Estimation robuste et apprentissage aveugle de modèles pour la séparation de sources sonores

Simon Arberet

N° d'ordre: 3832

THÈSE

Présentée devant

devant l'Université de Rennes 1

pour obtenir

le grade de : Docteur de l'Université de Rennes 1
Mention Traitement du Signal et Télécommunications

par

Simon ARBERET

Équipe d'accueil : METISS/IRISA
École Doctorale : Matisse
Composante universitaire : SPM

Titre de la thèse :

Estimation robuste et apprentissage aveugle de modèles pour la séparation de sources sonores

soutenue le 17 decembre 2008 devant la commission d'examen

M. :	Jean-Jacques	FUCHS	Président
MM. :	Manuel	DAVY	Rapporteurs
	Yannick	DEVILLE	
MM. :	Rémi	GRIBONVAL	Directeurs de thèse
	Frédéric	BIMBOT	

Remerciements

Je tiens d'abord à remercier les membres du jury, en particulier Frédéric Bimbot, Directeur de Recherches CNRS et Rémi Gribonval, Chargé de Recherches INRIA, qui m'ont encadré avec brio pendant ces trois années à l'IRISA (INRIA Rennes-Bretagne Atlantique), Manuel Davy, Chargé de Recherches CNRS au LAGIS de Lille et Yannick Deville, Professeur à l'Université de Toulouse 3, pour l'attention qu'ils ont portée à mes travaux en acceptant la tâche de rapporteurs. Je remercie également Jean-Jacques Fuchs, Professeur à l'Université de Rennes 1 pour avoir accepté de présider ce jury.

J'adresse un très grand merci à Alexey Ozerov et Emmanuel Vincent avec qui j'ai collaboré à différentes reprises et qui m'ont éclairé sur leurs travaux lors de multiples discussions scientifiques. Je remercie également Pierre Vandergheynst qui m'a accueilli pour un court séjour dans son équipe de l'EPFL ainsi que Anna Llagostera avec qui j'ai travaillé à cette occasion sur des aspects multimodaux de la séparation de sources, Sofi Chow qui a accepté que je l'encadre pour un stage, Ayush Bhandari pour ses conseils et son amitié. Merci bien entendu à mes autres collègues de l'équipe METISS de l'IRISA, Guillaume Gravier, Sylvain Lesage, Boris Mailhé, Gilles Gonon, Sacha Krstulovic, Daniel Moraru, Benjamin Roy, Mathieu Ben, Sylvain Busson, Prasad Sudhakar, Armando Muscariello, Pierre Cauchy, Stéphanie Lemaile, qui m'ont supporté et soutenu à un moment ou à un autre, ainsi qu'à mes amis et à ma famille.

Je conclus en remerciant à nouveau Rémi Gribonval pour ses qualités scientifiques et didactiques éminentes ainsi que la pertinence de ses critiques.

Table des matières

Notations

Notations mathématiques usuelles

$\Re\mathrm{e}$	Partie réelle
$\Im\mathrm{m}$	Partie imaginaire
$\mathbb{N}, \mathbb{N}^*$	Ensemble des entiers naturels, des entiers strictement positifs
$\mathbb{R}, \mathbb{R}_+$	Ensembles des réels et des réels positifs
$\mathbb{C}$	Ensemble des complexes
$P, \mathbb{E}$	Probabilité et espérance

Matrices et vecteurs

$\mathbf{A}$	Matrice
$\mathbf{a}$	Vecteur
$\mathbf{I}_M$	Matrice identité de dimension $M \times M$
$\mathbf{A^{-1}}$	Matrice inverse de la matrice $\mathbf{A}$
$\mathbf{A}^\dagger$	Matrice pseudo-inverse de la matrice $\mathbf{A}$
$\mathbf{A^T}$	Matrice transposée de la matrice $\mathbf{A}$
$\mathbf{A^H}$	Matrice transconjuguée de la matrice $\mathbf{A}$
$\det(\mathbf{A})$	Déterminant de la matrice $\mathbf{A}$
$\mathrm{diag}(\mathbf{a})$	Matrice diagonale dont la diagonale est le vecteur $\mathbf{a}$
$\mathrm{diag}(\mathbf{A})$	Matrice diagonale ayant la même diagonale que la matrice $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	Trace de la matrice $\mathbf{A}$

Notations particulières utilisées

t	Indice du temps
f	Indice de fréquence
$\mathcal{T}$	Fiabilité

Mélange

N	Nombre de sources
M	Nombre de canaux d'observation

$s_n(t) \in \mathbb{R}$	n-ième source à l'instant t
$x_m(t) \in \mathbb{R}$	m-ième observation à l'instant t
$\mathbf{s}(t) \in \mathbb{R}^N$	Vecteur des sources à l'instant t
$\mathbf{x}(t) \in \mathbb{R}^M$	Vecteur des observations du mélange à l'instant t
$\mathbf{S}(t,f) \in \mathbb{C}^N$	TFCT des sources
$\mathbf{X}(t,f) \in \mathbb{C}^M$	TFCT des observations du mélange
$S_n(t,f) \in \mathbb{C}$	TFCT de la n-ième source
$X_m(t,f) \in \mathbb{C}$	TFCT de la m-ième observation du mélange

Mélange instantané

$\mathbf{A} \in \mathbb{R}^{M,N}$	Matrice de mélange
a_{mn}	Gain du mélange de la n-ième source sur le m-iène canal d'observation
$\mathbf{a_n} \in \mathbb{R}^M$	n-ième colonne de la matrice de mélange $\mathbf{A}$

Mélange convolutif et anéchoïque

$\mathbf{A}(f) \in \mathbb{C}^{M,N}$	Matrice de mélange
$\mathbf{a_n}(f) \in \mathbb{C}^M$	n-ième colonne de la matrice de mélange $\mathbf{A}(f)$
δ_{mn}	Délai correspondant au chemin pris par la n-ième source pour atteindre le m-ième capteur
δ_n	Différence de temps d'arrivée de la n-ième source (cas ou $M = 2$)

Introduction

> Quand le soir approchait je descendais des cimes de l'île et j'allais volontiers m'asseoir au bord du lac sur la grève dans quelque asile caché ; là le bruit des vagues et l'agitation de l'eau fixant mes sens et chassant de mon âme toute autre agitation la plongeaient dans une rêverie délicieuse où la nuit me surprenait souvent sans que je m'en fusse aperçu. Le flux et reflux de cette eau, son bruit continu mais renflé par intervalles frappant sans relâche mon oreille et mes yeux, suppléaient aux mouvements internes que la rêverie éteignait en moi et suffisaient pour me faire sentir avec plaisir mon existence sans prendre la peine de penser.

JJ. Rouseau *Les Rêveries du promeneur solitaire* (posth. 1782), cinquième promenade.

J.J. Rousseau (JJR), dans cet extrait autobiographique des *Rêveries du promeneur solitaire*, analyse la scène sonore l'environnant, et identifie les sources sonores que sont « les vagues », « l'agitation de l'eau », « le flux et le reflux de cette eau ». Les représentations de ces objets sonores dans son esprit ont pour effet de le plonger dans une « rêverie délicieuse ». Cependant, si par malheur un ours venait à s'approcher, JJR serait immédiatement alerté par son système cognitif, de la présence du féroce animal. En effet, à partir des sons, du rugissement et du craquement de branches, produits par le dangereux mammifère, JJR vraisemblablement :

1. Analyserait la présence d'une nouvelle source sonore, qu'il chercherait à catégoriser parmi les objets sonores stockés dans sa mémoire.
2. Analyserait la direction de la nouvelle source sonore, à partir du signal bi-canal fourni par ses deux oreilles.
3. Orienterait sa tête dans la direction présumée de l'animal, et recommencerait un processus de localisation de la source, à partir d'une part des nouvelles informations auditives dues à la nouvelle orientation de sa tête, et d'autre part des informations visuelles corrélées aux informations auditives.

Mais comment fait-il pour localiser et se focaliser sur les sources de la scène sonore, alors que celles-ci sont mélangées lorsqu'elles parviennent à son cerveau ?

En effet, le signal reçu par le cerveau est le résultat de l'addition et de la propagation des sons de chacune des sources à chacun des capteurs que sont les oreilles. De plus, avant d'arriver aux oreilles, le son émis par les sources emprunte une multitude de chemins,

dus aux possibles réflexions des sons. Ces réflexions sont spécifiques à l'espace acoustique dans lequel se déroule la scène sonore. Cependant si l'on néglige les effets de réflexion des sons qui sont d'ailleurs moindres dans un espace extérieur, et que l'on considère uniquement le trajet direct des sources aux capteurs, alors on pourra remarquer que chacune des sources sonores présente des indices de localisation caractéristiques. En effet, à part cas particulier, la distance d'une source aux capteurs est différente d'un capteur à l'autre. Par conséquent, les temps de parcours, et donc les temps d'arrivée ainsi que les amplitudes des signaux reçus, sont également différents d'un capteur à l'autre. Ces indices acoustiques de localisation, caractéristiques de chacune des sources sonores, sont appelés : ITD (*Interaural Time Difference*) pour la différence de temps d'arrivée, et ILD (*Interaural Level Difference*) pour la différence de niveau d'amplitude.

Il existe également une deuxième catégorie d'indices de localisation, mais qui sont spécifiques au système de captation. Il s'agit des indices spectraux dus à la résonance et à la réflexion des sons sur le système de captation. Dans le cas du système d'écoute de l'humain, il s'agit principalement des interactions des sons avec la tête et les pavillons auditifs. Ces interactions se manifestent par une fonction de transfert, dite HRTF (*head-related transfer function*), qui est spécifique à chaque direction de l'espace. Ces indices spectraux permettent à l'humain notamment d'éviter la confusion entre l'avant et l'arrière, ainsi que d'estimer l'élévation de la source, ce que ne permettent pas de faire les indices ITD et ILD quand il n'y a que deux capteurs. On sait que les indices spectraux sont très utilisés par le système cognitif humain, mais malheureusement, pour pouvoir les exploiter, il est nécessaire de connaître la fonction de transfert HRTF qui est spécifique à chaque système de captation.

Une troisième possibilité permettant à l'humain d'affiner la localisation d'une source, consiste à orienter sa tête pendant le processus de captation, afin d'introduire des informations complémentaires de localisation éventuellement plus précises. Enfin une quatrième piste consiste à exploiter l'information multimodale disponible, dans notre exemple la vision. Cette dernière nous permettrait sûrement d'évaluer la position de l'animal, d'une façon plus précise que l'analyse auditive seule.

Comme nous l'avons déjà mentionné, la scène sonore est généralement composée de plusieurs sources sonores qui sont mélangées dans le signal reçu par le système de captation. En plus d'entendre les sons émis par notre ours, JJR perçoit les sons du ressac, des oiseaux, ainsi que probablement une multitude d'autres sources plus ou moins intenses. Cela a pour conséquence de compliquer la tâche d'identification, de localisation, et de toute autre analyse de notre source d'intérêt.

Comment l'humain fait-il alors pour se focaliser sur une unique source ?

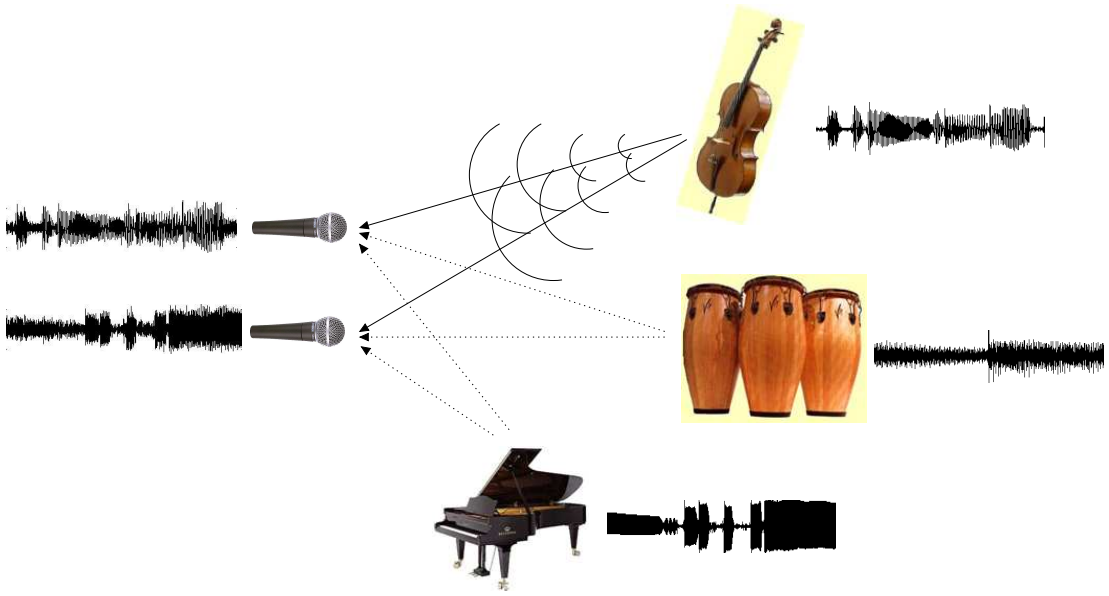
Tout d'abord, considérons le son émis par chacune des sources comme une combinaison d'atomes sonores, c'est-à-dire de petits morceaux de sons qui additionnés ensembles reconstitueraient le son dans son entier. Le son capté par l'auditeur serait alors le résultat de l'addition et de la propagation de l'ensemble des atomes sonores émis par chacune des sources. Au cours du processus de captation, certains atomes, peut être la majorité, se seront mélangés avec des atomes émis par d'autres sources. Cependant une partie

des atomes sonores ne se seront sans doute peu ou pas mélangés avec les atomes des autres sources. Ces atomes en question auront ainsi conservé les informations spatiales propres à la source qui les a émis. Ainsi, si l'on observe les indices de localisations de chacun des atomes reçus, on remarquera qu'un certain nombre d'atomes, précisément ceux qui ne se seront pas mélangés aux autres, auront les mêmes indices, ou des indices très proches, tandis que les autres auront des indices dont la valeur semble être aléatoire. Par conséquent, l'observation des indices de localisation de l'ensemble des atomes reçus présentera des amas correspondant aux directions des sources de la scène sonore. À une source donnée correspond donc un amas. Cependant, d'une part certains atomes d'une source peuvent s'être mélangés avec d'autres atomes et ainsi ne sont pas présents dans l'amas en question. D'autre part certains atomes d'une autre source peuvent s'être mélangés avec d'autres atomes et être présent dans l'amas en question. Par conséquent, si l'on additionne les atomes présents dans un amas, on reconstruit la source sonore d'une façon partielle et bruitée.

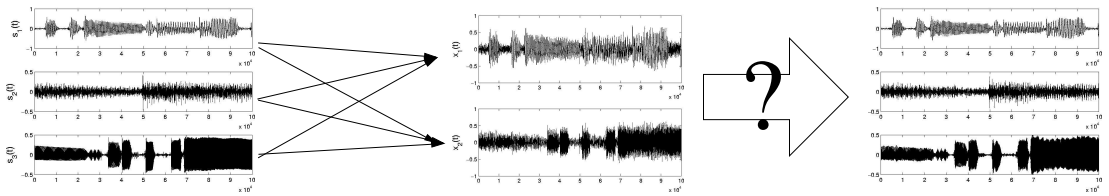
Nous venons de montrer comment les informations de nature spatiale pouvaient nous aider à discriminer les sources de la scène sonore mélangées lors du processus de captation. Nous avons aussi vu que l'information spatiale à elle seule est insuffisante pour reconstituer parfaitement les sources de scène sonore. Même si le système cognitif humain n'effectue pas une séparation parfaite des sources, il est capable de se focaliser sur une source et d'en extraire des informations de haut niveau, comme la reconnaissance de l'objet sonore, et éventuellement son contenu sémantique (le sens d'une phrase si par exemple la source sonore est un locuteur). Ces informations de haut niveau permettent également d'affiner la focalisation car elles permettent de désambigüiser les confusions dues aux mélanges des sources. Ces informations de haut niveau nécessitent évidemment d'avoir des connaissances sur les objets de la scène sonore. Il est évident que JJR avait une idée du son produit par l'agitation de l'eau, le son produit par le chant des oiseaux, ainsi que celui du craquement d'une branche et du rugissement de l'ours. Si bien même il ne connaissait pas le son exact du rugissement de l'ours, il était sûrement au courant qu'il pouvait croiser le chemin d'un ours ou simplement celui d'un gros animal capable de rugir dans ces contrées helvétiques. Il avait également sûrement une idée de la structure rythmique du son que produit le déplacement d'un tel animal. Il semble donc évident que les connaissances *a priori* sur les sources sonores jouent un rôle considérable dans l'analyse de la scène auditive. Il est néanmoins tout à fait possible qu'un humain soit face à un objet sonore inouï. En effet, JJR pourrait entendre un animal dont il ne connait pas le son, et dont il ignore tout de sa présence en ces lieux. A ce moment là, l'auditeur arriverait tout de même à se focaliser sur la source sonore inconnue, mais ne parviendrait pas à l'identifier. Il mémorisera un temps le son produit par cette source, et le jour où il verra l'animal de ses yeux en train de produire le son en question, il associera enfin le son à l'animal. Si donc les connaissances *a priori* sur les sources sonores jouent un rôle considérable dans l'analyse de la scène auditive, il en est de même du processus d'apprentissage de nouvelles sources sonores, qui permet alors au système d'analyse de la scène auditive de s'adapter à une infinité de scènes sonores. En particulier, si on n'a aucune connaissance *a priori* sur la scène sonore, et que l'on souhaite séparer, c'est-à-dire isoler, l'ensemble des sources de la scène sonore, on parle de séparation aveugle de sources sonores.

Le problème de séparation de sources :

Dans le cadre de cette thèse, nous nous intéressons à l'analyse de scènes sonores, et en particulier à la séparation aveugle de sources appliquée aux signaux audio. Nous considérons la configuration suivante : une scène est composée de sources sonores et observée par l'intermédiaire de capteurs, typiquement des microphones. Le signal observé est un mélange, résultat de la propagation et de l'addition des ondes de chacune des sources jusqu'aux capteurs. Le problème de la séparation de sources consiste à extraire chacun des signaux sources à partir de l'observation du mélange de ces sources. La figure 1 illustre la problématique de la séparation de sources dans le cas d'un mélange de trois sources captées par deux microphones.



(a) scène sonore composée de 3 sources captées à l'aide de 2 microphones.



(b) La séparation des 3 sources à partir du mélange sur 2 canaux

FIG. 1 – La problématique de la séparation de sources

Il existe un grand nombre de méthodes de séparation de sources, adaptées à des problématiques différentes. Ces problématiques se caractérisent par des hypothèses sur le mélange et sur les sources qui varient suivant l'application et le degré de réalisme que l'on souhaite atteindre. Ainsi, la difficulté du problème de séparation de sources audio (SSA) varie selon :

1. **la nature du mélange** : c'est à dire les hypothèses que l'on fait sur l'environnement acoustique dans lequel l'enregistrement est effectué. En particulier, les enregistrements qui ont lieu dans un milieu naturel sont affectés par de multiples échos qui ont pour conséquence d'introduire des directions multiples pour chacune des sources (voir l'exemple de la figure 2(a)). Si l'enregistrement est effectué dans une salle anéchoïque comme sur l'exemple de la figure 1(a), alors on peut supposer qu'il n'y a plus d'écho, par contre les distances d'une source à chaque microphone étant différentes, les signaux des sources n'arrivent pas au même instant aux différents microphones. Les sources sonores peuvent aussi avoir été enregistrées séparement, comme cela se fait souvent en studio d'enregistrement, puis mélangées artificiellement via une console de mixage à l'aide de potentiomètres. C'est la cas de l'exemple de la figure 2(b). La console de mixage permet alors d'ajuster l'intensité de chacune des sources via un potentiomètre de gain, ainsi que la balance des sources entre les canaux de sortie via un potentiomètre de panoramique.

