sources s_k et s_m sont actives (cas 2), le mélange stéréo est réduit à

$$\begin{bmatrix} X_L(p,q) \\ X_R(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} & a_{1m} \\ a_{2k} & a_{2m} \end{bmatrix} \begin{bmatrix} S_k(p,q) \\ S_m(p,q) \end{bmatrix} \tag{4}$$

et les molécules estimées correspondantes sont obtenues par

$$\begin{bmatrix} \hat{S}_k(p,q) \\ \hat{S}_m(p,q) \end{bmatrix} = \begin{bmatrix} a_{1k} & a_{1m} \\ a_{2k} & a_{2m} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{X}_L^W(p,q) \\ \tilde{X}_R^W(p,q) \end{bmatrix} \tag{5}$$

Finalement, les signaux sources temporels sont reconstruits à partir des estimées des molécules MDCT correspondantes par MDCT inverse (bloc 12 de la Figure 1).

[1]Dans les équations qui suivent, $X(p,q)$ représente une molécule MDCT $1 \times T$. Les équations appliquées séparément à chaque coefficient MDCT sont identiques à celles définies pour une molécule entière. Les indices L/R indiquent la voie de gauche, respectivement de droite, du mélange.

3 Expérimentations et Résultats

3.1 Données et mesures

Les résultats de SSI par indexation présentés dans cette section sont effectués sur des signaux de musique échantillonnés à 44100Hz avec des mélanges chant + 3 instruments (basse, batterie, et piano ou guitare[2]). On mélange ainsi deux jeux de 4 signaux sources de 10 secondes jouant en harmonie (les pistes des différentes sources sont issues de morceaux grand public de genre *jazz* et *pop-rock*). Les résultats sont moyennés sur les 24 mélanges correspondant à toutes les permutations possibles des vecteurs colonnes (normalisés) de la matrice de mélange suivante $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 0.93 & 0.80 & 0.60 & 0.37 \\ 0.37 & 0.60 & 0.80 & 0.93 \end{bmatrix} \tag{6}$$

On cherche ainsi à s'affranchir de possibles effets liés à la position des instruments. La durée totale des signaux tests est de 8 minutes.

La qualité des signaux sources estimés est évaluée à la fois par des tests d'écoute subjectifs informels, et des tests objectifs de mesures de performances comme ceux définis dans [13]. Le ratio source-à-distortion (SDR) fournit un critère global de performance de séparation, le ratio source-à-interferences (SIR) mesure le niveau d'interférences des autres sources dans l'estimée d'un signal source, et le ratio source-à-artéfacts (SAR) mesure le niveau d'artéfacts dans l'estimée d'une source. Nous fournissons également le SIR d'entrée de manière à pouvoir quantifier le pouvoir de réjection d'une méthode de séparation en mesurant la différence entre le SIR en entrée et le SIR en sortie de notre système.

Notre méthode est comparée avec la méthode BZ développée dans la configuration sous-déterminée dans [2]. Dans [2], à chaque atome TF, deux signaux sources (parmi 4 dans le cas présent) sont estimés en trouvant la combinaison linéaire de deux vecteurs colonnes de $\mathbf{A}$ conduisant au plus court chemin de l'origine au vecteur de mélange observé $\mathbf{x}$. Par exemple, Figure 3, on considère que le vecteur de mélange $\mathbf{x}$ est une combinaison des sources 1 et 2. Notons cependant qu'une telle méthode géométrique ne fournit pas l'ensemble des combinaisons possibles de sources. Par exemple, si $\mathbf{x}$ est en fait une combinaison des sources 1 et 3, cette méthode retournera soit la combinaison (1,2), soit la combinaison (2,3), toutes deux erronées. Le tatouage inséré dans le signal de mélange en SSI par indexation fournit une solution à ce problème, et ce même dans le cas de vecteurs sources $\mathbf{a}_i$ très proches les uns des autres, cas de figure où, pour la méthode BZ, le risque de mauvaise sélection du couple de sources formant le mélange est accru. Notons que les 24 permutations de $\mathbf{A}$ limitent de possibles artéfacts dus au cas malencontreux où la combinaison "impossible" de la méthode BZ concernerait systématiquement les sources de plus hautes énergies. Notons aussi que la comparaison avec cette dernière méthode n'est effectuée que pour la phase d'estimation des sources, la matrice de mélange étant supposée connue dans tous les cas[3].

[2]Dans toute la suite, on désigne cette quatrième source par *piano*.