   On définit ainsi trois hypothèses sur le mélange qui correspondent aux trois scénarios précédents :

   (a) La plus simple est le cas du mélange *linéaire instantané*, dans lequel les sources sont uniquement caractérisées par une différence de niveau d'intensité entre les canaux.

   (b) Une généralisation du cas instantané est le cas *anéchoïque*, dans lequel on considère que les sources arrivent à chaque capteur à des instants différents, si bien que chaque source est caractérisée par un délai entre les canaux.

   (c) Le cas anéchoïque peut lui même être généralisé au cas *convolutif*, dans lequel on considère des chemins multiples entre chaque source et chaque capteur.

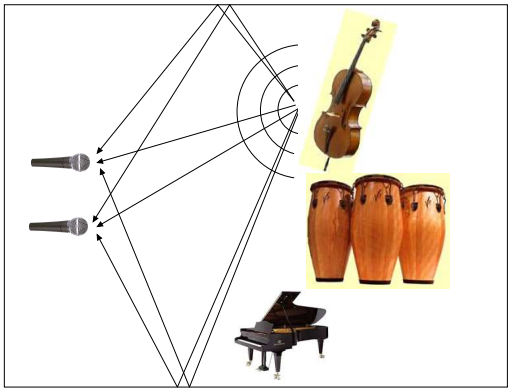
2. **la détermination du mélange** : c'est à dire le rapport entre le nombre de sources et le nombre de capteurs. Si le nombre de sources est égal au nombre de capteurs, ont parlera de mélange *déterminé*. Si le nombre de sources est inférieur au nombre de capteurs, on parlera de mélange *sur-déterminé*. Enfin si le nombre de sources est supérieur au nombre de capteurs, on parlera de mélange *sous-déterminé*.

3. **le niveau de connaissance *a priori*** que l'on possède sur les sources et le mélange. Il est plus simple d'estimer le mélange, si l'on connait le nombre de sources, l'emplacement physique des sources et des capteurs, ou bien les caractéristiques et la position relative des capteurs (cas du HRTF). De même un certain nombre de connaissances *a priori* sur les sources, comme par exemple, des informations de nature spectrale sur les sources, permettent de résoudre le problème de SSA de façon plus efficace.

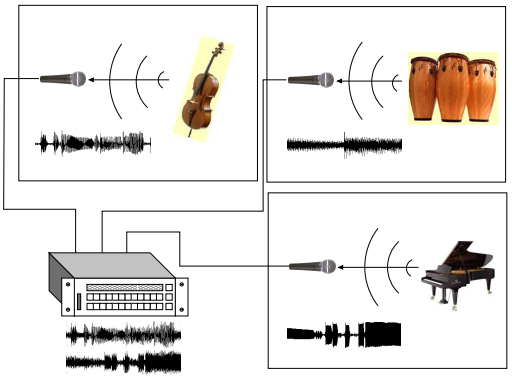
Afin d'introduire la problématique de la thèse, nous allons présenter les approches « standard » utilisées pour la séparation de sources, ainsi que leurs limites.

Résolution du problème dans les cas triviaux (inversibles) :

Le cas le plus simple et le plus largement traité dans le littérature est le cas du mélange linéaire instantané, dans lequel les observations issues des M capteurs s'écrivent



(a) scène sonore enregistrée dans une salle à l'aide de 2 microphones. Nous avons représenté uniquement les premières réflexions du son émis par la première source sur les paroies latérales.



(b) scène sonore où les sources sont isolées dans différentes salles et mélangées à l'aide d'une console de mixage.

FIG. 2 – Enregistrement sur 2 canaux d'une scène sonore composées de 3 sources

comme une combinaison linéaire des N sources : $x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t), 1 \leq m \leq M$. Cette relation entre les sources s_n et les observations x_m peut s'écrire sous la forme matricielle $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$.

Si le mélange est dit déterminé, c'est à dire que le nombre de sources est égal au nombre d'observations ($N = M$), et que l'on connait la matrice de mélange $\mathbf{A}$, alors l'estimation des sources est triviale puisqu'il suffit d'inverser la matrice $\mathbf{A}$ (à condition que celle-ci soit inversible, ce qui est assuré si les sources ont des directions différentes) pour retrouver les sources à une permutation et un facteur d'échelle près.

Si le mélange est sur-déterminé ($N \leq M$), on peut estimer les sources avec la pseudo-inverse de la matrice de mélange : $\widehat{\mathbf{s}}(t) = \mathbf{A}^{\dagger}\mathbf{x}(t)$.

Dans les cas déterminés et sur-déterminés, l'estimation des sources est équivalente à l'estimation de la matrice de mélange. C'est à partir de cette problématique que s'est développée l'*Analyse en Composantes Indépendantes* (ICA), qui fait l'hypothèse que les sources sont indépendantes (et non gaussiennes).

Mélanges sous-déterminés :

Si l'on est dans le cas sous-déterminé ($N > M$), c'est à dire qu'il y a plus de sources que d'observations, le nombre de solutions du système linéaire $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ est infini. Aussi, il n'est pas possible de séparer parfaitement les sources de façon linéaire [GBVF03]. Des hypothèses supplémentaires sur les sources sont nécessaires pour pouvoir fournir des solutions pertinentes. Une hypothèse généralement admise, et que nous allons expliquer aux prochains paragraphes, est la parcimonie des sources. Pour cela, nous allons dans un premier temps introduire la notion de support disjoint des sources.

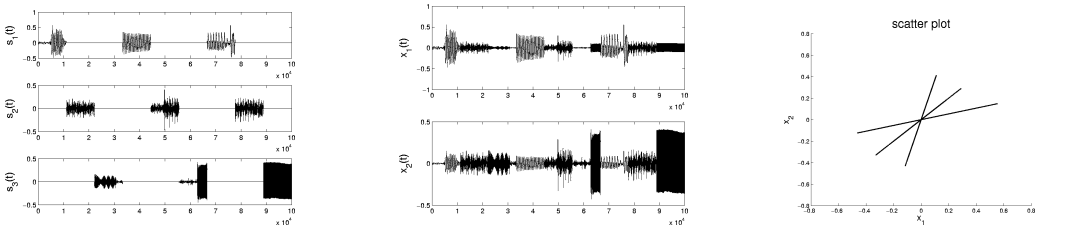
Hypothèse de parcimonie[1] :

Une hypothèse simplificatrice permet de résoudre le problème de SAS dans le cas linéaire instantané sous-déterminé. Il s'agit de l'hypothèse selon laquelle les sources ont des supports disjoints, c'est à dire, qu'à chaque instant t, une seule source est active. Ainsi, pour l'ensemble des instants $t \in \Omega_n$ où la source s_n est la seule source active, on a :

$$\mathbf{x}(t) = \mathbf{a}_n s_n(t) \tag{1}$$

où $\mathbf{a}_n$ désigne la n-ième colonne de la matrice de mélange $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N]$. On en déduit que l'ensemble des points $\mathbf{x}(t), t \in \Omega_n$ doit être aligné le long de la droite passant par l'origine et de vecteur directeur $\mathbf{a}_n$. Par la suite, nous désignons par le terme de *direction* le vecteur $\mathbf{a}_n$. Comme on peut le voir sur l'équation (1), si l'on ne connait ni $\mathbf{a}_n$ ni $s_n(t)$, alors il y a une indétermination d'amplitude entre $\mathbf{a}_n$ et $s_n(t)$. Pour lever cette indétermination nous supposons par la suite que les directions sont normées. Ainsi, dans le cas d'un mélange stéréophonique ($M = 2$), la direction $\mathbf{a}_n$ peut s'écrire $\mathbf{a}_n = [\cos(\theta_n),\ \sin(\theta_n)]^T$ et ne dépend que d'un seul paramètre : θ_n. Pour cette raison, dans le cas stéréophonique nous désignons également par θ_n la direction de la source s_n.

Pour illustrer l'hypothèse du support disjoint des sources, nous représentons les points $\mathbf{x}(t)$ issus d'un mélange stéréophonique ($M = 2$) de trois sources ayant un support temporel disjoint, dans un *diagramme de dispersion* (voir figure 3). Le diagramme de dispersion est la représentation graphique des points $(x_1(t), x_2(t))$ pour l'ensemble des valeurs de t.



(a) sources disjointes dans le domaine temporel (b) mélange des sources (c) diagrammme de dispersion

FIG. 3 – Principe de la séparation de sources basée sur la parcimonie pour des sources à supports temporels disjoints. Le diagramme de dispersion fait clairement apparaître les directions du mélange.

Le diagramme de dispersion de la figure 3 fait clairement apparaître des alignements de points le long de chacune des directions de la matrice de mélange. L'hypothèse des supports disjoints des sources permet donc :

– d'estimer la matrice de mélange $\mathbf{A}$ à partir des points du diagramme de dispersion, Un algorithme de *clustering* sur les directions $\theta(t) = \tan^{-1}(x_2(t)/x_1(t))$ des points

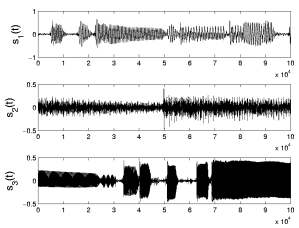
[1] Cette partie est inspiré de [Gri07]

$\mathbf{x}(t)$ permet de facilement estimer la valeur des directions $\mathbf{a}_n = [\cos(\theta_n),\ \sin(\theta_n)]^T$ du mélange.

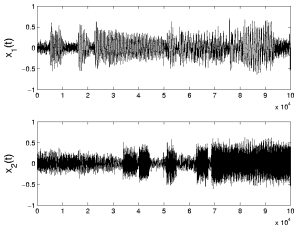
– d'estimer les sources $s_n(t)$ à partir des clusters $\widehat{\Omega}_n$ et de l'estimation de la matrice de mélange $\widehat{\mathbf{A}}$, par la technique du masquage binaire (utilisée par la méthode DUET [YR04]). Si l'étape de clustering s'est bien passé, $\widehat{\Omega}_n \approx \Omega_n$. Si bien que pour estimer la source s_n, il suffit d'inverser le mélange (1) pour $t \in \widehat{\Omega}_n$. Pour les valeur de $t \notin \widehat{\Omega}_n$, la source est supposée à valeur nulle :

$$\widehat{s}_n(t) \stackrel{\triangle}{=} \begin{cases} \widehat{\mathbf{a}}_n^T \mathbf{x}(t) & \text{si } t \in \widehat{\Omega}_n \\ 0 & \text{sinon} \end{cases} \tag{2}$$

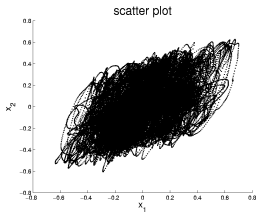
En réalité, les sources audio n'ont que très rarement des supports temporels disjoints. Ainsi, comme on peut le voir sur la figure 4, le diagramme de dispersion du mélange ne fait pas apparaitre les directions du mélange, et la technique ci-dessus ne peut s'appliquer telle quelle.



(a) sources musicales (non-disjointes) dans le domaine temporel



(b) mélange des sources



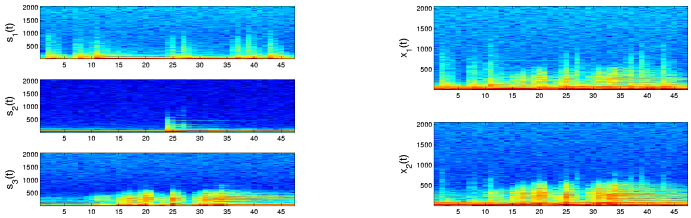
(c) diagrammme de dispersion

FIG. 4 – Principe de la séparation de sources basée sur la parcimonie pour des sources musicales dans le domaine temporel. Le diagramme de dispersion ne fait plus apparaître les directions du mélange.

Cependant bien que l'hypothèse des supports disjoints des sources dans le domaine temporel ne soit pas valide pour les signaux audio, il existe des représentations dites parcimonieuses, pour lesquelles les coefficients des sources audio ont des supports essentiellement disjoints.

La figure 5 montre les représentations temps-fréquence des sources obtenues en prenant la valeur absolue des coefficients d'une Transformée de Fourier à Court Terme (TFCT), ainsi que le diagramme de dispersion des points $(\Re e x_1(t,f), \Re e x_2(t,f))$ représentant la partie réelle des coefficients de la TFCT du mélange. On peut obverver que les directions du mélange sont de nouveau discernables.

Les transformées temps-fréquences comme la TFCT, sont des représentations parcimonieuses pour les signaux audio, c'est à dire des représentations dans lesquelles l'hypothèse de parcimonie est vérifiée. L'hypothèse de parcimonie suppose que pour chaque source, la majorité des coefficients sont nuls ou très petits, tandis que seuls quelques coefficients sont significatifs [Rao98, DE03, KDR99, BZ01, YR04, Gri07]. Etant donnée qu'un faible nombre de coefficients ont une intensité significative, on comprend bien



(a) sources musicales dans le domaine temps-fréquence (b) mélange des sources (c) diagrammme de dispersion

FIG. 5 – Principe de la séparation de sources basée sur la parcimonie pour des sources musicales dans le domaine temps-fréquence (En faisant une TFCT). Les directions du mélanges sont à peu près discernables sur le diagramme de dispersion.

que, si le nombre de sources n'est pas trop élevé, et que les sources ne sont pas trop corrélées, l'hypothèse des supports disjoints des sources puisse être valide.

La matrice de mélange, et les coefficients des sources peuvent donc à nouveau être estimés par l'approche précédente. On peut désormais dresser un schéma fonctionnel standard des méthodes de séparation de sources basées sur la parcimonie (voir figure 6), avec les étapes suivantes :

1. Transformation du mélange dans un domaine parcimonieux ;
2. Estimation de la matrice de mélange par un algorithme de clustering ;
3. Estimation des coefficients des sources dans le domaine transformé ;
4. Reconstruction des sources par la transformation inverse.

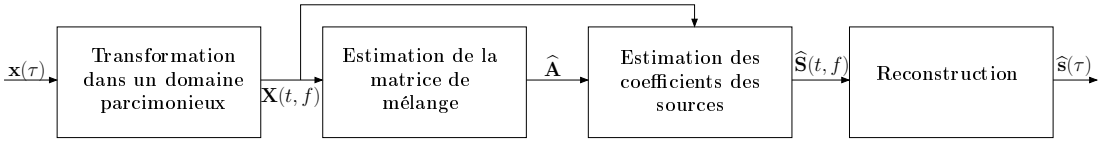


FIG. 6 – Schéma fonctionnel standard des méthodes de séparation de sources basées sur la parcimonie

Limites de l'approche standard et problématique de la thèse

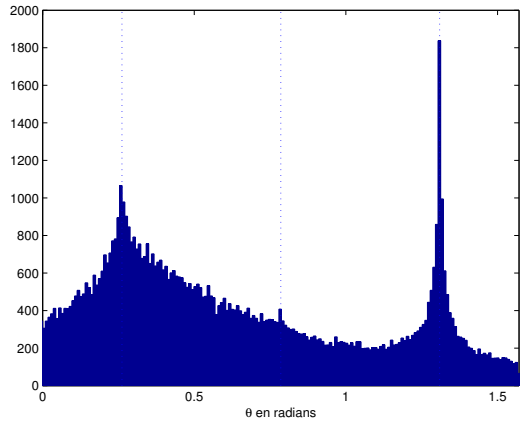
En général, comme on peut le constater sur la figure 9(c), l'hypothèse des supports disjoints des sources, qui permet de faire clairement apparaître les directions du mélange sur le diagramme de dispersion, n'est que rarement vérifiée y compris dans le domaine temps-fréquence. Ainsi on constate que cette hypothèse ne peut être supposée vraie que pour une partie seulement des coefficients du mélange.

Cette constatation a une implication double :

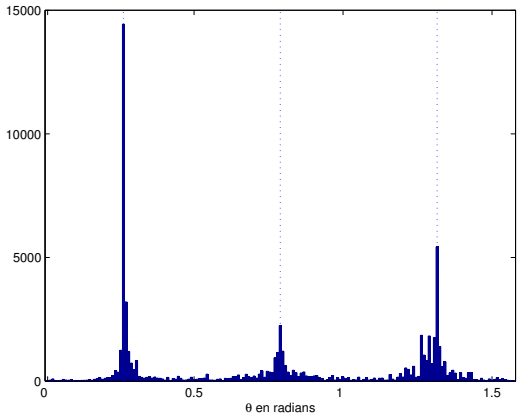
1. D'une part, comme les points du diagramme de dispersion ne sont pas systématiquement alignés sur les directions du mélange, il devient plus délicat d'estimer la matrice de mélange de façon automatique ;
2. D'autre part, même si l'on connait la matrice de mélange, il est difficile d'estimer quelles sont les sources actives, et quelle est la contribution respective de chacune des sources, pour chaque coefficient du mélange.

Limite des méthodes d'estimation des directions du mélange :

Une première approche pour estimer la matrice de mélange consiste à représenter les points du diagramme de dispersion dans un histogramme de directions, et d'estimer les pics de l'histogramme.



(a) Histogramme de directions non pondéré. Certaines directions sont clairement visibles, tandis que d'autres sont très peu discernables.



(b) Histogramme de directions pondéré par l'intensité des points du diagramme de dispersion. Les directions sont plus discernables que dans l'histogramme non pondéré.

FIG. 7 – Histogramme de directions sur l'ensemble des points du diagramme de dispersion. Les vrais directions sont indiquées par des lignes pointillées.

Il est possible par cette méthode de voir apparaitre dans l'histogramme quelques directions des colonnes de la matrice de mélange, mais bien souvent, comme on peu le voir sur la figure 7(a), une partie des directions n'est pas visible.

Cela est dû au fait qu'une grande partie des points ont une faible intensité et ont souvent une direction peu corrélée avec l'une des directions du mélange. Une deuxième approche illustrée par la figure 7(b) (c'est le cas par exemple de la méthode DUET [YR04]), consiste alors à prendre en compte l'intensité des points du diagramme de dispersion, afin d'accroître le rôle joué par les points de grande intensité par rapport aux points d'intensité négligeable. Bien que cette approche fasse davantage apparaître les directions du mélange, certaines directions du mélange restent peu visibles dans l'histogramme du fait que les sources ont des intensités diverses. De plus, comme l'hypothèse des supports disjoints des sources n'est que rarement vérifiée, il existe des points de

grande intensité qui ont une direction ne correspondant à aucune des directions du mélange. Ceci à pour effet de faire apparaitre dans l'histogramme, des pics ne correspondant à aucune direction du mélange. Enfin l'utilisation d'un histogramme pose le problème du choix de la largeur des classes, qui dans le cas de l'histogramme de direction correspond à la résolution angulaire. Une résolution élevée a pour effet de faire apparaître davantage de pics ne correspondant à aucune direction du mélange, tandis qu'une faible résolution ne permet pas d'estimer précisément les directions du mélange, et si deux directions du mélange sont proches, il y a de grandes chances qu'elles soient confondues.

Une troisième approche proposée par Abrard et Deville avec la méthode TIFROM [AD03], et illustrée par la figure 10, consiste à détecter des régions du plan temps-fréquence où une seule source est active, et à estimer les directions du mélange à partir de ces régions. Pour détecter ces régions, la variance des directions des points de chacune des régions est estimée, et celles qui ont les plus petites variances sont sélectionnées. L'avantage de la méthode TIFROM est qu'elle permet de s'affranchir de l'hypothèse des supports disjoints des sources, puisqu'il suffit seulement que pour chaque source, il existe au moins une région du plan temps-fréquence où cette source est la seule active. Cependant la façon de définir si deux estimations de direction correspondent à la même direction du mélange, se fait par un seuillage dur sur la distance entre les directions. La méthode est donc assez sensible à la distance qu'il y a entre les directions du mélange.