[3]Les auteurs de [2] supposent, dans la phase d'estimation des

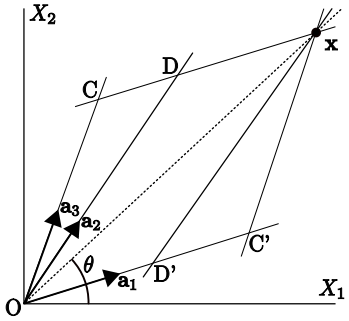

Figure 3 – Méthode géométrique du *plus-court chemin* de l'origine au vecteur de mélange **x** introduit dans [2].

3.2 Mesure de la superposition des sources

La superposition ou non-superposition des signaux sources dans le domaine TF demeure un problème majeur pour les méthodes de séparation basées sur la parcimonie des sources. Dans le but de justifier la pertinence de l'approche proposée sur des signaux de musique, une mesure de la superposition des signaux sources est effectuée. À chaque canal fréquentiel, la superposition des signaux sources est mesurée par le critère de norme l_ε^0 utilisé dans [1] et défini par

$$\| s(f,t) \|_{0,\epsilon(f)} = \text{card}(\{i, |s_i(f,t)| \geq \epsilon(f)\}) \qquad (7)$$

La norme l_ε^0 représente donc le nombre de signaux sources d'amplitude supérieure à ε sur un canal fréquentiel f donné. Ceci permet d'obtenir le pourcentage moyen de trames temporelles pour lesquelles $\|s(f,t)\|_{0,\epsilon(f)} = 0, \cdots, N$, 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 apparait que pour plus de 80% des trames temporelles, seuls deux des quatre signaux sources sont actifs. De plus, une étude plus poussée de la distribution énergétique de chaque source en fonction du rang de son ratio énergétique (1), présentée à la Table 1 montre que 97.1% (pour la batterie) à 99.4% (pour le chant) de l'énergie de chaque signal source correspond aux molécules où cette source est l'une des deux sources les plus énergétiques. Ces résultats justifient la sélection d'au plus deux signaux sources dans le processus de séparation à l'échelle moléculaire. Encore une fois, si trois au quatre sources se superposent, les trois et quatrième sources sont considérées comme du bruit. Si la matrice inverse n'est pas mal dimensionnée, la séparation obtenue en (5) fournit de bons résultats même dans cette configuration "bruitée".

3.3 Résultats de séparation

La Table 2 fournit les résultats moyens de séparation obtenus avec la méthode de SSI par indexation, pour la séparation des vingt-quatre mélanges 2×4 de test. Cette table montre que la méthode SSI par indexation

signaux sources, que la matrice de mélange est connue car supposée parfaitement estimée de manière aveugle.

Rang	Bass	Chant	Batterie	Piano
1	**82.3**	**95.4**	**78.5**	**94.7**
2	**16.0**	**4.0**	**18.6**	**4.3**
3	1.6	0.5	2.7	0.8
4	7.10^{-4}	3.10^{-4}	0.2	0.15

Table 1 – Pourcentage de l'énergie totale des signaux sources en fonction de son rang énergétique dans le mélange. Étude à l'échelle d'une molécule 1×4.

fournit de très bons résultats de séparation pour les quatre signaux sources. Une augmentation du SIR de (-8.5) à (-0.5)dB en entrée à 33.3 à 37.4dB en sortie est obtenue, soit une amélioration du SIR de 36.2 à 44dB. Ces bons résultats sont confirmés par des tests d'écoute qui montrent une très bonne réjection des interférences dues aux autres sources : chaque instrument est clairement isolé. La qualité d'écoute des signaux sources n'est bien évidemment pas parfaite (SDR/SAR dans une fourchette de 9.5 à 14.7dB), et un certain niveau de bruit musical demeure. Cependant, les signaux sources estimés peuvent être ajoutés ou soustraits au signal de mélange (un des post-traitements possibles), et dans ce cas, le bruit musical est largement masqué par le mélange.

La Table 2 fait apparaître l'avantage de la présente méthode de SSI par indexation sur la méthode BZ, et ce pour chacune des mesures de performances, démontrant ainsi l'avantage d'utiliser l'information sur les signaux sources. Les scores de SDR obtenus par SSI par indexation sont de 4.8dB (pour le chant) à 6.3dB (pour le piano) supérieurs à ceux obtenus par BZ, avec une moyenne de 5.4dB d'amélioration. En ce qui concerne le SAR, l'amélioration varie entre 3.9dB (pour la basse et le chant) à 4.9dB (pour le piano), avec une moyenne de +4.2dB. La qualité audio des signaux sources estimés est, en conséquence, bien meilleure avec la méthode de SSI par indexation. Quelques exemples de résultats peuvent être téléchargés à l'adresse suivante `http ://www.icp.inpg.fr/~girin/Stereo-ISS-demo.rar`.