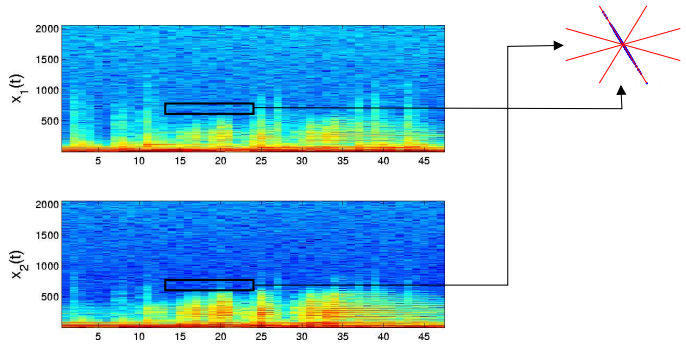


FIG. 8 – Approche locale de Deville. les directions sont estimées à partir du diagramme de dispersion des points contenus dans une région temps-fréquence où une seule source est active.

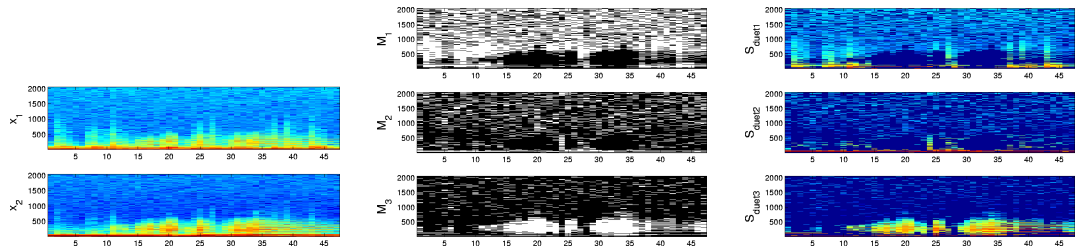
Ainsi, les méthodes existantes exploitent soit le diagramme de dispersion *global* de la figure 5(c) c'est à dire comprenant l'ensemble des points du mélange, soit un ensemble de diagrammes de dispersions *locaux* comprenant les points appartenant à des régions du plan temps-fréquence. La méthode globale a pour principal défaut qu'elle suppose que les sources ont des supports disjoints, et ne permet pas toujours de faire apparaître l'ensemble des directions du mélange. L'approche locale de TIFROM repose sur une hypothèse plus souple, mais pose le problème délicat de la sélection des "bonnes" régions pour l'estimation des directions du mélange. Il s'agit en effet de sélectionner une et une seule région par direction du mélange, avec le risque d'oublier des directions et d'estimer

deux fois la même direction. Enfin la question de l'estimation du nombre de directions nécessaire à une analyse totalement aveugle de la scène sonore a très peu été traitée dans la littérature.

Pour cette raison, les méthodes d'estimation du mélange ne fonctionnent correctement que lorsque le nombre de sources est relativement limité par rapport au nombre de canaux, que chaque source a une énergie (pour les méthodes de type DUET) ou une "visibilité" (pour la méthode TIFROM) suffisante dans le plan temps-fréquence, et que les directions des sources ne sont pas trop proches les unes des autres.

Limites des méthodes spatiales d'estimation des sources :

La seconde étape du problème de séparation de source par la parcimonie est l'estimation des sources à l'aide des directions du mélange.



(a) mélange stéréophonique

(b) masques binaires de chacune des sources, estimés par la méthode DUET

(c) sources estimées après application des masques sur le mélange

FIG. 9 – Principe de l'estimation des sources par la technique du masquage binaire.

La technique du masquage binaire utilisée par la méthode DUET et illustrée sur la figure 9, consiste à estimer la source supposée la plus active en sélectionnant la source dont la direction est la plus corrélée avec l'observation du mélange. Ensuite, la méthode affecte de l'énergie uniquement à cette source. Ainsi des sources qui sont significativement actives peuvent se voir affecter un coefficient nul. De plus il est possible que la source supposée comme étant la plus active ne soit en fait que très faiblement active, voire pas active du tout. Cette source se verra alors affecter de l'énergie alors qu'elle n'était pas active. L'exemple de la figure 10 illustre ce cas de figure.

Ainsi, pour les coefficients du mélange pour lesquels plusieurs sources sont actives, des erreurs d'estimation sont commises. L'erreur globale sur l'estimation des sources sera d'autant plus importante qu'il y aura un nombre important de points temps-fréquence pour lesquels plusieurs sources sont actives simultanément.

Il existe des méthodes d'estimation des sources qui relaxent l'hypothèse des supports disjoints des sources, et supposent qu'un certain nombre de sources peuvent être actives simultanément. C'est le cas des approches par minimisation de norme l_1, qui supposent qu'il y a jusqu'à $J = M$ sources actives. Les sources qui sont considérées actives sont dans le cas où $M = 2$, les deux directions voisines qui « entourent » la direction du mélange. Autrement dit, deux directions non voisines, ne seront jamais considérées comme

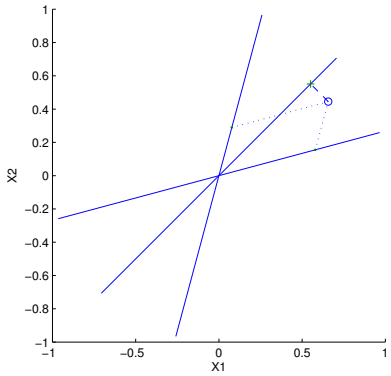
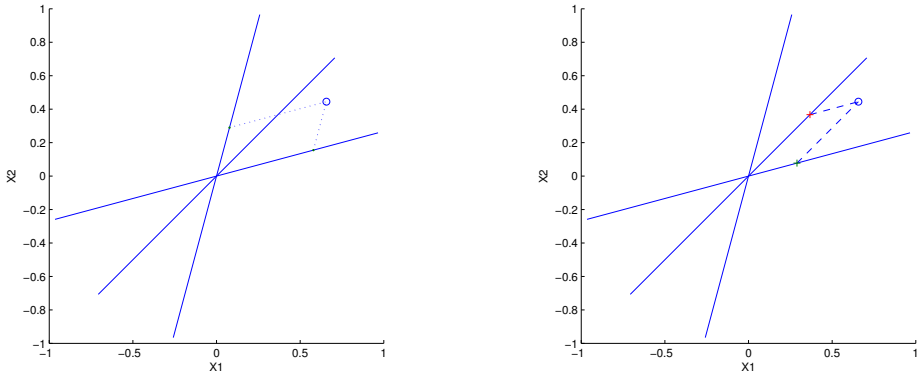


FIG. 10 – Estimation des coefficients des sources par la méthode DUET, pour un point temps-fréquence donné. Le coefficient du mélange ($\circ$) est projeté orthogonalement sur la direction la plus proche (ici la direction du milieu correspondant à la source S_2). La distance du point projeté ($+$) à l'origine est la valeur attribuée à la source S_2, alors qu'une valeur nulle est attribuée aux autres sources. Les valeurs estimées par cette méthode sont les plus vraisemblables si l'on suppose qu'une seule source au plus est active (a une valeur non nulle). Or dans l'exemple ci-dessus, les sources S_1 et S_3 sont actives, tandis que la source S_2 est inactive.

actives simultanément. La figure 11 illustre le principe de l'estimation des coefficients des sources par la minimisation de norme l_1.



(a) Exemple où les sources S_1 et S_3 sont les 2 sources actives.

(b) Estimation des sources par minimisation de norme l_1. Les 2 sources considérées actives sont S_1 et S_2.

FIG. 11 – Principe de l'estimation des coefficients des sources en supposant que 2 sources au plus sont actives, pour un point temps-fréquence donné. Les deux directions qui sont considérées comme actives par la minimisation de norme l_1 sont les deux directions voisines qui « entourent » la direction du mélange.

Cette approche a été utilisée par Bofill et Zibulevsky [BZ01] sur des mélanges audio

stéréophoniques. Cependant les résultats obtenus ne sont pas significativement meilleurs que DUET. La raison étant que cette approche se base uniquement sur l'information spatiale ponctuelle (chaque coefficient étant traité de façon indépendante), laquelle justement est mise en défaut pour les coefficients où l'hypothèse des supports disjoints des sources est également mise en défaut.

Proposition d'une approche spatio-spectrale d'estimation des sources :

Les approches d'estimation des sources basées sur la parcimonie que l'on vient de présenter, exploitent uniquement l'information spatiale ponctuelle. Ponctuelle, car l'estimation des coefficients des sources se fait de façon indépendante pour chaque point temps-fréquence, et spatiale, car les seules informations utilisées sont les directions des sources et le coefficient du mélange. Par conséquent, pour les points temps-fréquence où l'information spatiale est insuffisante parce qu'il y a plusieurs sources actives, il serait sans doute profitable de chercher à exploiter des informations complémentaires, par exemple des informations sur la forme spectrale des sources. De telles informations ont été utilisées avec succès dans le cas de la séparation de source mono-capteur ($M = 1$) [Ben03, Oze06], problématique dans laquelle on ne dispose d'aucune information spatiale.

Nous discutons maintenant de l'aspect architectural des approches de séparation de sources, c'est à dire des différentes façon de décomposer le problème de séparation de sources en tâches élémentaires interconnectées.

Dans le cas multicanal sous-déterminé, la séparation de sources est, comme on vient de le voir, effectuée en deux étapes qui s'appuient sur l'hypothèse de parcimonie des sources. L'utilisation de l'hypothèse de parcimonie revient à exploiter la diversité spatiale des sources. L'étape d'analyse spatiale, a pour tâche d'estimer les directions du mélange à partir de l'observation du mélange. La seconde étape consiste à estimer les sources à partir de l'observation du mélange, mais également des directions du mélange qui ont été estimées lors de l'étape d'analyse spatiale (voir figure 12). Bien que dans les cas de mélanges (sur-)déterminés, il soit possible d'estimer les sources en une seule étape par les approches de l'Analyse en Composantes Indépendantes (ACI), l'approche en deux étapes que nous venons de décrire est toujours valable.

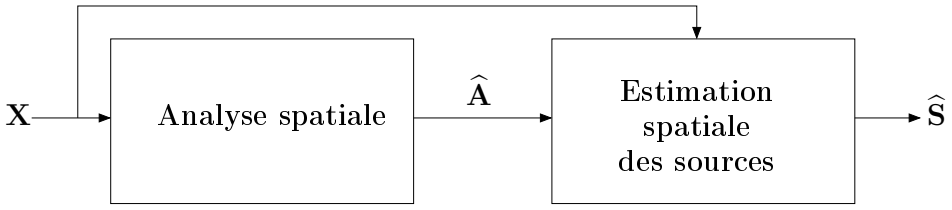


FIG. 12 – Architecture des méthodes de séparation de sources aveugles multicanal basées sur la parcimonie. $\mathbf{X}$ est le mélange mulicanal, $\widehat{\mathbf{A}}$ est l'estimation de la matrice de mélange, $\widehat{\mathbf{S}}$ est l'estimation des sources.

Dans le cas d'un mélange monocanal, où l'information spatiale n'est pas dispo-

nible, des modèles spectraux des sources peuvent être utilisés pour séparer les sources [BB03, OPBG07]. Cependant, cette approche nécessite pour effectuer la séparation, de connaître *a priori* les modèles des sources contenues dans le mélange. Une tâche d'apprentissage des modèles de sources à partir d'exemples ayant des propriétés spectrales similaires aux sources du mélange, doit avoir été effectuée avant la tâche d'estimation spectrale des sources (voir figure 13).

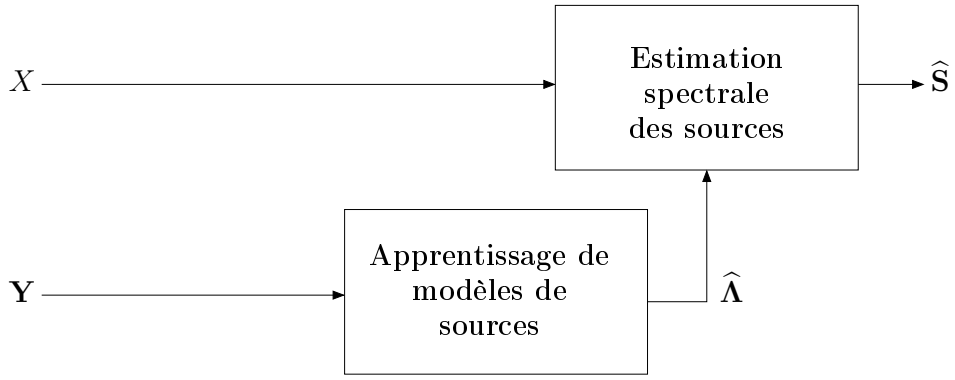


FIG. 13 – Architecture des méthodes de séparation de sources monocanal basées sur des modèles de sources. X est le mélange momocanal, $\mathbf{Y}$ est un ensemble de sources d'apprentissage représentatif des sources contenues dans le mélange, $\widehat{\mathbf{\Lambda}}$ est l'estimation des modèles de sources, $\widehat{\mathbf{S}}$ est l'estimation des sources.

Dans cette thèse, nous étudions la façon de combiner l'approche spatiale basée sur la parcimonie et l'approche spectrale basée sur des modèles de sources, dans une architecture unique afin d'une part de pouvoir exploiter à la fois la diversité spectrale et la diversité spatiale des sources, et d'autre part d'exploiter le caractère aveugle de l'approche spatiale afin d'affranchir l'approche par modèles de sources, des connaissances *a priori*.

Il s'agirait alors d'inférer les modèles spectraux des sources à partir de l'analyse spatiale du mélange (voir figure 14), et d'estimer les sources en prenant en compte les modèles spectraux des sources en plus de l'observation du mélange et de ses directions.

Chacune des tâches composant cette nouvelle architecture est différente de celles des architectures de l'état de l'art illustrées par les figure 12 et 13.

En particulier, un formalisme doit être trouvé afin de pouvoir apprendre les modèles des sources à partir du mélange.

Une des contributions principales de cette thèse est le développement de modèles de la distribution locale du mélange, afin d'une part d'estimer de façon robuste les directions du mélange, et d'autre part d'estimer, en utilisant la connaissance sur les directions du mélange, la distribution des sources en chaque point temps-fréquence.

Plan de la thèse

Cette thèse ce compose de trois parties :

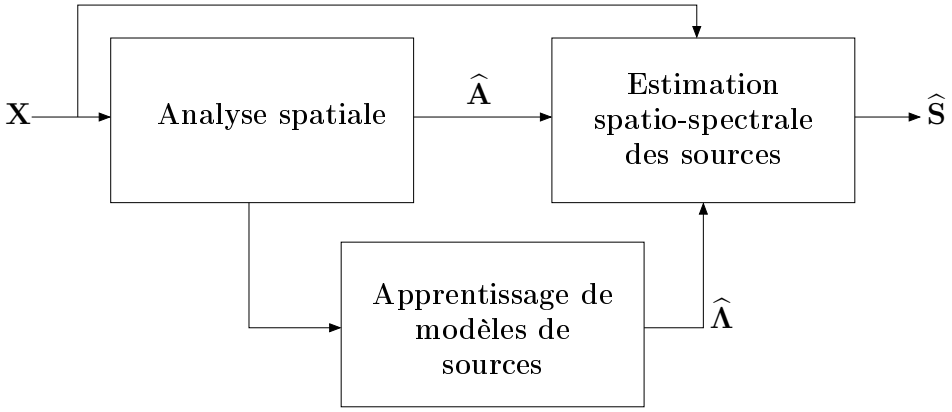


FIG. 14 – Architecture de la méthode de séparation spatio-spectrale que nous proposons. $\mathbf{X}$ est le mélange multicanal, $\widehat{\mathbf{A}}$ est l'estimation de la matrice de mélange, $\widehat{\mathbf{\Lambda}}$ est l'estimation des modèles de sources, $\widehat{\mathbf{S}}$ est l'estimation des sources.

- Etat de l'art ;
- Contributions ;
- Conclusion et Perspectives.

Etat de l'art

Dans cette partie, nous présentons les différentes approches de séparation de sources de l'état de l'art. Cela permet d'introduire les principaux concepts, ainsi que de positionner notre problématique qui sera traitée dans la deuxième partie.

Au chapitre 1, nous introduisons les grandes familles d'approches de séparation de sources, ainsi que les hypothèses sur lesquelles elle se basent. En particulier, nous verrons que l'hypothèse d'indépendance des sources, seule, qui est utilisée par les méthodes de l'Analyse en composantes Indépendante déterminée (ACI-déterminée) n'est pas suffisante pour traiter le cas sous-déterminé qui nous intéresse, mais que celui-ci l'est à condition d'introduire des hypothèses sur la distribution des sources. Si ces distributions sont supposées différentes pour chacune des sources, alors il est nécessaire d'estimer les paramètres de ces distributions en plus des paramètres du mélange, ce qui peut s'avérer difficile voire imposssible dans le cas d'une modélisation gausiennne des sources. Une autre approche consiste à supposer une distribution parcimonieuse des sources identique pour chacune des sources. Cette dernière approche appelée Analyse en Composantes Parcimonieuses permet d'estimer les paramètres du mélange et les sources sans avoir à recourir à une étape d'apprentissage des paramètres de la distribution des sources. La Factorisation en Matrices Non Négative (FMNN) est une autre approche, qui sera briévement présentée dans ce chapitre, mais ne sera pas exploitée dans la suite de cette thèse.

Cette présentation nous permet de discuter du choix d'une architecture en deux étapes pour résoudre le problème de séparation de sources. Les deux étapes étant l'es-

timation des paramètres du mélange suivi de l'estimation des sources.

Au chapitre 2, nous présentons les approches basées sur l'hypothèse de parcimonie des sources pour l'estimation des paramètres du mélange. Nous distinguons l'approche globale, largement utilisée dans l'état de l'art en particulier par la méthode DUET, qui cherche à estimer les paramètres du mélange à partir de l'ensemble des points temps-fréquence, de l'approche locale, utilisée par les méthodes TIFROM et TIFCORR, qui se base sur la notion de régions temps-fréquence. Nous proposons également, dans le cas des mélanges anéchoïques, une méthode permettant l'estimation de délais, qui tout comme la méthode TIFROM [PD05, PD06] permet de dépasser la limitation à des délais inférieurs à un échantillon de la méthode DUET.

Au chapitre 3, nous présentons les approches basées sur l'hypothèse de parcimonie des sources et la connaissance des paramètres du mélange, pour l'estimation des sources du mélange. Nous verrons que ces approches, que se soit dans le cas du masquage binaire de DUET ou dans le cas des techniques de minimisation de norme l_p, sont essentiellement spatiales.

Au chapitre 4, nous présentons une approche spectrale d'estimation des sources à l'aide de Modèles de Mélanges de Gaussiennes (MMG). Cette approche est utilisée dans le cas monophonique, où aucune information spatiale n'est disponible. Nous verrons que cette approche est généralisable au cas multicanal, mais que dans tous les cas une étape d'apprentissage préalable des sources est nécessaire.

Au chapitre 5, nous expliquons comment les modèles MMG sont estimés dans la pratique. Une première approche consite à utiliser un algorithme d'apprentissage EM (*Expectation-Maximisation*) à partir de données d'apprentissage, tandis qu'une seconde approche cherche à estimer les modèles ainsi que les paramètres du mélange, à l'aide d'un algorithme EM, directement à partir de l'observation du mélange. Ces deux approches, en particulier la seconde, comportent cependant des limitations qui les rendent difficilement utilisables en pratique.

Contributions

Cette partie présente les contributions apportées au cours de mon doctorat au problème de séparation de sources sonores.

Au chapitre 6, nous présentons une nouvelle approche d'estimation des paramètres du mélange, basée sur la notion de régions temps-fréquence et sur l'hypothèse de parcimonie des sources. A partir de l'étude d'un modèle de la distribution du mélange dans ces régions temps-fréquence, nous proposons des estimateurs de direction qui ont des propriétés de robustesse, grâce notament à l'utilisation d'une mesure de fiabilité. Cette mesure de fiabilité permet notamment de détecter les régions temps-fréquence où essentiellement une seule source est active, c.-à-d. les régions dans lesquelles l'estimation locale de direction est particulièrement *fiable*. L'estimateur de direction locale que nous proposons a la propriété d'invariance par rotation, ce qui permet d'estimer une direction qui est proche d'un canal de façon identique à une direction qui est placée au centre (c.-à-d. qui a la même intensité sur les différents canaux). Les estimateurs de direction locale et de fiabilité que nous proposons peuvent être utilisés par un algorithme standard

de *clustering* comme le K-Means afin d'estimer les directions du mélange. Cependant ces algorithmes nécessitent que l'on spécifie le nombre de sources du mélange. Nous proposons un nouvel algorithme de *clustering* appelé DEMIX qui permet d'estimer les directions du mélanges et leur nombre. Nous proposons également une méthode d'estimation des délais, qui généralise la méthode GCC-PHAT au cas multi-sources, et qui permet d'estimer des délais supérieurs à un échantillon.

Au chapitre 7, nous présentons des résultats d'experiences qui démontrent la robustesse de l'approche proposée pour estimer les directions de mélanges instantanés et anéchoïques. En particulier nous évaluons la robustesse de la méthode DEMIX avec des méthodes de l'état de l'art dans des situations difficiles, c.-à-d. quand les directions sont proches les une des autres, quand le nombre de sources à séparer est important, ou quand les délais sont grands.

Au chapitre 8, nous présentons une nouvelle famille d'appoches, pour l'estimation des sources à partir de la connaissance des paramètres du mélange. Cette famille d'approches est basée sur l'exploitation de l'information spatiale contenue dans les régions temps-fréquence, afin d'estimer la distribution des sources. Nous proposons une première approche appelée Modèle Gaussien Local (MGL), où la distribution des sources est supposée localement gaussienne. Nous proposons ensuite des approches pour estimer les modèles MMG spectraux introduits au chapitre 4 en aveugle, c.-à-d. sans avoir à recourir à d'autres informations que l'observation du mélange et les paramètres du mélange. Nous présentons également des résultats d'experiences qui montrent une augmentation significative des performances en séparation des approches spatio-spectrales que nous proposons par rapport aux méthodes de l'état de l'art basées sur la parcimonie.

Conclusion et Perspectives

Au chapitre 9, nous concluons la thèse, et au chapitre 10, nous proposons entre autres quelques perspectives pour améliorer l'étape d'estimation des modèles spectraux du chapitre 8, ainsi que des extensions des approches d'estimation des sources et des paramètres du mélanges pour le cas convolutif.

Listes des articles relatifs aux contributions de cette thèse

Cette thèse a donné lieu à quatres articles de conférences et un article de revue :

- S. Arberet, R. Gribonval, F. Bimbot. A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In *ICA*, 2006.
- S. Arberet, R. Gribonval, F. Bimbot. A robust method to count and locate audio sources in a stereophonic linear anechoic mixture. In *ICASSP*, 2007.
- S. Arberet, A. Ozerov, R. Gribonval, F. Bimbot. Blind spectral-GMM estimation for underdetermined instantaneous audio source separation. Submitted to *ICA*, 2009.
- E. Vincent, S. Arberet, R. Gribonval. Underdetermined instantaneous audio source separation via local gaussian modeling. Submitted to *ICA*, 2009.

– S. Arberet, R. Gribonval, F. Bimbot. A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture. submitted to *IEEE Transactions on Signal Processing*.

Demonstration et Evaluation

La méthode DEMIX d'estimation des directions dans le cas de mélanges instantanés, ainsi que les méthodes d'estimation des sources MGL et MGL-MMG qui constituent une contribution de cette thèse ont été soumises à la campagne d'évaluation SICEC 2008. Les résultats de la campagne ainsi que les sources estimées seront en ligne à partir de Decembre 2008.

Première partie

Etat de l'art

Chapitre 1

Les grandes familles d'approches

Dans ce chapitre, nous discutons des grandes familles d'approches cherchant à résoudre le problème de séparation de sources.

Nous considérons dans ce chapitre uniquement le modèle de mélange instantané, qui est la forme la plus simple du problème de séparation de sources, mais qui est suffisant pour présenter les principales familles d'approches de l'état de l'art.

Le problème de séparation de sources aveugle dans le cas linéaire instantané consiste à factoriser une matrice d'observation $\mathbf{X}$ de taille $M \times T$ en deux matrices $\mathbf{A}$ de taille $M \times N$ et $\mathbf{S}$ de taille $N \times T$, avec la relation d'égalité :

$$\mathbf{X} = \mathbf{AS} \tag{1.1}$$

Les lignes de la matrice $\mathbf{S}$ sont les signaux discrétisés des sources, et chaque colonne de la matrice $\mathbf{A}$ contient les gains appliqués à l'une des sources sur chacun des canaux du mélange.

Modèle bruité ? On peut éventuellement considérer un bruit additif dans la modélisation, mais dans ce cas, il est toujours possible de se ramener à un problème de factorisation, soit en considérant le bruit comme M sources, soit en se contentant d'une approximation de cette factorisation $\mathbf{X} \approx \mathbf{AS}$.

Séparation en 1 ou 2 étapes Certaines approches cherchent à estimer $\mathbf{A}$ et $\mathbf{S}$ conjointement, tandis que d'autres séparent le problème de factorisation en deux sous-problèmes : estimer $\mathbf{A}$, puis estimer $\mathbf{S}$ connaissant $\mathbf{A}$. Nous discutons dans ce chapitre des aspects architecturaux des méthodes de séparation aveugle de sources multicanal, ainsi que des hypothèses générales et des méthodes qui permettent d'estimer les inconnues $\mathbf{A}$ et $\mathbf{S}$ du problème.

Cas déterminés Dans le cas déterminé et sur-déterminé, c'est à dire quand le nombre de source N est égal ou inférieur au nombre de capteurs M, il est trivial dans le cas instantané (par une simple inversion matricielle) d'estimer les sources si l'on connait la

matrice de mélange $\mathbf{A}$:

$$\widehat{\mathbf{S}} = \mathbf{A}^{\dagger}\mathbf{X}.$$

De même il est trivial d'estimer la matrice de mélange si l'on connait les sources :

$$\widehat{\mathbf{A}} = \mathbf{X}\mathbf{S}^{\dagger}.$$

Les techniques de l'Analyse en Composantes Indépendantes (ACI), qui sont historiquement les premières à avoir été proposées pour résoudre le problème de séparation de sources, cherchent à estimer les sources directement à partir du mélange, en supposant que celles-ci sont indépendantes. Malheureusement les méthodes de l'ACI supposent que le mélange est déterminé ou sur-déterminé.

Cas sous-déterminé Dans le cas sous-déterminé qui nous intéresse particulièrement, l'estimation des paramètres du mélange n'est pas suffisant pour reconstruire les sources.

Généralement la séparation de source dans le cas sous-déterminé est effectuée en deux étapes disinctes qui exploitent l'hypothèse de parcimonie des sources :

- Estimation les paramètres du mélange $\mathbf{A}$, à partir de l'observation du mélange $\mathbf{X}$;
- Estimation des sources $\mathbf{S}$ à partir de l'observation du mélange $\mathbf{X}$ et des paramètres du mélange $\widehat{\mathbf{A}}$ estimés à l'étape précédente.

L'intérêt de cette architecture en deux étapes est sa modularité qui permet de combiner n'importe quelle méthode de l'étape 1 avec n'importe quelle méthode de l'étape 2.

Il existe cependant des méthodes qui estiment simultanément les paramètres du mélange et les sources :

Une extension de l'ACI pour des mélanges sous-déterminés bruités, permet, à l'aide de l'algorithme EM d'estimer les sources conjointement à la matrice de mélange. Cependant cette approche a des limites. D'une part, il est nécessaire de spécifier une distribution sur les sources qui soit suffisamment simple pour permettre à l'algorithme EM de converger vers une solution pertinente en temps raisonable, mais qui soit suffisamment complexe pour pouvoir modéliser les sources convenablement. D'autre part, avec cette modélisation, l'algorithme EM ne converge plus quand le bruit devient nul.

Les méthodes de Factorisation en Matrices Non-Négatives (FMNN), cherchent également à estimer conjointement les sources et les paramètres du mélange, en leur imposant une contrainte de positivité.

Plan du chapitre Dans ce chapitre, nous présentons dans un premier temps le principe des techniques de l'ACI, ainsi que son extension au cas sous-déterminé bruité. Nous donnons comme exemple une spécification gaussienne des sources. Nous présentons ensuite les méthodes de FMNN et enfin l'Analyse en Composantes Parcimonieuses, qui est à la base des approches que nous proposons dans cette thèse.

1.1 L'Analyse en Composantes Indépendantes (ACI)

L'Analyse en Composantes Indépendantes regroupe un ensemble de méthodes d'analyse statistique de données multivariées, visant à extraire de données observées $\mathbf{X} = [\mathbf{x}(1)\dots\mathbf{x}(T)]$, des composantes linéaires « aussi indépendantes que possible ». Dans le cadre de la séparation de sources, il s'agit d'estimer linéairement les sources $\mathbf{S} = [\mathbf{s}(1)\dots\mathbf{s}(T)]$ à partir du mélange $\mathbf{X} = \mathbf{AS}$, avec $\mathbf{A}$ de taille $M \times M$ inversible mais inconnue, en maximisant un critère d'indépendance des sources estimées. Autrement dit, le problème est [Car02] :

$$\widehat{\mathbf{s}}(t) = \widehat{\mathbf{A}}^{-1}\mathbf{x}(t), \quad \widehat{\mathbf{A}} = \underset{\mathbf{A}\in GL(M)}{\arg\min}\ I(\mathbf{Y} = \mathbf{A}^{-1}\mathbf{X}) \tag{1.2}$$

où $I(\mathbf{Y})$ est une mesure de la dépendance des lignes de la matrice aléatoire $\mathbf{Y}$ de taille $M \times T$ et $GL(M)$ est le groupe général linéaire de degré M. Il existe un certain nombre de critères de dépendance. Citons par exemple l'information mutuelle

$$I(\mathbf{Y}) \triangleq K(P_{\mathbf{Y}}|\prod_n P_{y_n}),$$

où $P_{\mathbf{Y}}$ est la distribution de $\mathbf{Y}$, P_{y_n} est la distribution marginale de la ligne n de $\mathbf{Y}$, et $K(f|g)$ est la divergence de Kullback-Leibler des distributions f et g définie par :

$$K(f|g) \triangleq \int f(x)\log\frac{f(x)}{g(x)}dx.$$

L'information mutuelle est une quantité qui est toujours positive qui s'annule si et seulement si les sources sont indépendantes. Il est donc naturel de minimiser cette quantité pour résoudre le problème de l'ACI.

Il existe des liens importants entre l'information mutuelle et l'estimation de $\mathbf{A}$ (ou de façon équivalente $\mathbf{S}$) au maximum de vraisemblance, quand l'on suppose que les sources sont indépendantes, c.-à-d. que $P_{\mathbf{S}}(\mathbf{S}) = \prod_n P_{s_n}(s_n)$. On peut montrer [Car07] que maximiser la vraisemblance consiste à minimiser la divergence de Kullback-Leibler $K(P_{\mathbf{Y}}|P_{\mathbf{S}})$ entre la distribution $P_{\mathbf{Y}}$ des sources estimées linéairement, et la distribution $P_{\mathbf{S}}$ des sources donnée par le modèle. Or on peut montrer [Car07] que :

$$K(P_{\mathbf{Y}}|P_{\mathbf{S}}) = I(\mathbf{Y}) + \sum_n K(P_{y_n}|P_{s_n}) \tag{1.3}$$

Ce qui signifie que l'estimation au maximum de vraisemblance consiste à minimiser l'information mutuelle, tout en cherchant à rapprocher les distributions des sources estimées de celles du modèle. Si l'on a aucune connaissance sur la distribution des sources, il est naturel de choisir le modèle de source le plus vraisemblable (c.-à-d. $P_{s_n} = P_{y_n}, \forall n$). Dans ce cas, le second terme de (1.3) s'annule, et il y a équivalence entre la minimisation de l'information mutuelle et maximisation de la vraisemblance. Par contre, si la distribution des sources est connue, alors la maximisation de la vraisemblance donnera de meilleures performances en séparation que le minimisation de l'information mutuelle [MC07].

Il existe un lien important entre l'information mutuelle et la corrélation. On définit la corrélation de $\mathbf{Y}$, notée $C(\mathbf{Y})$, comme l'information mutuelle de son approximation gaussienne [Car02] :

$$C(\mathbf{Y}) \stackrel{\triangle}{=} I(\mathbf{Y}^G) = K(\mathcal{N}(0, \bar{\mathbf{R}}_y) | \mathcal{N}(0, \operatorname{diag} \bar{\mathbf{R}}_y)) \tag{1.4}$$

où $\mathbf{Y}^G$ est une matrice dont les colonnes sont des vecteurs aléatoires iid de loi gaussienne $\mathcal{N}(\mu_y, \bar{\mathbf{R}}_y)$ avec les même moments d'ordre 1 et 2 que $\mathbf{Y}$, $\bar{\mathbf{R}}_y = \frac{1}{T} \sum_t \mathbb{E}\left\{\mathbf{y}(t)\mathbf{y}(t)^T\right\}$ (rappelons que $\mathbf{y}(t), 1 \leq t \leq T$ sont des vecteurs aléatoires qui ne sont pas *a priori* identiquement distribués) de taille $M \times M$ étant la moyenne des matrices de covariance de colonnes de $\mathbf{Y}$, et $\operatorname{diag} \bar{\mathbf{R}}_y$ étant la matrice diagonale ayant la même diagonale que $\bar{\mathbf{R}}_y$. En fait la corrélation ne dépend pas du moment d'ordre 1, μ_y, et mesure l'écart de $\bar{\mathbf{R}}_y$ à la diagonalité [Car02]. On voit bien que si les sources sont iid gaussiennes, alors l'information mutuelle est égale à la corrélation, et donc l'ACI ne fait dans ce cas là qu'une simple décorrélation. Notons que contrairement à l'Analyse en Composantes Principales (ACP) qui effectue une décorrélation unique grâce à la contrainte d'orthogonalité des composantes ($\widehat{\mathbf{A}}^T \widehat{\mathbf{A}} = I$), le nombre de solutions de l'ACI dans le cas gaussien est infini. Ainsi, la minimisation de la corrélation ne suffit pas à elle seule à déterminer le mélange $\mathbf{A}$.

Il est aussi possible de relier l'information mutuelle à la corrélation dans le cas général en introduisant la non-gaussiannité $G(\mathbf{Y})$ comme la divergence entre la distribution de $\mathbf{Y}$ et sa meilleure approximation iid gaussienne $\mathbf{Y}^G$: $G(\mathbf{Y}) \stackrel{\triangle}{=} K(\mathbf{Y}|\mathbf{Y}^G)$. Notons que cette définition est aussi valable pour un vecteur. L'information mutuelle peut alors s'écrire :

$$I(\mathbf{Y}) = C(\mathbf{Y}) - \sum_n G(\mathbf{y}_n) + cst$$

Ainsi, minimiser l'information mutuelle consiste à minimiser la corrélation $C(\mathbf{Y})$ entre les sources, tout en maximisant leur "non-iid-gaussianité" $G(\mathbf{y}_n)$.

À l'annexe A, nous montrons qu'il n'est pas possible d'identifier la matrice de mélange et les sources dans le cas où les sources sont gaussiennes iid, mais que par contre si les sources sont gaussiennes à matrice de covariance diagonale dans une base de Fourier, alors l'identification des sources (et du mélange) est possible à condition que les sources aient des spectres différents. Cette dernière modélisation est particulièrement adaptée à la séparation de sources audio.

Algorithmes de l'ACI Dans le cas où l'on suppose une distribution gaussienne des sources diagonale dans une base, l'information mutuelle correspond à un critère de diagonalisation conjointe, qui peut être optimisé par l'algorithme de Pham [Pha01] ou ceux de Cardoso et Souloumiac [CS93, CS96]. Dans le cas où les modèles des sources sont iid-non gaussiens, on peut citer l'algorithme fastICA [Hyv99], l'algorithme HJ [JH91] qui est à l'origine de l'ICA et propose une approche *neuromimétique*, les algorithmes de Comon [COM94] basées sur la maximisation d'un contraste qui approxime l'information mutuelle par une fonction des cumulants des observations, et l'algorithme JADE [Car99, CS93] qui utilise un critère voisin.

Limites de l'ACI La principale limitation de l'ACI est qu'elle repose sur le fait que la matrice de mélange $\mathbf{A}$ soit inversible. L'ACI est donc limitée à la séparation dans les cas déterminés ou sur-déterminés. Il faut aussi noter que la prise en compte d'un bruit additif dans le modèle de mélange casse la structure "inversible" du modèle de l'ACI, brisant ainsi les liens entre la vraisemblance et l'information mutuelle de l'équation (1.3).

1.2 ACI sous-déterminée pour des mélanges bruités

Partant de l'hypothèse de l'ACI, nous discutons dans cette section des conséquences de l'adjonction du bruit gaussien dans le modèle de mélange. En particulier nous étudions la forme que prend la vraisemblance dans le cas bruité, ainsi que les algorithmes qui permettent de la maximiser. Le modèle bruité, que nous allons présenter, est particulièrement interressant, car il ne fait pas apparaitre dans l'expression de la vraisemblance de différences significatives entre le cas déterminé sur-déterminé et sous-déterminé [Car07, BC99a]. Pour ces raisons nous appelons l'approche décrite dans la présente section ACI sous-déterminée, par opposition à l'ACI classique qui a été présentée à la section 1.1, que nous appelons ACI quand il n'y a pas d'ambiguïté, ou bien ACI déterminée.

Le modèle du mélange est le suivant :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{b}(t) \tag{1.5}$$

La matrice de mélange $\mathbf{A}$ est une matrice de rang plein et de taille $M \times N$, avec éventuellement $N > M$, les sources sont des processus iid de densité de probabilité $P_{\mathbf{s}}(\mathbf{s}) = \prod_n P_{s_n}(s_n)$, et le bruit est un processus gaussien iid de moyenne nulle $\mathbf{b} \sim \mathcal{N}(\mathbf{b}; 0, \mathbf{R}_b)$ et indépendant des sources. La distribution des sources et du bruit étant iid, il en est de même du mélange $P(\mathbf{X}) = \prod_t P(\mathbf{x}(t))$, et $\mathbf{x}(t)$ sont différentes observations de la variable aléatoire $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{b}$.

1.2.1 Estimation de la matrice du mélange et du bruit

Nous détaillons à l'annexe A.2.1 la dérivation des estimateurs de $\mathbf{A}$ et de $\mathbf{R}_b$ au maximum de vraisemblance, dont les formules sont données aux équations (A.15) et (A.21) :

$$\widehat{\mathbf{A}} = \widehat{\mathbf{R}}_{xs|x,\xi}\widehat{\mathbf{R}}_{s|x,\xi}^{-1} \tag{A.15}$$

$$\widehat{\mathbf{R}}_b = \widehat{\mathbf{R}}_x - \widehat{\mathbf{R}}_{xs|x,\xi}\widehat{\mathbf{A}}^T \tag{A.21}$$

$\xi = (\mathbf{A}, \mathbf{R}_b, \mathbf{\Lambda})$ désigne l'ensemble des paramètres du modèle bruité, $\mathbf{\Lambda}$ étant l'ensemble des paramètres de la densité de probabilité $P_{\mathbf{s}}$ des sources, et les quantités $\widehat{\mathbf{R}}_{xs|x,\xi}$, $\widehat{\mathbf{R}}_{s|x,\xi}$ et $\widehat{\mathbf{R}}_x$ sont données par les équations (A.17) (A.18), (A.19) et (A.20) :

$$\widehat{\mathbf{R}}_x = \frac{1}{T}\sum_t \mathbf{x}(t)\mathbf{x}(t)^T \tag{A.17}$$

$$\widehat{\mathbf{R}}_{sx|x,\xi} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}|\mathbf{x}(t);\xi\right\}\mathbf{x}(t)^T \tag{A.18}$$

$$\widehat{\mathbf{R}}_{s|x,\xi} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x}(t);\xi\right\} \tag{A.19}$$

$$\widehat{\mathbf{R}}_{xs|x,\xi} = \widehat{\mathbf{R}}_{sx|x,\xi}^T \tag{A.20}$$

Afin d'estimer les paramètres $\mathbf{A}$ et $\mathbf{R}_b$, il est donc nécessaire d'estimer les quantités (dites statistiques suffisantes) $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$.