Signaux	SIR entrée	SSI par indexation			BZ		
		SDR	SIR	SAR	SDR	SIR	SAR
bass	-5.4	11.5	33.3	11.5	6.6	14.3	7.6
chant	-0.5	14.7	35.7	14.7	9.9	17.7	10.8
batterie	-8.5	9.5	34.2	9.5	3.8	10.9	5.3
piano	-6.6	12.7	37.4	12.7	6.4	12.8	7.8

Table 2 – Performances de séparations moyennées sur 24 mélanges de test (8 minutes de musique).

4 Conclusion

La méthode de SSI par indexation des signaux sources décrite dans cet article n'appartient pas aux méthodes de séparation de sources classiques. Contrairement au cas de méthodes de séparation aveugles, les signaux sources initiaux sont ici disponibles avant qu'ait lieu le mélange, et des applications spécifiques telle que

l'écoute active de CD-audio sont particulièrement visées. Après les résultats prometteurs obtenus dans le cadre d'un signal monophonique dans [10], le présent article fournit des résultats significatifs dans la configuration stéréophonique. La simplicité de l'information utilisée (et de fait la faible ressource nécessaire pour tatouer cette information sur le signal de mélange) comparée à l'approche par codage des signaux sources proposée dans [10] est compensée par une exploitation efficace de la parcimonie des signaux sources, et de l'information spatiale. L'information insérée par tatouage permet de relaxer l'hypothèse, trop forte, d'une seule source active [15] pour l'étendre à l'hypothèse de deux sources prédominantes dans chaque région du plan TF. Une sélection correcte des deux sources prédominantes (qui n'est pas assurée par [2]), permet ainsi un processus d'estimation des sources relativement simple et efficace.

On peut noter que l'oracle que nous avons utilisé dans ce travail pour sélectionner les sources est un oracle "a priori" (calculé avant le mélange), et donc sous-optimal. Dans des travaux futurs il pourra être remplacé par un oracle "a posteriori" (calculé en utilisant le mélange), optimal, conduisant à la meilleure estimation possible des signaux sources sous l'hypothèse de parcimonie [14] [8]. La combinaison de cette approche par indexation avec l'approche par codage des signaux sources de [10] laisse entrevoir une suite logique à notre travail : un système hybride qui permettrait la séparation d'un grand nombre de signaux sources à partir d'un mélange stéréophonique. Par exemple, si quatre sources sont localement prédominantes (et donc se superposent) dans un mélange de, disons, huit sources, deux d'entre elles peuvent être extraites par une approche codage, et les deux autres sources peuvent être estimée par la présente méthode après que les deux premières sources décodées aient été soustraites du signal de mélange. De futurs travaux se concentreront également sur des types de mélanges plus complexes et plus réalistes comme des mélanges convolutifs.

Remerciements

Ces travaux sont soutenus par l'Agence Nationale de la Recherche française (ANR) dans le cadre du projet DReaM (ANR 09 CORD 006).

Références

[1] S. Araki, H. Sawada, and S. Makino. *K-means Based Underdetermined Blind Speech Separation*, pages 243–270. in S. Makino and al. (Eds), Blind Source Separation, Springer, 2007.

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

[3] J.F. Cardoso. Blind signal separation : Statistical principles. *Proc. IEEE*, 9(10) :2009–2025, 1998.

[4] B. Chen and G. Wornell. Quantization index modulation : A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inform. Theory*, 47(4) :1423–1443, 2001.

[5] P. Comon and C. Jutten. *Séparation de sources - Au-delà de l'aveugle et applications.* Hermès-Lavoisier, 2007.

[6] P. Comon and C. Jutten. *Séparation de sources - Concepts de base et analyse en composantes indépendantes.* Hermès-Lavoisier, 2007.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley & Sons, 2001.

[8] A. Nesbit and M. Plumbley. Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation. In *ICASSP*, 2008.

[9] 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)*, pages 101–104, 2009.

[10] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Trans. Audio, Speech, and Language Process.*, 2010. accepted.

[11] J. Pinel, L. Girin, and C. Baras. Une technique de tatouage "haute-capacité" pour signaux musicaux au format cd-audio. In *Actes du Congrès Français d'Acoustique*, 2010.

[12] John P. Princen and Alan B. Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. In *IEEE Trans. Acoust. Speech Sig. Proc.*, volume 64, pages 1153–1161, 1986.

[13] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech Audio Process.*, 14(4) :1462–1469, 2005.

[14] E. Vincent, R. Gribonval, and M.D. Plumbleys. Oracle estimators for the benchùarking of source separation algorithms. *Signal Processing*, 87(2007) :1933–1950, 2007.

[15] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.*, 52(7) :1830–1847, 2004.