1.2.2 Estimation des sources

Le meilleur estimateur linéaire des sources au sens du Minimum de l'Erreur Quadratique Moyenne (MEQM en français, LMS en anglais), c.-à-d. minimisant :

$$\mathbb{E}\{(\widehat{\mathbf{s}} - \mathbf{s})^2\},$$

est donné par l'espérance conditionnelle des sources connaissant les observations :

$$\widehat{\mathbf{s}} = \mathbb{E}\{\mathbf{s}|\mathbf{x};\xi\} \tag{1.6}$$

Dans le cas où les sources sont Gaussiennes, la probabilité $P(\mathbf{s}|\mathbf{x},\xi)$ étant Gaussienne, cet estimateur est également celui qui maximise le critère MAP (Maximum A Posteriori) et est donné par :

$$\widehat{\mathbf{s}}^G = \boldsymbol{\rho}(\mathbf{x}) \tag{1.7}$$

où $\boldsymbol{\rho}(\mathbf{x})$ est défini par l'équation (A.25).

1.2.3 Limites de l'ACI sous-déterminé

Il est très difficile d'évaluer les espérances conditionnelles $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$ dans le cas général [MCG97] où l'on n'impose pas un modèle pour les sources. Cependant, il est relativement aisé de calculer ces espérances lorsque les sources sont supposées avoir des distributions gaussiennes, ou bien des distributions mélanges de Gaussiennes (MMG).

1.3 ACI sous-déterminée et algorithme EM pour des sources gaussiennes

On suppose que les sources $\mathbf{s}$ sont des Gaussiennes iid de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$:

$$\mathbf{s} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Les sources étant des Gaussiennes indépendantes, la matrice de covariance $\mathbf{\Sigma}$ est diagonale : $\mathbf{\Sigma} = \text{diag}\left(\left[\sigma_1^2, \ldots, \sigma_N^2\right]\right)$.

Nous dérivons à l'annexe A.2.2 les expressions des espérances conditionnelles $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$ nécessaires à l'estimation de $\mathbf{A}$ et $\mathbf{R}_b$ et $\mathbf{s}$. Nous présentons ces formules ci-dessous :

$$\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\} = \boldsymbol{\rho}(\mathbf{x}) \tag{A.31}$$

$$\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\} = \boldsymbol{\rho}(\mathbf{x})\boldsymbol{\rho}(\mathbf{x})^T + \mathbf{C} \tag{A.32}$$

où $\boldsymbol{\rho}(\mathbf{x})$ et $\mathbf{C}$ sont définis par :

$$\boldsymbol{\rho}(\mathbf{x}) = \boldsymbol{\mu} + \mathbf{W}\left(\mathbf{x} - \mathbf{A}\boldsymbol{\mu}\right) \tag{A.25}$$

$$\mathbf{W} = \mathbf{\Sigma}\mathbf{A}^T\mathbf{B} \tag{A.26}$$

$$\mathbf{C} = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{A}^T\mathbf{B}\,\mathbf{A}\mathbf{\Sigma} \tag{A.27}$$

$$\mathbf{B} = \left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{R}_b\right)^{-1} \tag{A.28}$$

Pour pouvoir estimer $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$, il est nécessaire de connaitre les paramètres $\xi_s = (\boldsymbol{\mu}, \mathbf{\Sigma})$ des sources, dont l'estimation au maximum de vraisemblance est donnée par les équations (A.33) et (A.34) :

$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}|\mathbf{x}(t);\xi\right\} \tag{A.33}$$

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x}(t);\xi\right\} - \text{diag}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T\right) \tag{A.34}$$

Ainsi, l'estimation des paramètres $\xi = (\mathbf{A}, \mathbf{R}_b, \boldsymbol{\mu}, \mathbf{\Sigma})$ du mélange au maximum de vraisemblance, conduit à des estimateurs qui dépendent des espérances conditionnelles $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$, dont l'expression dans le cas gaussien est donnée par les équations (A.31) et (A.32). Or celles-ci dépendent des paramètres ξ.

Pour résoudre ce problème, il est possible d'estimer alternativement les espérances conditionnelles $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$ connaissant les paramètres $\xi = (\mathbf{A}, \mathbf{R}_b, \boldsymbol{\mu}, \mathbf{\Sigma})$ du modèles, puis les paramètres ξ connaissant $\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$. C'est ce que fait l'algorithme EM [DLR], que l'on développe à la section 1.3.1.

1.3.1 Principe de l'algorithme EM

L'algorithme EM (Expectation-Maximization) [DLR], est une méthode itérative de recherche du maximum de vraisemblance, qui à chaque itération ré-estime les paramètres du modèle en garantissant une croissance monotone de la vraissemblance. L'algorithme EM cherche à maximiser une vraisemblance, non pas des données observées, mais des données observées auxquelles s'ajoute des variables aléatoires du modèle, dites variables cachées, et qui si elles étaient observées, permettraient une identification explicite du modèle.

Au lieu de maximiser la log-vraisemblance $\log P(\mathbf{x};\xi)$, la méthode EM cherche à maximiser la fonctionnelle :

$$Q(\xi \,|\, \xi') = \mathbb{E}\left\{ \log P(\mathbf{X}, \mathbf{Z}|\, \xi) |\, \mathbf{X}; \xi' \right\} \tag{1.8}$$

où $\mathbf{X}$ sont les données observées, et $\mathbf{Z}$ les variables cachées. Une propriété importante de cette fonctionnelle est que si $Q(\xi \,|\, \xi') \geq Q(\xi' \,|\, \xi')$, alors $\log P(\mathbf{x};\xi) \geq \log P(\mathbf{x};\xi')$. Par conséquent, la séquence $\left\{\xi^{(0)}, \xi^{(1)}, \ldots\right\}$ des paramètres obtenues par l'itération :

$$\xi^{(l+1)} = \arg\max_{\xi} Q\left(\xi \,\middle|\, \xi^{(l)}\right) \tag{1.9}$$

fait croître la vraisemblance de façon monotone. La dérivée de la fonctionnelle EM est :

$$\frac{\partial Q(\xi \,|\, \xi')}{\partial\, \xi} = \mathbb{E}\left\{ \frac{\partial \log P(\mathbf{X}, \mathbf{Z}; \xi)}{\partial\, \xi} \,\middle|\, \mathbf{X}; \xi' \right\} \tag{1.10}$$

Il est interressant de la comparer avec celle de la log-vraisemblance, donnée par l'équation (A.9). Dans l'équation (A.9), les sources $\mathbf{S}$ ne sont pas observées, et correspondent aux variables cachées $\mathbf{Z}$ de la fonctionnelle EM de l'équation (1.8). La seule différence entre les deux formules est que dans la fonctionnelle EM, l'espérance est calculée en prenant les paramètres ξ' obtenues à l'itération précédante, au lieu des paramètres "libres" ξ de la vraisemblance. Il apparait clairement qu'un point fixe de l'algorithme EM ($\xi' = \xi$ et $\frac{\partial Q(\xi \,|\, \xi')}{\partial\, \xi} = 0$), est un point stationnaire de la vraisemblance. L'intérêt de l'algorithme EM, est que la maximisation de la fonctionnelle EM obtenue par l'annulation de l'équation (1.10), contrairement à celle de la log-vraisemblance, peut s'obtenir de manière explicite.

1.3.2 Formules de ré-estimation dans le cas gaussien

Les formules de ré-estimation des paramètres, découlent des formules (A.15), (A.21), (A.33) et (A.34), en prenant pour valeur de ξ, celles estimées à l'itération précédente. On obtient donc pour la matrice de mélange et la covariance du bruit, les ré-estimations suivantes :

$$\mathbf{A}^{(l+1)} = \widehat{\mathbf{R}}_{xs|x,\xi^{(l)}} \widehat{\mathbf{R}}^{-1}_{s|x,\xi^{(l)}} \tag{1.11}$$

$$\mathbf{R}_b^{(l+1)} = \widehat{\mathbf{R}}_x - \widehat{\mathbf{R}}_{xs|x,\xi^{(l)}} \left(\mathbf{A}^{(l+1)}\right)^T \tag{1.12}$$

et pour les paramètres des sources gaussiennes :

$$\boldsymbol{\mu}^{(l+1)} = \frac{1}{T} \sum_t \mathbb{E}\left\{ \mathbf{s}|\mathbf{x}(t); \xi^{(l)} \right\} \tag{1.13}$$

$$\boldsymbol{\Sigma}^{(l+1)} = \widehat{\mathbf{R}}_{s|x,\xi^{(l)}} - \mathrm{diag}\left(\boldsymbol{\mu}^{(l+1)} \boldsymbol{\mu}^{(l+1),T}\right) \tag{1.14}$$

Les valeurs de $\widehat{\mathbf{R}}_x$, $\widehat{\mathbf{R}}_{s|x,\xi}$ et $\widehat{\mathbf{R}}_{xs|x,\xi}$ sont données par les équations (A.17), (A.19), (A.20).

Limites d'identifiabilité : Le modèle gaussien qui déjà n'était pas identifiable dans le cas de l'ACI déterminé, n'est évidemment pas identifiable dans le cas sous-déterminé (voir annexe A.2.2 pour plus de détailles sur les limites d'identifiabilité du modèle gaussien). Cependant, si l'on connait certains paramètres du modèles, comme les variances et les moyennes des sources, ou bien la matrice de mélange, et que le modèle devient identifiable, alors l'algorithme EM qui vient d'être décrit peut être utilisé en fixant la valeur de ces paramètres.

1.3.3 Comportement de l'algorithme EM quand le bruit tend vers zéro

Nous étudions maintenant le fonctionnement de l'algorithme EM, quand la covariance du bruit tend vers zéro. Nous considérons la limite $\mathbf{R_b} = \sigma_b^2\mathbf{I}, \sigma_b^2 \to 0$. Analysons dans un premier temps ce que devient la probabilité des sources conditionnellement aux observations lorsque le bruit est supposé nul, puis nous verrons ensuite ce que deviennent les formules de ré-estimation de la matrice de mélange.

1.3.3.1 Distribution *a posteriori* des sources

Pour une densité gaussienne des sources et pour un bruit non nul, la densité de probabilité $P(\mathbf{s}|\mathbf{x};\xi)$ est une Gaussienne dont l'équation est donnée en annexe (voir équation (A.24)). Sans perte de généralité, mais afin de simplifier les explications qui suivent, nous considérons que la moyennes des sources est nulle $\boldsymbol{\mu} = \mathbf{0}$. La densité de probabilité *a posteriori* des sources est alors donnée par :

$$P(\mathbf{s}|\mathbf{x};\xi) = \mathcal{N}(\mathbf{s}; \mathbf{W}\mathbf{x}, \mathbf{C})$$

Quand $\mathbf{R_b} = \sigma_b^2\mathbf{I}, \sigma_b^2 \to 0$, les équations du filtre de Wiener [Wie64] $\mathbf{W}$ et de la covariance *a posteriori* des sources $\mathbf{C}$ s'écrivent sous des formes différentes selon la détermination du mélange.

Cas (sur-)déterminés :

– cas déterminé :

$$\mathbf{W} = \mathbf{A}^{-1}$$
$$\mathbf{C} = \mathbf{0}$$

– cas sur-déterminé :

$$\mathbf{W} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$$
$$\mathbf{C} = \mathbf{0}$$

La covarariance des sources $\mathbf{C}$ s'annule, rendant cette distribution singulière : $P(\mathbf{s}|\mathbf{x};\xi) = \delta[\mathbf{x} - \boldsymbol{\rho}(\mathbf{x})]$. Autrement dit, la structure spatiale est suffisante pour imposer à elle seule une solution unique pour les sources.

Cas sous-déterminé :

$$\mathbf{W} = \mathbf{\Sigma A}^T \left(\mathbf{A \Sigma A}^T\right)^{-1} \tag{1.15}$$

$$\mathbf{C} = \left(\mathbf{I}_N - \mathbf{WA}\right) \mathbf{\Sigma} \tag{1.16}$$

Dans le cas sous-déterminé, la structure spatiale n'est pas suffisante pour déterminer à elle seule les sources. La distribution des sources conditionnellement aux observations reste gaussienne, cependant on peut remarquer que contrairement au cas bruité, cette distribution n'est plus définie sur l'ensemble de l'espace de dimention N des sources, mais dans le sous-espace des solutions $\mathbf{x} = \mathbf{As}$ de dimension $N - M$. On peut montrer (voir annexe A.2.3) que l'image de $\mathbf{C}$ est le noyau de $\mathbf{A}$.

$$\mathrm{Im}\mathbf{C} = \mathrm{Ker}\mathbf{A}$$

Exemple d'un mélange sous-déterminé monophonique : Prenons maintenant un exemple très simple où l'on a un mélange monophonique ($M = 1$) $\mathbf{x_0}$, composé de deux sources gaussiennes $s_1 \sim \mathcal{N}(s_1; 0, \sigma_1^2)$ et $s_2 \sim \mathcal{N}(s_1; 0, \sigma_2^2)$. La matrice de mélange est alors $\mathbf{A} = [1, 1]$. En utilisant les formules (1.15) et (1.16), on obtient les équations du filtre de wiener [Wie64]

$$\mathbb{E}\{\mathbf{s}|\mathbf{x_0}; \xi\} = \mathbf{W}\mathbf{x_0} = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \\ \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} \end{bmatrix} \cdot \mathbf{x_0}$$

La variance *a posteriori* de $\mathbf{s}$ dans l'espace de Ker$\mathbf{A}$ est donnée par l'unique valeur propre non nulle de $\mathbf{C}$ qui vaut :

$$\sigma^2_{\mathbf{s};\xi} = \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

Les densités de probabilité marginales des sources connaissant le mélange sont alors : $P(s_1|\mathbf{x}_0; \xi) = \mathcal{N}\left(s_1; \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\mathbf{x}_0, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$ et $P(s_2|\mathbf{x_0}; \xi) = \mathcal{N}\left(s_2; \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mathbf{x}_0, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$.

La figure 1.1 fait apparaitre les distributions a priori de chacune des sources aini que la distribution des sources *a posteriori* qui est une Gaussienne dans l'espace des solutions $\mathbf{x_0} = \mathbf{As}$, qui dans notre exemple est une droite passant par les points $(0, \mathbf{x_0})$ et $(\mathbf{x_0}, 0)$. la solution au maximum a posteriori des sources (MAP) est alors la solution obtenue par l'estimateur MEQM et est donnée par $\widehat{\mathbf{s}} = \mathbb{E}\{\mathbf{s}|\mathbf{x_0}; \xi\}$.

1.3.3.2 Formule de ré-estimation de la matrice de mélange

La formule de réestimation de la matrice de mélange donnée par l'équation (1.11) devient :

$$\widehat{\mathbf{R}}_{xs|x,\xi^{(l)}} \widehat{\mathbf{R}}^{-1}_{s|x,\xi^{(l)}} = \widehat{\mathbf{R}}_x \mathbf{A} \left(\mathbf{A}^T \widehat{\mathbf{R}}_x \mathbf{A}\right)^{-1} \cdot \left(\mathbf{A}^T \mathbf{A}\right)$$

Si bien que si $\mathbf{A}$ est carré (cas déterminé), la formule de réestimation devient :

$$\mathbf{A}^{(l+1)} = \mathbf{A}^{(l)} \tag{1.17}$$

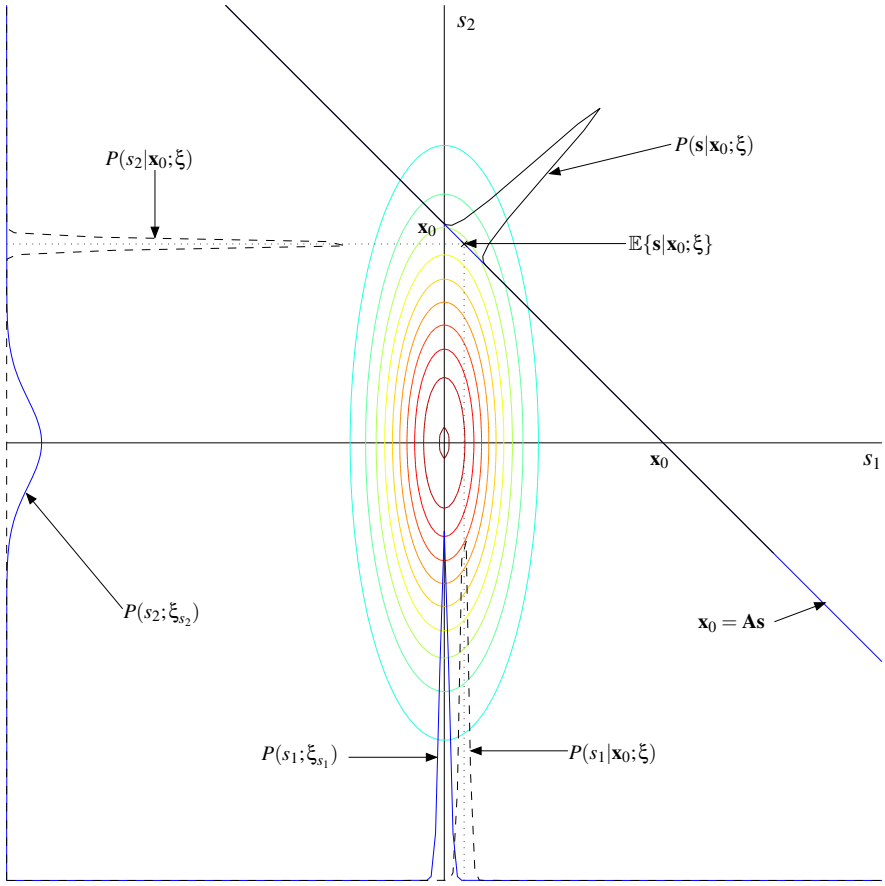


FIG. 1.1 – probabilité a posteriori des sources connaisant le mélange et la distribution des sources

On peut montrer (voir annexe A.2.3) qu'il en est de même dans le cas sous-déterminé, et ceci, quelque soit la distribution sur les sources. Dans le cas sur-déterminé, on peut montrer qu'il y a une infinité de matrice $\mathbf{A}$ qui sont des points stationnaires de la formule de réestimation du mélange.

Discussion : On a vu que lorsque l'on considérait un bruit nul dans le modèle, l'algorithme EM développé pour le modèle bruité ne permettait pas d'estimer la matrice de mélange, et ceci quelque soit la détermination du mélange et la distribution des sources.

Bermond [Ber00, BC99b] a étudié le comportement de l'algorithme EM pour le modèle bruité, dans le contexte d'un bruit faible, en faisant une approximation de Taylor des moments a posteriori impliqués dans les formules de réestimation. La formule de réestimation de la matrice de mélange obtenue prend une forme de gradient classique avec un pas d'apprentissage proportionnel à la covariance du bruit. Si bien que la convergence est d'autant plus lente que le bruit est faible.

1.3.4 Conclusion et limites

Nous avons présenté l'algorithme EM pour estimer conjointement la matrice de mélange et les sources d'un mélange bruité de sources gaussiennes. Cependant comme dans le cas de l'ACI déterminé, l'identification de sources gaussiennes n'est possible que si l'on connait la matrice de mélange.

Estimation de la matrice de mélange : Même si l'on choisit une modélisation non-gaussienne des souces, l'estimation de la matrice de mélange n'est possible que si le bruit est non-nul, et la convergence de l'algorithme EM est très lente pour un bruit faible [BC99b]. Ainsi, il est peut conseillé d'utiliser l'approche EM pour estimer la matrice de mélange.

Initialisation et Convergence de l'algorithme EM : Un autre problème important en général de l'approche EM est sa convergence, qui non seulement peut être très lente, mais en plus dépend énormément de son initialisation. En effet, bien que l'algorithme EM soit assuré de converger dans un minimum de la vraisemblance, rien ne guarantit que ce minimum soit le minimum global. Ainsi le minimum dans lequel l'algorithme va converger dépend fortement de son initialisation. La quantité de minimum locaux, ainsi que la vitesse de convergence dépend de la proportion de variables cachées et de la complexité du modèle [JK97]. Autrement dit, on a intérêt d'une part à avoir un modèle pas trop complexe pour les sources, surtout si le nombre de source est important, et d'autre part à ne pas utiliser l'algorithme EM pour estimer la matrice de mélange et le bruit en plus des paramètres du modèle des sources.

Estimation d'un Modèle de Mélange de Gaussiennes (MMG) : Au chapitre 5, nous développerons les formules de ré-estimation de l'algorithme EM dans le cas d'une distribution mélange de Gaussiennes (MMG) des sources. Comme nous le verrons au chapitre 4, les modèles MMG sont particulièrement bien adaptés pour la modélisation des spectres à court terme des sources audio.

1.4 La Factorisation en Matrices Non-Négatives (FMNN)

Nous avons déjà fait remarquer que le problème de séparation de sources aveugle est un problème de factorisation, qui si l'on ne définit pas plus de contraintes, a une infinité de solutions. l'ACI propose alors dans le cas où $\mathbf{A}$ est une matrice carrée, de supposer que les lignes de $\mathbf{S}$ sont des réalisations de vecteurs aléatoires indépendants entre eux. La factorisation en matrices non-négatives (FMNN), quant à elle, exploite un critère de non-négativité des matrices $\mathbf{X}$, $\mathbf{A}$ et $\mathbf{S}$. Plus formellement, la problématique de la FMNN est de factoriser une matrice non-négative $\mathbf{X} \in \Re\mathrm{e}^{M \times T}$ en deux matrices non-négatives $\mathbf{A} \in \Re\mathrm{e}^{M \times N}$ et $\mathbf{S} \in \Re\mathrm{e}^{N \times T}$ par l'une des fonctions de coût suivante à minimiser [LS00a] :

$$f(\mathbf{A}, \mathbf{S}) = \|\mathbf{X} - \mathbf{AS}\|^2, \text{sous contrainte que } \mathbf{A}, \mathbf{S} \geq 0 \tag{1.18}$$

$$f(\mathbf{A},\mathbf{S}) = \mathrm{D}_{\mathrm{KL}}\left(\mathbf{X}, \mathbf{AS}\right), \text{sous contrainte que } \mathbf{A},\mathbf{S} \geq 0 \tag{1.19}$$

où $\mathbf{A},\mathbf{S} \geq 0$ signifie que l'ensemble des éléments de $\mathbf{A}$ et de $\mathbf{S}$ doivent être supérieurs ou égaux à zéro, et D_{KL} est la divergence de Kullback-Leibler entre deux distributions gaussiennes définie à l'équation (A.3) de l'annexe A.

Ce problème d'optimisation est convexe en $\mathbf{A}$ et en $\mathbf{S}$, mais non en $(\mathbf{A},\mathbf{S})$ conjointement. Il existe par conséquent plusieurs mimima locaux. D'autre part, pour une valeur du critère (1.18), il n'y a pas de solution unique [XLG03, BBL$^+$07]. Il existe plusieurs algorithmes qui résolvent le problème de FMNN, dont le plus célèbre est celui de Lee & Seung [LS99, LS00a].

Généralement, on suppose que $N < M$ de façon à ce que la factorisation $\mathbf{AS}$ représente une approximation "'compressée"', de rang inférieur, à la matrice des oservations $\mathbf{X}$. On peut alors interpréter la FMNN comme une décomposition de $\mathbf{X}$ en une combinaison linéaire de quelques vecteurs de base $\mathbf{a}_i$ ($\mathbf{a}_i$ étant une colonne de la matrice $\mathbf{A}$), dont les poids sont donnés par les coefficients de la matrice $\mathbf{S}$. Dans le cas où l'on choisit une valeur de $N \geq M$, et si aucun autre critère n'est utilisé, l'algorithme converge vers une solution triviale sans intérêt [EK04]. Par exemple, si $N = M$ l'algorithme peut converger vers la solution $\mathbf{A} = \mathbf{I}_M$ et $\mathbf{S} = \mathbf{X}$, qui n'a effectivement pas grand intérêt.

La FMNN suppose que toutes les composantes sont non-négatives, alors que dans le cas des signaux sonores, les composantes peuvent évidemment prendre des valeurs négatives. Pour ces raisons, l'utilisation de la FMNN appliquée à l'audio se fait plutôt dans le domaine de la densité spectrale de puissance (DSP). De cette manière, les composantes sont toutes non-négatives, D'autre part il est connu que le domaine spectral est dans le cas des signaux audio beaucoup plus parcimonieux que dans le domaine temporel, et par conséquent le domaine spectral est plus approprié à une décomposition de $\mathbf{X}$ en un faible nombre de vecteurs de base, comme le suggère la FMNN.

La FMNN a été utilisé en audio principalement dans le cas monocanal, et plus particulièrement pour la transcription d'instruments de musique [SB03, BBR07]. La FMNN permet alors de factoriser la matrice des observations (dans le domaine de la DSP) $\mathbf{X}$, en une matrice $\mathbf{A}$ dont les colonnes sont des formes spectrales élémentaires, par exemples des notes de musiques, et une matrice $\mathbf{S}$ dont les lignes sont les coefficients d'activité correspondant à ces formes spectrales.

Dans certains cas particuliers, on peut considérer que la FMNN effectue une séparation des sources sonores, si l'on considère que les sources sonores se résument à une unique forme spectrale. Cependant, bien que certains instruments de musique comme les idiophones puissent se résumer approximativement à une unique forme spectrale, la plupart des sources sonores produisent un large ensemble de formes spectrales. Ainsi, pour que la FMNN puisse être utilisée pour la séparation de sources sonores, il est nécessaire de regrouper les composantes spectrales de chacune des sources. Cette étape de clustering s'avère particulièrement délicate [Vir07] si elle ne se base pas sur des connaissances a priori sur les sources.

Dans le cas multicanal, une autre piste consiste à exploiter l'information spatiale pour faire le clustering des formes spectrales. A partir de cette idée, différent forma-

lismes étendant la FMNN ont vu le jour. Parry [PE06] en supposant un mélange instantané, propose de factoriser la matrice des observations de taille $MF \times T$, (M étant le nombre de canaux, F le nombre de fréquences de la DSP, T le nombre de trames de la DSP) en non plus 2, mais 3 matrices. La matrice A des formes spectrales (multicanal) étant maintenant décomposée en 2 matrices, l'une contenant les formes spectrales (monocanal) de sources, l'autre contenant les différences d'intensité spatiale. Fitzgerald [FCC05a] exploite la même idée, mais formalise le problème à l'aide de tenseurs. Fitzgerald [FCC05b] propose également une extension de la FMNN pour la séparation monocanal d'intruments harmoniques, en supposant que les composantes spectrale d'une même source sont les translations d'une forme spectrale de base.

Remarque sur la parcimonie : On a vu que la contrainte $N < M$ entraine une réduction du rang de $\mathbf{X}$ (on suppose que $\mathbf{X}$ est de rang plein), et que cette contrainte, associée à la minimisation de l'erreur de reconstruction (1.18), a pour conséquence que les formes spectrales qui émergent dans $\mathbf{A}$ représentent bien $\mathbf{X}$ (au sens de la parcimonie de la représentation). Cette décomposition ressemble à une décomposition en dictionnaire parcimonieux, bien qu'il n'y ait pas de critère explicite de parcimonie dans la décomposition FMNN. D'ailleurs la méthode K-SVD a été comparée à la FMNN pour la transciption musicale dans l'article [BBR07]. Cependant le fait que la FMNN produise un dictionnaire parcimonieux, n'est qu'un effet « secondaire » de la méthode et n'est pas controlé explicitement par un critère [EK04]. Dans certains cas et en particulier dans le cas dégénéré $M \geq N$, un critère de parcimonie doit être ajoutée à la fonction de coût (1.18), pour éviter d'obtenir des solutions triviales [EK04].

1.5 L'Analyse en Composantes Parcimonieuses (SCA)

L'analyse en composantes parcimonieuses, contrairement à l'ACI et à la FMNN, se décompose en deux étapes distinctes. Premièrement estimer la matrice de mélange $\mathbf{A}$, puis estimer les sources $\mathbf{S}$, connaissant la matrice de mélange. Ces deux étapes s'appuient toutes les deux sur l'hypothèse de parcimonie des sources. L'hypothèse de parcimonie suppose que la plupart des coefficients des sources sont nuls ou proches de zéro. La distribution des valeurs des coefficients présente alors un pic étroit en zéro et une queue longue [OPR05]. Une telle distribution est souvent assimilée à la distribution Laplacienne.

Contrairement à la FMNN, qui ne possède pas de solution unique, l'identification de $\mathbf{A}$ et de $\mathbf{S}$ par la SCA est unique (aux indéterminations de permutation et d'échelle près) sous certaines conditions sur la structure de $\mathbf{A}$ et sur la parcimonie de $\mathbf{S}$ [GTC05]. Cette propriétée est vraie y compris dans le cas sous-déterminé, la parcimonie compensant la limite sur le nombre de capteurs. Les sources étant supposées indépendantes, les conséquences d'une distribution parcimonieuse des sources sont que la probabilité d'avoir plusieurs sources actives (c.a.d. qui aient une valeur non nulle) en même temps est relativement faible. Ainsi la plupart des coefficients du mélange contiennnent essentiellement l'énergie d'une unique source. Certaines méthodes font l'hypothèse que ceci

est vrai pour l'ensemble des coefficients du mélange, autrement dit que les sources ont des supports disjoints. Cette dernière hypothèse a pour conséquence comme on la vu dans l'introduction de faire ressortir la structure de la matrice de mélange qui peut alors être estimée avec un algorithme de clustering. Comme on le verra dans le chapitre suivant, d'autres méthodes font l'hypothèse (plus faible), que les sources ont des supports disjoints seulement dans un sous-ensemble (ayant possiblement une petite taille) de l'ensemble des coefficients du mélange.

L'étape d'estimation des sources consiste à maximiser la parsimonie des sources sous la contrainte d'égalité $\mathbf{x} = \mathbf{As}$.

En pratique on définit la mesure de parcimonie d'un vecteur, par le nombre de composantes non nulles de ce vecteur, c'est à dire sa norme l_0 (qui en réalité n'est pas une norme).

$$\|\mathbf{s}\|_0 \stackrel{\triangle}{=} \#\{k, \mathbf{s}_k \neq 0\} = \sum_k |\mathbf{s}_k|^0 \tag{1.20}$$

Bien que ce critère garantisse l'unicité et l'exactitude de la solution, si les sources ont une certaine parcimonie et sous certaines conditions sur $\mathbf{A}$ [GTC05], ce problème d'obtimisation est combinatoire, et en pratique impraticable pour des matrices de grandes dimensions. Dans le cas des signaux audio, les matrices de mélange sont en général de petite dimension, mais par contre l'hypothèse de parcimonie au sens de la norme l_0 est peu vérifiée en pratique. Pour ces raisons, on utilise en général une forme relaxée de la mesure de parcimonie qui n'est autre que la norme l_p :

$$\|\mathbf{s}\|_p \stackrel{\triangle}{=} \sum_k |\mathbf{s}_k|^p \tag{1.21}$$

Le critère à obtimiser est donc le suivant :

$$f(\mathbf{A}, \mathbf{s}) = \arg\min_{\mathbf{s}|\mathbf{x}=\mathbf{As}} \|\mathbf{s}\|_p \tag{1.22}$$

Les valeurs de p comprises entre 0 et 1 inclus produisent des solutions parcimonieuses assez semblables, mais les algorithmes mis en jeu ne sont pas les mêmes. Pour des valeurs de p telles que $1 < p \leq 2$, le problème est strictement convexe, et contient par conséquent un minimum global, mais les solutions obtenues ne sont plus vraiment parcimonieuses. Pour des valeurs de p telles que $p < 1$, le problème est non convexe et contient des minimums locaux. Cependant pour une valeur de $p = 1$ le problème reste convexe et les solutions obtenues restent parcimonieuses, ce qui explique pourquoi ce cas particulier a été beaucoup étudié dans la littérature.

Aussi il existe une interprétation bayésienne au choix de p. L'optimisation du critère (1.22), correspond au calcul des coefficients $\mathbf{s}$ les plus vraisemblables sachant que $\mathbf{x} = \mathbf{As}$ et sous l'hypothèse que les sources soient des Gaussiennes généralisées iid de paramètre p (voir section 3.3 pour plus de détails sur le modèle gaussien généralisé). Si $p = 2$, la Gaussienne généralisée se ramène à une distribution gaussienne, tandis que pour $p = 1$, la Gaussienne généralisée se ramène à une distribution laplacienne.

Pour tenir compte d'un enventuel bruit additif, ou de l'inexactitude stricto sensus de l'hypothèse de parcimonie, le problème se pose sous la forme lagrangienne suivante :

$$f(\mathbf{A}, \mathbf{s}) = \|\mathbf{x} - \mathbf{A}\mathbf{s}\|^2 + \lambda \|\mathbf{s}\|_p \tag{1.23}$$

où le paramètre λ détermine le compromis entre la qualité de l'approximation de la reconstruction représentée par le premier terme du critère, et l'hypothèse de parcimonie des sources représentée par le second terme.

Comme on l'a vu dans l'introduction, les signaux audio sont très peu parcimonieux dans le domaine temporel (leur domaine d'origine), par contre il existe des transformations linéaires qui permettent d'obtenir des représentations dans lesquelles les coefficients des sources sont parcimonieux. C'est le cas par exemple de la Transformée de Fourier à Court Terme (TFCT). La transformation étant linéaire, la matrice de mélange garde la même forme dans le domaine transformé que dans le domaine temporel : $\mathbf{X}(t, f) = \mathbf{A}\mathbf{S}(t, f)$. La SCA se divise alors en quatre étapes :

- Transformation du mélange $\mathbf{x}(t)$ dans une représentation parcimonieuse $\mathbf{X}(t, f)$;
- Estimation de la matrice de mélange $\widehat{\mathbf{A}}$ à partir de $\mathbf{X}(t, f)$;
- Estimation des sources $\widehat{\mathbf{S}}(t, f)$ à partir de $\widehat{\mathbf{A}}$ et $\mathbf{X}(t, f)$ par un critère de parcimonie comme (1.22) ou (1.23) ;
- Reconstruction des sources dans le domaine temporel $\widehat{\mathbf{s}}(t)$ en appliquant la transformation inverse.

1.6 Conclusion

ACI pour les mélanges (sur-)déterminés : On a vu que pour résoudre le problème de séparation de sources dans le cas instantané (sur-)déterminé, il suffisait d'estimer la matrice de mélange, car les sources peuvent être reconstruites à partir des observations et de la matrice de mélange. Il existe un grand nombre de méthodes qui proposent de résoudre le problème de séparation de sources dans ce contexte, en particulier l'ACI que nous avons présenté à la section 1.1, qui prend comme hypothèse l'indépendance et la non-gaussianité des sources.

ACI pour les mélanges sous-déterminés : Dans le cas d'un mélange sous-déterminé, la connaissance de la matrice de mélange en complément des observations, n'est pas suffisante pour reconstruire les sources qui sont alors contenues dans un espace affine de dimension $N - M$. Des hypothèses plus fortes sur les sources, comme supposer que les sources ont une distribution gaussienne, permettent à l'aide de l'algorithme EM, d'estimer les sources à condition de connaître la matrice de mélange. En théorie, à condition que les sources soient non-gaussiennes, il est possible à l'aide de l'algorithme EM d'estimer à la fois les sources conjointement à la matrice de mélange. Cependant, lorsque le bruit est faible l'estimation de la matrice de mélange devient critique, et en particulier si le bruit est nul, alors la suite des ré-estimées de la matrice de mélange fournie par l'algorithme EM est stationnaire.

Factorisation en Matrices Non-Négatives (FMNN) : La FMNN est une autre approche qui consiste à factoriser conjointement une matrice observée en deux matrices comme dans l'équation (1.1), avec comme contrainte la non-négativité de chacune des matrices de l'équation. Seulement d'une part la solution à ce problème d'optimisation n'est pas unique, d'autre part les signaux audio peuvent prendre des valeurs négatives. Il est tout de même possible de mettre à profit cette technique de factorisation en travaillant avec la densité spectrale des signaux afin d'avoir des coefficients à valeur non-négative. Cependant d'une part l'information de phase des signaux est perdue, d'autre part le problème contient maintenant une dimension supplémentaire, la dimension des fréquences. Le formalisme doit alors être étendu à une factorisation en trois matrices [PE06], ou à un problème de factorisation en tenseurs non-négatifs [FCC05a]. Une étape de clustering est également nécessaire afin de regrouper les composantes spectrales (obtenues par la factorisation) qui appartiennent à la même source.

L'Analyse en Composantes Parcimonieuses : Enfin l'Analyse en Composantes Parcimonieuses prend comme hypothèse que les sources sont parcimonieuses dans une représentation et propose de diviser le problème de séparation de sources en deux étapes bien disctinctes : l'estimation de la matrice de mélange connaissant les observations, puis l'estimation des sources connaissant les observations et la matrice de mélange. L'avantage de cette architecture modulaire séparant clairement une première étape d'estimation de la matrice de mélange et une seconde d'estimation des sources, est qu'il est alors possible de connecter n'importe quelle méthode de la première étape avec n'importe quelle méthode de la seconde. Nous allons voir aux chapitres suivants quelles sont les méthodes de l'analyse en composantes parcimonieuses, pour l'estimation de la matrice de mélange, et l'estimation des sources. Reste alors à savoir dans quelle mesure l'hypothèse de parcimonie est vérifiée pour les signaux audio, et quelle est la robustesse des algorithmes de la SCA quand l'hypothèse de parcimonie n'est pas parfaitement respectée. Nous avons vu que les méthodes de la SCA exploitent généralement une hypothèse plus forte que l'hypothèse de parcimonie des sources : l'hypothèse du support disjoint des sources qui se justifie par la conjonction de l'hypothèse de parcimonie des sources et de l'indépendance des sources. Cependant il est évident que plus le nombre de sources augmente, moins l'hypothèse du support disjoint des sources n'a de chance d'être validée.

La SCA permet de traiter le cas sous-déterminé, à l'exception du cas mono-canal car celui-ci ne présente plus de diversité spatiale. Cependant nous verrons au chapitre 4 qu'il est possible de séparer les sources dans le cas monocanal à l'aide de modèles MMG des spectres des sources, à condition d'avoir effectué une étape d'apprentissage préalable des modèles de chacune des sources par un algorithme EM similaire à celui que nous avons présenté à la section 1.2.

Positionnement de la thèse : L'hypothèse d'indépendance des sources, « seule » (on suppose néanmoins que les sources ne sont pas gaussiennes), permet de résoudre le problème de séparation de sources dans les cas déterminés.

Par contre dans les cas sous-déterminés qui nous intéressent particulièrement, il est nécessaire de faire une hypothèse plus forte sur les sources. Nous avons vu dans ce chapitre deux familles d'approhe qui supposent une certaine distribution des sources :

1. Une première approche où l'on suppose que les sources ont une certaine distribution, par exemple gaussienne ou MMG, dont les paramètres sont différents pour chacune des sources. Le problème de séparation de sources revient alors essentiellement à estimer les paramètres du mélange et les paramètres des sources. L'approche EM propose d'estimer tous ces paramètres en même temps, mais cela pose un certain nombre de problèmes de convergence qui rendent difficile son utilisation pratique.
2. L'approche de l'Analyse en Composantes Parcimonieuses qui suppose que les sources ont une distribution parcimonieuse, par exemple laplacienne, avec des paramètres identiques pour chacune des sources. De ce fait l'estimation des sources est effectuée, par la minimisation d'un critère de parcimonie sans qu'il soit nécessaire d'estimer les paramètres de la distribution de chacune des sources. Par contre il est néccessaire d'avoir estimé la matrice de mélange auparavant.

Ainsi, quelque soit la distribution que l'on fait sur les sources, il semble qu'une architecture en deux étapes, où l'on estime d'abord la matrice de mélange, puis les sources, soit le meilleur choix dans le cas sous-déterminé.

Nous allons voir dans les prochains chapitres les principales approches permettant l'estimation de la matrice de mélange, puis les approches d'estimation des sources par la parcimonie. Nous verrons ensuite une approche d'estimation des sources par des modèles MMG spectraux, dans le cas on l'on connait les paramètres du mélanges et de la distribution des sources. Cette approche, contrairement aux approches de l'analyse en composantes parcimonieuses basées uniquement sur la diversité spatiale, permet la séparation de sources dans le cas particulier où l'on ne dispose que d'un seul capteur.

Chapitre 2

Exploitation de la parcimonie pour estimer les paramètres du mélange

Dans ce chapitre, nous nous intéressons à l'état de l'art pour la première étape de l'analyse en composantes parcimonieuses qui a été introduite à la section 1.5 et qui consiste à estimer les paramètres du mélange à partir des observations du mélange. Par la suite, nous employons parfois le terme de *directions du mélange* pour parler des *paramètres du mélange*.

Contrairement aux méthodes de l'ACI que nous avons présentées à la section 1.1, les méthodes d'estimation qui réposent sur l'hypothèse de parcimonie permettent d'estimer les paramètres du mélange dans le cas éventuellement sous-déterminé.

L'hypothèse de parcimonie n'étant pas valide dans le domaine temporel, nous devons effectuer un changement de représentation du mélange. Dans le domaine du traitement du signal audio on emploie fréquement la Transformée de Fourier à Court Terme (TFCT). La TFCT du signal $x_m(\tau), 0 \leq \tau < T$, définie par l'équation (2.1) est une transformée linéaire pour laquelle il existe des algorithmes très rapides.

$$X_m(t,f) = \sum_{\tau=0}^{T-1} x_m(\tau) h_{\mathrm{L}}(\tau - t) \exp\left(-i2\pi f \tau\right) \tag{2.1}$$

La fenêtre $h_{\mathrm{L}}(.)$ est une fonction de support $[-\mathrm{L}/2, \mathrm{L}/2]$ de norme unité $\|h_{\mathrm{L}}\| = 1$, et f est la fréquence réduite. Les coefficients de la TFCT du mélange $\mathbf{X}(t,f) = [X_1(t,f), \cdots, X_M(t,f)]^T$ sont calculés sur une grille discrète telle que : $t = k\,\mathrm{L}/2, k \in \mathbb{Z}$, et $f = l/\mathrm{L}, 0 \leq l \leq \mathrm{L}/2$, permettant une reconstruction parfaite du signal $\mathbf{x}(t) = [x_1(t), \cdots, x_M(t)]^T$. Pour plus de détails sur ces transformées, on peut se référer à [Mal00].

La TFCT offre une représentation parcimonieuse pour les signaux audio, et est pourvue d'une structure en grille dans le plan temps-fréquence qui peut être mise à profit pour analyser localement le signal comme on le verra par la suite.

Il est possible d'estimer la matrice de mélange en cherchant à minimiser le critère global de parcimonie (1.22), qui peut s'interpréter comme une estimation au maximum de vraisemblance de la matrice de mélange et des sources, en supposant que les sources

ont une distribution gaussienne généralisée. L'estimation globale est difficile comme on l'a vu à la section 1.5. Une approche similaire consiste à alterner estimation des sources et estimation de la matrice de mélange [LS00b].

Une autre famille d'approches [YR04, BZ01, AD03] plus intuitive que nous allons étudier dans ce chapitre consiste à utiliser un algorithme de *clustering* sur les points du diagramme de dispersion. On peut alors diviser la procédure d'estimation des paramètres du mélange en 2 étapes :

1. L'étape d'*extraction des caractéristiques*, qui consiste à extraire des paramètres à partir desquels les directions du mélange sont estimées à la seconde étape ;
2. L'étape de *classification*, dont le rôle est d'estimer les directions du mélange par un algorithme de *clustering* à partir des paramètres extraits pendant la première étape.

Nous allons d'abord nous intéresser à l'estimation des paramètres d'un mélange *linéaire instantané*, puis nous étudierons ensuite les approches de l'état de l'art qui permettent de traiter le cas du mélange *anéchoïque*.

2.1 Mélange linéaire instantané

Dans cette section nous présentons le modèle de mélange du cas instantané, ainsi que les deux grandes familles d'approches de l'état de l'art qui permettent d'estimer les paramètres du mélange. La première dite approche globale, estime les directions à partir du diagramme de dispersion (global) du mélange en supposant que les sources ont un support quasi-disjoint, autrement dit, que pour chaque point temps-fréquence, il y ait au plus une seule source active. La deuxième se base sur une hypothèse de parcimonie plus souple, puisqu'elle suppose qu'il existe seulement quelques régions temps-fréquence où le support des sources est quasi-disjoint.

2.1.1 Le modèle de mélange instantané

L'enregistrement d'un mélange instantané de N sources et M capteurs s'écrit sous forme d'une équation linéaire dans le domaine temporel :

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t) + b_m(t), \qquad 1 \leq m \leq M \tag{2.2}$$

ou de façon équivalente :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{b}(t) \tag{2.3}$$

Le paramètre $a_{mn} \in \mathbb{R}$ représente le gain (ou intensité) qui est appliqué à la source n sur le canal m. Le signal $\mathbf{b}(t)$ est un bruit additionel qui est souvent négligé.

On souhaite estimer le nombre de sources N, et les paramètres d'intensité a_{mn}, à partir de l'observation du mélange $\mathbf{x}(t)$. La plupart des méthodes de séparation de sources aveugles considèrent néanmoins que le nombre de sources N est une donnée connue

[PO05]. Nous nous plaçons dans le cas où le nombre de sources peut être supérieur au nombre de capteurs.

Etant donné que nous ne connaissons ni $\mathbf{A}$ ni $\mathbf{s}(t)$, l'identification de $\mathbf{A}$ n'est possible qu'à une permutation et à un facteur d'échelle près [PO05]. Sans perte de généralité, nous supposons que $\sum_{m=1}^{M} a_{mn}^2 = 1$ et que $a_{1n} \geq 0$ pour $1 \leq n \leq N$, afin de fixer l'indétermination du facteur d'échelle.

Grâce à la linéarité de la TFCT $X_m(t,f)$ de chacun des canaux $x_m(t)$ du mélange, l'équation (2.3) peut s'écrire dans le domaine transformé selon l'équation (2.4), où $\mathbf{X}(t,f)$ et $\mathbf{S}(t,f)$ désignent respectivement les vecteurs colonnes $[X_1(t,f),\ldots,X_M(t,f)]^T$ et $[S_1(t,f),\ldots,S_N(t,f)]^T$.

$$\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f) + \mathbf{B}(t,f) \tag{2.4}$$

La matrice $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_N]$ est de taille $M \times N$. Les vecteurs colonnes $\mathbf{a}_n = [a_{1n}, a_{2n},\ldots,a_{Mn}]^T$ de $\mathbf{A}$ sont les directions des sources du mélange.

Dans le cas stéréophonique ($M = 2$), chaque colonne de $\mathbf{A}$ est un vecteur à deux dimensions qui peut s'écrire :

$$\mathbf{a}_n = \begin{bmatrix} \cos\theta_n \\ \sin\theta_n \end{bmatrix} \in \mathbb{R}^2. \tag{2.5}$$

Le paramètre $a_{2n}/a_{1n} = \tan\theta_n$ est appelé *différence d'intensité* (DI) de la direction n. Le paramètre $\theta_n \in]-\pi/2, \pi/2]$ est le *paramètre d'intensité* de la direction n. Celui-ci peut être négatif si $\theta_n < 0$. Dans le cas stéréophonique linéaire instantané, l'unique paramètre caractérisant une direction est le paramètre d'intensité θ_n. Etant donné la relation biunivoque entre la DI et θ_n, nous employons parfois le terme de DI pour parler de θ_n.

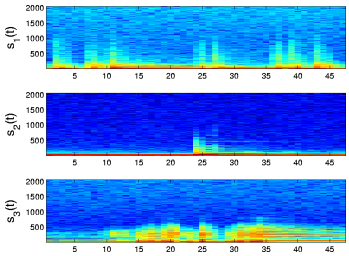
2.1.2 1re étape : Extraction des caractéristiques par l'approche globale

Avant de discuter des « caractéristiques » à partir desquelles on compte estimer les directions du mélange, nous donnons un aperçu d'un *diagramme de dispersion* obtenu à partir d'un mélange stéréophonique à trois sources (voir figure 2.1). Le diagramme de dispersion est la représentation graphique du nuage des points représentant les couples de valeurs $(x_1(t), x2(t))$, pour l'ensemble des échantillons t d'un mélange sur $M = 2$ canaux. Nous étendons cette définition à tous types de représentations. Sur le diagramme de dispersion de la figure 2.1, on a tracé les points
$\mathbf{X}^{\Re\mathrm{e}}(t,f) \triangleq [\Re\mathrm{e}X_1(t,f), \Re\mathrm{e}X_2(t,f)]^T$ et $\mathbf{X}^{\Im\mathrm{m}}(t,f) \triangleq [\Im\mathrm{m}X_1(t,f), \Im\mathrm{m}X_2(t,f)]^T$.

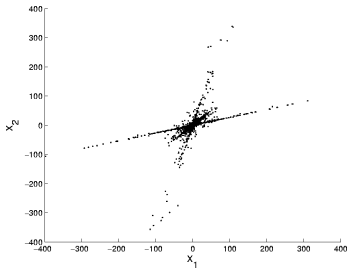
On peut observer que les points sont à peu près alignés le long de trois directions qui correspondent aux directions $\mathbf{a}_n$ de $\mathbf{A}$.

L'hypothèse exploitée par la méthode DUET [YR04] est que le support des sources est quasi-disjoint dans le domaine de la TFCT.

Hypothèse du support disjoint des sources : Si les sources avaient des supports totalement disjoints, alors tous les points du diagramme de dispersions seraient alignés



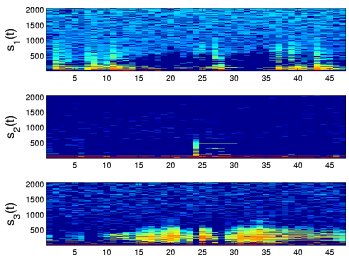
(a) sources musicales dans le domaine temps-fréquence



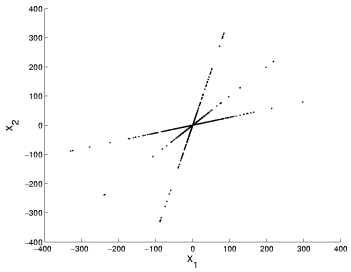
(b) diagramme de dispersion

FIG. 2.1 – Diagramme de dispersion d'un mélange musical stéréophonique composé de trois sources.

le long de l'une des directions. Le diagramme de dispersion ressemblerait alors à celui de la figure 2.2 où pour chaque point temps-fréquence (t, f), on a, avant de les mélanger, volontairement et artificiellement mis à zéro les coefficients de toutes les sources, excepté la source la plus dominante $n(t, f) = \arg\max_{n'} |S_{n'}(t, f)|$, afin que les sources aient un support disjoint dans le domaine



(a) sources musicales dans le domaine temps-fréquence



(b) diagramme de dispersion

FIG. 2.2 – Diagramme de dispersion d'un mélange musical stéréophonique composé de trois sources, qui ont été artificiellement modifiées afin de rendre l'hypothèse de support disjoint valide.

En effet si l'on suppose que pour chaque point temps-fréquence (t, f), une seule source est active, alors il y a un index $1 \leq n(t, f) \leq N$ tel que $\forall n \neq n(t, f)\ |S_n(t, f)| = 0$, et l'équation du mélange se simplifie :

$$\mathbf{X}(t, f) = \mathbf{A}\mathbf{S}(t, f) = \mathbf{a}_{n(t,f)} \cdot S_{n(t,f)}(t, f).$$

Par conséquent :

$$\begin{cases} \Re\mathrm{e}\mathbf{X}(t, f) &= \mathbf{a}_{n(t,f)} \cdot \Re\mathrm{e}S_{n(t,f)}(t, f) \\ \Im\mathrm{m}\mathbf{X}(t, f) &= \mathbf{a}_{n(t,f)} \cdot \Im\mathrm{m}S_{n(t,f)}(t, f) \end{cases} \tag{2.6}$$

et

$$\frac{\Re\mathrm{e}X_2(t,f)}{\Re\mathrm{e}X_1(t,f)} = \frac{\Im\mathrm{m}X_2(t,f)}{\Im\mathrm{m}X_1(t,f)} = \frac{X_2(t,f)}{X_1(t,f)} = \tan\theta_{n(t,f)} \in \mathbb{R} \tag{2.7}$$

Ainsi, si le support des sources était disjoint, alors chaque point non nul fournirait la valeur de l'une des directions, et il suffirait pour trouver le nombre de sources de compter le nombre de valeurs différentes du rapport

$$R_{21}^{\mathbb{R}}(t,f) \triangleq \frac{\mathbb{R}X_2(t,f)}{\mathbb{R}X_1(t,f)}, \tag{2.8}$$

où $\mathbb{R}$ dans l'équation (2.8) représente indifféremment la partie réelle $\Re\mathrm{e}$ ou bien la partie imaginaire $\Im\mathrm{m}$.

Comme on peut le voir sur la figure 2.1, les points ne sont pas rigoureusement alignés sur les directions de $\mathbf{A}$, car malheureusement l'hypothèse du support disjoint des sources n'est pas valide. Il est par conséquent nécessaire de se baser sur des hypothèses de parcimonie moins fortes.

Hypothèse du support quasi-disjoint des sources : L'hypothèse du support quasi-disjoint des sources [YR04] suppose que si une source $S_n(t,f)$ a une énergie importante au point temps-fréquence (t,f), alors la contribution des autres sources a de grandes chances d'être faible. Autrement dit, l'hypothèse suggère qu'il y ait un index $1 \leq n(t,f) \leq N$ tel que $|S_{n(t,f)}(t,f)| \gg |S_n(t,f)|$, $n \neq n(t,f)$, et que par conséquent on ait les approximations suivantes :

$$\mathbf{X}(t,f) \approx \mathbf{a}_{n(t,f)}(f) \cdot S_{n(t,f)}(t,f)$$

et :

$$\frac{\Re\mathrm{e}X_2(t,f)}{\Re\mathrm{e}X_1(t,f)} \approx \frac{\Im\mathrm{m}X_2(t,f)}{\Im\mathrm{m}X_1(t,f)} \approx \tan\theta_{n(t,f)} \in \mathbb{R} \tag{2.9}$$

L'hypothèse du support quasi-disjoint des sources suggère ainsi que les points du diagramme de dispersion soient groupés autour des directions $\mathbf{a}_{n(t,f)}(f)$, ce qui est vérifié par l'observation de la figure 2.1. Afin d'estimer les directions, une approche commune [BZ01] consiste alors à exploiter la structure géométrique des points du diagramme de dispersion par l'utilisation d'un algorithme de clustering.

2.1.3 2$^{\mathrm{e}}$ étape : Estimation des paramètres du mélange par *clustering*

Bien que l'on puisse directement utiliser un algorithme de clustering comme le K-means sur les données du diagramme de dispersion, l'approche [BZ01] couramment utilisée pour estimer les directions consiste à calculer une *fonction potentielle* $\Phi_\lambda(\theta)$ à partir des points du diagramme de dispersion, de façon a obtenir des pics à l'endroit des directions du mélange. L'étape suivante consiste alors à trouver les pics de cette fonction qui correspondent aux directions du mélange.

2.1.3.1 La fonction potentielle

La construction de la *fonction potentielle* est similaire à la méthode de Parzen, qui est une généralisation de la méthode d'estimation par histogramme. La *fonction potentielle* est la somme pondérée de fonctions locales de base h_λ centrées sur les valeurs $\theta(t,f) \triangleq \tan^{-1} R_{21}^{\mathbb{R}}(t,f)$:

$$\Phi_\lambda(\theta) \triangleq \sum_{t,f} \omega(t,f) \cdot h_\lambda\left(\theta - \theta(t,f)\right) \tag{2.10}$$

où $\omega(t,f)$ est un coefficient qui attribue plus de poids aux points qui sont considérés comme étant plus fiables [BZ01], et h_λ est la fonction de base dont le support dépend du facteur de dilatation λ : $h_\lambda(\theta) \triangleq \frac{1}{\sqrt{\lambda}} h\left(\frac{\theta}{\lambda}\right)$, où h est une fonction à support local fixe. La valeur de $\omega(t,f)$ est en général fonction de l'amplitude $\rho^{\mathbb{R}}(t,f) = \sqrt{(\mathbb{R}X_1(t,f))^2 + (\mathbb{R}X_2(t,f))^2}$ du point temps-fréquence : $\omega(t,f) = f(\rho^{\mathbb{R}}(t,f))$. En général, $f(\rho) = \rho$ ou $f(\rho) = \rho^2$. L'idée étant que : à perturbation égale, l'erreur angulaire est plus faible pour les points du diagramme de dispersion qui sont loins de l'origine [BZ01]. DUET utilise la valeur de pondération $\omega(t,f) = |X_1(t,f) \cdot X_2(t,f)|^p$, où la valeur $p = 1$ est suggérée [YR04]. Remarquons que dans le cas où une source $S_n(t,f)$ domine les autres au point temps-fréquence (t,f),

$$|X_1(t,f) \cdot X_2(t,f)|^p \approx \left| \frac{\sin(2\theta_n)}{2} \cdot |S_n(t,f)|^2 \right|^p .$$

Le poids $\omega(t,f)$ de DUET dépend par conséquent de la direction θ_n de la source, et devient nul dans les cas extrêmes où la source n'apparait que sur un canal ($\theta_n = 0$ ou $\theta_n = \pi/2$).

Outre le choix du poid $\omega(t,f)$, il reste à choisir la fonction de base h_λ, ainsi que son facteur de dilatation λ. Bofill utilise une fonction triangulaire [BZ01] pour la fonction h, mais d'autres fonctions sont possibles. Le choix de λ est très délicat car il influe sur le nombre de maximums de la fonction potentielle et dépend des données [BZ01].

Aussi, pour des raisons pratiques, la fonction potentielle $\Phi_\lambda(\theta)$ doit être discrétisée. Celle-ci est généralement opérée sur une grille à espacement régulier $\theta(j) = j\pi/J$, avec $-\frac{J}{2} < j \leq \frac{J}{2}$.

L'utilisation de la fonction potentielle nécessite par conséquent de choisir les paramètres :

- du poids $\omega(t,f)$;
- de la fonction de base h ;
- du facteur de dilatation λ ;
- la grille de discrétisation de $\theta(j)$.

qui dépendent pour la plupart fortement des données et pour lesquels il ne semble pas exister à ce jour d'étude comparative [Gri07].

Exemple de fonction potentielle : A titre d'exemple, nous donnons les paramètres de la fonction potentielle qui est utilisée dans la méthode de Bofill et Zibulevsky [BZ01] :

- Le poids $\omega(t,f)$ est donné par l'amplitude du point temps-fréquence $\omega(t,f) = \rho(t,f) = \sqrt{|X_1(t,f)|^2 + |X_2(t,f)|^2}$; un seuillage sur l'amplitude est utilisé pour réduire le nombre de points à traiter. Ce seuillage consiste à supprimer les points tels que $\rho(t,f) < \eta$, avec η reglé à $1/3$ de la valeur maximale du signal.
- La fonction de base h est une fonction triangulaire définie par :

$$h(\theta) = \begin{cases} 1 - \frac{\theta}{\pi/4} & ,|\theta| < \pi/4 \\ 0 & \text{ailleurs} \end{cases}$$

- Le facteur de dilatation est un paramètre fixé « à la main » en fonction du signal. Sur les exemples de mélanges traités dans l'article [BZ01], l'ordre de grandeur des valeurs prises par ce paramètre varie de 10^{-3} à 10^2 (lorsque θ est en radians).
- La grille de discrétisation est elle aussi choisie « à la main » en fonction du signal. Sur les exemples de mélanges traités dans l'article [BZ01], les valeurs prises par ce paramètre varient entre $J = 30$ et $J = 540$.

2.1.3.2 Limites de l'approche globale

Bien que le diagramme de dispersion global (calculé sur l'ensemble des points $\mathbf{X}^{\mathbb{R}}(t,f)$) de la figure 2.1 fasse apparaître les directions du mélange, leur estimation par un algorithme de clustering tel que le K-means, ou bien par la détection des pics de la fonction potentielle peut s'avérer délicat pour plusieurs raisons :

- Les sources du mélange peuvent avoir des intensités très diverses, et par conséquent les sources de faible intensité sont peu visibles dans le diagramme de dispersion global ; c'est d'ailleurs le cas de la source du milieu sur le diagramme de dispersion de la figure 2.1(b).
- Si des directions du mélange sont proches, il y a un risque important de les confondre lorsque λ est trop grand.
- Si les représentations temps-fréquences des sources ne sont pas assez disjointes, il y aura vraisemblablement des points temps-fréquence de grande intensité où plusieurs sources contribuent significativement au mélange. Il y a par conséquent un risque de créer un cluster (pour le K-means), ou bien de détecter un pic (pour la fonction potentielle) avec ces points qui ne correspondent pas à une direction du mélange. Il faut aussi noter que plus le nombre de sources est important dans le mélange, plus la probabilité que ce type de problème se produise augmente.

2.1.4 Extraction des caractéristiques par l'approche locale

Afin de rendre plus robuste l'estimation des directions, Abrard et Deville [AD03] se basent sur une hypothèse beaucoup plus faible que celle du support quasi-disjoint des sources. Ils supposent simplement qu'il existe pour chaque source au moins une région temps-fréquence où l'intensité de cette source est très supérieure à la contribution des autres sources. Leur méthode TIFROM ne se base pas sur le diagramme de dispersion global qui a été utilisé précédemment, mais sur des diagrammes de dispersion locaux

contruits à partir des points contenus dans des régions temps-fréquence $\Omega_{t,f}$ de taille (fixe) $|\Omega| = 2\,\mathcal{L}+1$ et définies par :

$$\Omega_{t,f} = \{(t + k\,\mathrm{L}\,/2, f) \qquad | \qquad |k| \leq \mathcal{L}\} \tag{2.11}$$

La figure 2.3 illustre l'approche locale de TIFROM, où les diagrammes de dispersions sont obtenus à partir de régions temps-fréquence.

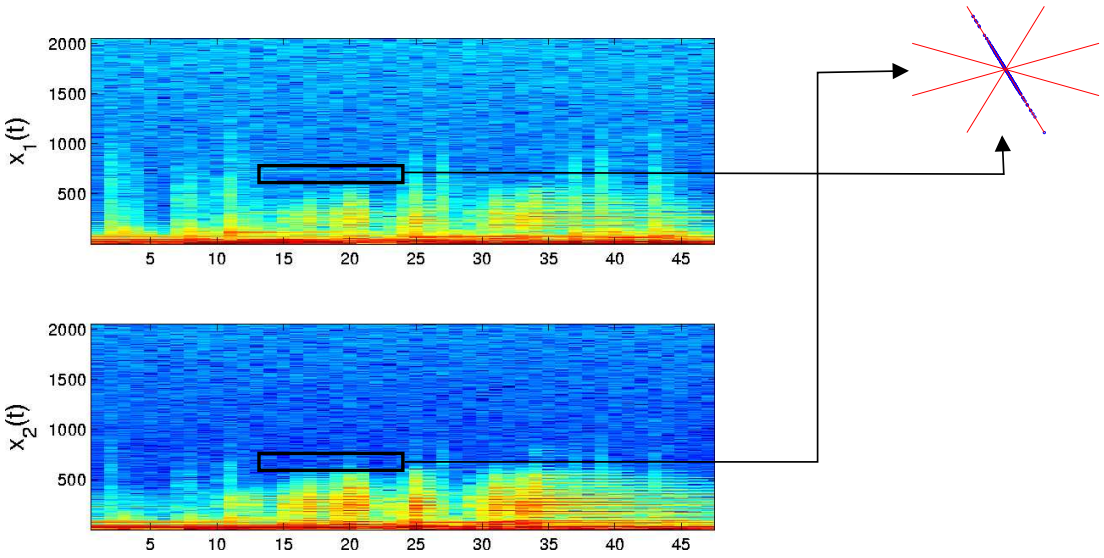


FIG. 2.3 – Approche locale de TIFROM. Les directions sont estimées à partir du diagramme de dispersion des points contenus dans une région temps-fréquence où une seule source est active.

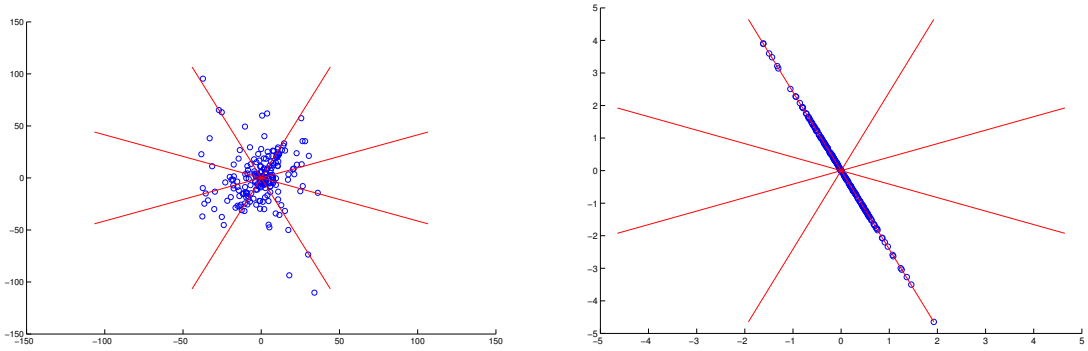
La figure 2.4(a) montre le diagramme de dispersion local d'une région temps-fréquence où plusieurs sources contribuent au mélange. On ne voit pas apparaître d'alignements ou de clusters de points le long d'une quelconque direction du mélange. Par contre, sur le diagramme de dispersion de la figure 2.4(b), on observe très clairement l'alignement des points le long de l'une des directions du mélange. Ainsi, à condition de savoir détecter ces régions, il est facile d'identifier les directions du mélange, si pour chaque source il existe au moins une région comme celle-ci.

Pour détecter ces régions, l'algorithme TIFROM [AD03] calcule la variance du rapport

$$R_{21}(t,f) \triangleq \frac{X_2(t,f)}{X_1(t,f)}, \tag{2.12}$$

des points contenus dans chaque région $\Omega_{t,f}$:

$$\mathrm{Var}\left(R_{21}\left(\Omega_{t,f}\right)\right) = \frac{1}{|\Omega_{t,f}|} \sum_{(\tau,\omega)\in\Omega_{t,f}} \left|R_{21}(\tau,\omega) - \bar{R}_{21}(\Omega_{t,f})\right|^2 \tag{2.13}$$



(a) Région temps-fréquence où plusieurs sources sont actives.

(b) Région temps-fréquence où une essentiellement une seule source est active.

FIG. 2.4 – Diagrammes de dispersion locaux de deux régions temps-fréquence. Les droites indiquent les vraies positions des sources du mélange. La fenêtre de la TFCT est de taille L = 4096, et la taille du voisinage est de $|\Omega| = 100$.

et où $\bar{R}_{21}(\Omega_{t,f})$ est la moyenne du rapport R_{21} dans la région $\Omega_{t,f}$:

$$\bar{R}_{21}(\Omega_{t,f}) = \frac{1}{|\Omega_{t,f}|} \sum_{(\tau,\omega)\in\Omega_{t,f}} R_{21}(\tau,\omega). \tag{2.14}$$

Ainsi, si dans le voisinage $\Omega_{t,f}$, une source a une intensité très supérieure à la contribution des autres, alors la variance $\mathrm{Var}\left(R_{21}\left(\Omega_{t,f}\right)\right) \approx 0$ et $\bar{R}_{21}(\Omega_{t,f}) \approx \tan\theta_{n(t,f)}$. Dans le cas opposé, si plusieurs sources ont des intensités non négligeables, alors $\mathrm{Var}\left(R_{21}\left(\Omega_{t,f}\right)\right)$ est significativement différente de zéro [AD03].

2.1.4.1 L'algorithme TIFROM

La méthode TIFROM consiste alors à :

1. sélectionner la région $\Omega_{t,f}$ qui a la plus petite variance $\mathrm{Var}(R_{21})$ parmi les régions temps-fréquence qui n'ont pas encore été sélectionnées ;
2. créer une nouvelle direction $\theta_K = \tan^{-1}\left(\Re\mathrm{e}\bar{R}_{21}(\Omega_{t,f})\right)$ si celle-ci n'est pas à une distance $|R_{21}(\Omega_{t,f}) - R_{21}(\Omega_k)|$ inférieure au seuil ζ d'une direction précédemment créée θ_k ;
3. s'arrêter si le nombre de directions créées a atteint le nombre K de directions spécifiées par l'utilisateur, sinon incrémenter K et revenir à l'étape 1.

2.1.4.2 Critique de la méthode TIFROM

Choix du seuil ζ : Le seuil ζ dépend des données et doit être choisi de manière à ce que la même direction ne soit pas sélectionnée plusieurs fois, mais que deux directions proches l'une de l'autre soient quand même distinguées. Le réglage du paramètre ζ s'avère par conséquent délicat.

Estimateurs symétriques : Nous définissons la propriété de *symétrie* de l'estimateur de variance par le fait que celui-ci donne le même résultat si l'on permute les canaux. L'estimateur de variance $V\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right)$ est symétrique si et seulement si

$$V\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right) = V\left([\pi/2-\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right).$$

Nous définissons la propriété de *symétrie* de l'estimateur de direction $E\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right)$ par le fait qu'il soit équivalent de permuter les canaux sur l'ensemble des points sur lesquels est calculé l'estimateur de direction que sur l'estimateur de direction lui-même, c.-à-d. quand :

$$\pi/2 - E\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right) = E\left([\pi/2-\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right).$$

Estimateurs invariants par rotation : Nous définissons la propriété d'*invariance par rotation* de l'estimateur de variance par le fait que celui-ci donne le même résultat si l'on applique une rotation à l'ensemble des points sur lesquels est calculé l'estimateur de variance. L'estimateur de variance $V\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right)$ est symétrique si et seulement si

$$V\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right) = V\left([\theta(\tau,\omega)+\alpha]_{(\tau,\omega)\in\Omega_{t,f}}\right), \forall\alpha.$$

Nous définissons la propriété d'*invariance par rotation* de l'estimateur de direction $E\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right)$ par le fait qu'il soit équivalent d'appliquer une rotation à l'ensemble des points sur lesquels est calculé l'estimateur de direction que sur l'estimateur de direction lui-même, c.-à-d. quand :

$$E\left([\theta(\tau,\omega)]_{(\tau,\omega)\in\Omega_{t,f}}\right) + \alpha = E\left([\theta(\tau,\omega)+\alpha]_{(\tau,\omega)\in\Omega_{t,f}}\right), \forall\alpha.$$

Remarquons que la propriété d'invariance par rotation implique la propriété de symétrie.

Estimateur de variance de TIFROM : Remarquons que l'estimateur de variance $\mathrm{Var}\left(R_{21}\left(\Omega_{t,f}\right)\right)$ n'est pas symétrique, car à perturbation angulaire égale, les perturbations sur R_{21} seront plus faibles pour les directions qui sont proches du canal X_1 que de celles proches de X_2. En effet, dans le cas où une source S_n domine les autres, $R_{21} \approx \tan\theta_n$, or dans ce cas la perturbation sur R_{21} est approximativement : $\Delta\tan(\theta_n) = \left(1+\tan^2\left(\theta_n+\Delta\theta\right)\right)\Delta\theta$. Par conséquent les perturbations sur R_{21} ne sont pas proportionnelles aux perturbations des directions angulaires $\Delta\theta$ mais dépendent très fortement de la valeur de la direction θ_n. En particulier, si la direction est sur le canal X_1 ($\theta_n = 0$), alors $\Delta R_{21} \approx \Delta\theta$, mais quand la direction est pratiquement sur le canal X_2, ($\theta_n \to \pi/2$), alors $\Delta R_{21} \to \infty$.

Afin d'éviter ce problème, Puigt et Deville ont proposé une version symétrisée de TIFROM [PD05] qui consiste à considérer en plus du rapport R_{21} le rapport $R_{12} = 1/R_{21}$.

2.1.4.3 La méthode TIFCORR

Deville propose une variante de TIFROM appelée TIFCORR [Dev03]. Cette approche est similaire à TIFROM, mais utilise des estimateurs différents pour les étapes de détection des régions temps-fréquence et d'identification des directions.

Détection des régions temps-fréquence : Dans la méthode TIFCORR, la détection des régions temps-fréquence, est basée sur le coefficient de corrélation inter-canal (aussi appelé cohérence inter-canal [AJ02, Vin04]) donné par :

$$\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right) \triangleq \frac{\widehat{R}_{X_1X_2}\left(\Omega_{t,f}\right)}{\sqrt{\widehat{R}_{X_1X_1}\left(\Omega_{t,f}\right)\widehat{R}_{X_2X_2}\left(\Omega_{t,f}\right)}} \tag{2.15}$$

avec :

$$\widehat{R}_{X_iX_j}\left(\Omega_{t,f}\right) \triangleq \frac{1}{\left|\Omega_{t,f}\right|}\sum_{(\tau,\omega)\in\Omega_{t,f}} X_i(\tau,\omega)X_j^*(\tau,\omega) \tag{2.16}$$

Le coefficient $\left|\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right)\right| \leq 1$ vaut 1 si une seule source est active, et à condition que cette source ne soit pas présente uniquement sur l'un des canaux. En effet, dans ce dernier cas, la corrélation est nulle bien qu'une seule source soit active. Dans le cas où une source a une intensité très supérieure à la contribution des autres, et si la direction θ de la source dominante n'est pas trop proche d'un des canaux ($\theta \not\to 0, \theta \not\to \pi/2$), alors la valeur du coefficient de corrélation sera proche de la valeur 1 : $\left|\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right)\right| \lessapprox 1$. Cet estimateur est symétrique, mais par contre il n'est pas invariant par rotation, car plus la direction se rapproche de l'un des canaux ($\theta \to 0, \theta \to \pi/2$), plus la source dominante doit avoir une intensité forte par rapport aux autres sources, afin que le coefficient de corrélation $\left|\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right)\right|$ soit proche de la valeur 1.

Identification des directions : L'estimation de la direction de la source dominante θ dans les régions où $\left|\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right)\right| \lessapprox 1$ est faite à l'aide de l'estimateur : $\widehat{\theta} = \tan^{-1}(\widehat{z})$, où $\widehat{z} \triangleq \Re\mathrm{e}\left(\frac{\widehat{R}_{X_1X_2}(\Omega_{t,f})}{\widehat{R}_{X_1X_1}(\Omega_{t,f})}\right)$. L'estimateur $\widehat{z}$ est en fait l'estimateur des moindres carrés du cœfficient directeur de la droite de régression linéaire passant par l'origine :

$$\widehat{z} = \arg\min_{z\in\mathbb{R}} \left\|\mathbf{X}_2\left(\Omega_{t,f}\right) - z\mathbf{X}_1\left(\Omega_{t,f}\right)\right\|^2,$$

où $\mathbf{X}_m\left(\Omega_{t,f}\right) = \left[X_m(\tau,\omega)\right]_{(\tau,\omega)\in\Omega_{t,f}}$. Cet estimateur est par conséquent asymétrique. En effet, si l'estimateur $\widehat{z}$ était symétrique, alors l'égalité suivante devrait être vraie pour tout ensemble de points $\Omega_{t,f}$:

$$\Re\mathrm{e}\left(\frac{\widehat{R}_{X_1X_2}\left(\Omega_{t,f}\right)}{\widehat{R}_{X_1X_1}\left(\Omega_{t,f}\right)}\right) = 1 \Big/ \Re\mathrm{e}\left(\frac{\widehat{R}_{X_1X_2}\left(\Omega_{t,f}\right)}{\widehat{R}_{X_2X_2}\left(\Omega_{t,f}\right)}\right) \tag{2.17}$$

Or si $\Re\mathrm{e}\left(\widehat{R}_{X_1X_2}\right) = \widehat{R}_{X_1X_2}$, alors l'égalité (2.17) n'est vérifiée que dans le cas particulier où $\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right) = \pm 1$.

2.2 Mélange anéchoïque

Nous présentons dans cette section le modèle de mélange anéchoïque, ainsi que l'extension des approches du cas instantané pour l'estimation des paramètres du mélange anéchoïque, en particulier le paramètre de délai de chacune des directions.

2.2.1 Le modèle de mélange anéchoïque

Afin de modéliser un enregistrement d'une scène auditive effectué à l'aide de microphones placés dans une salle dans laquelle les parois sont supposées anéchoïques, on utilise le modèle de mélange suivant :

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}) + b_m(t), \qquad 1 \leq m \leq M \tag{2.18}$$

La seule différence par rapport au modèle instantané de l'équation (2.2), est l'introduction des délais δ_{mn}. Le délai δ_{mn} représente l'intervalle de temps entre l'émission du son par la source n et la captation de celui-ci par le microphone m. Ce délai est indéterminable dans l'absolu, mais on peut le caractériser par la différence de temps d'arrivée entre les canaux. Pour fixer cette indétermination, on supposera que $\delta_{1n} = 0$.

De la même façon que dans le cas instantané, nous supposons que $\sum_{m=1}^{M} a_{mn}^2 = 1$ et que $a_{1n} \geq 0$ pour $1 \leq n \leq N$, afin de fixer l'indétermination du facteur d'échelle.

L'équation (2.18) du mélange anéchoïque peut s'écrire de façon approximative dans le domaine de la TFCT par l'équation :

$$\mathbf{X}(t, f) = \mathbf{A}(f)\mathbf{S}(t, f) + \mathbf{B}(t, f),$$

où $\mathbf{A}(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_n(f)]$ est la matrice $M \times N$ dont les vecteurs colonnes $\mathbf{a}_n(f) = [a_{1n}, a_{2n}e^{-2i\pi\delta_{2n}f}, \ldots, a_{Mn}e^{-2i\pi\delta_{Mn}f}]^T$ sont appelés les directions des sources.

Dans le cas stéréophonique ($M = 2$), chaque colonne de $\mathbf{A}(f)$ est un vecteur complexe à deux dimensions qui peut s'écrire :

$$\mathbf{a}_n(f) = \begin{bmatrix} \cos\theta_n \\ \sin\theta_n \cdot e^{-2i\pi\delta_n f} \end{bmatrix} \in \mathbb{C}^2. \tag{2.19}$$

Le paramètre $\theta_n \in]-\pi/2, \pi/2]$ est le *paramètre d'intensité* de la direction n qui est défini de la même façon que dans le cas instantané. Le paramètre $\delta_n \in \mathbb{R}$ caractérise le délai de la source n entre les deux canaux. Une direction n est alors entièrement définie par la paire (θ_n, δ_n). Dans le cas où le mélange est sur $M > 2$ canaux, on peut généraliser cette paire, en définissant le *profil d'intensité* et le vecteur des délais $\Delta_n = [\delta_{1n}, \delta_{2n}, \ldots, \delta_{Mn}]^T \in \mathbb{R}^M$ avec $\delta_{1n} = 0$, de la direction n. Nous définissons le *profil d'intensité* de la direction n par $\mathrm{abs}\,(\mathbf{a}_n(f)) \in \mathbb{R}^M$ avec $\|\,\mathrm{abs}(\mathbf{a}_n(f))\|^2 = 1$, où $\mathrm{abs}(\cdot)$ est une fonction de $\mathbb{C}^M$ dans $\mathbb{R}^M$ qui à tout élément du vecteur de $\mathbb{C}^M$ fait correspondre le module de cet élément.

2.2.2 Extraction des paramètres d'intensité et des délais

La plupart des méthodes d'estimation des directions d'un mélange anéchoïque se basent, de la même façon que pour le cas instantané, sur l'hypothèse du support quasi-disjoint des souces dans la représentation temps-fréquence. Si l'on suppose qu'au point temps-fréquence (t, f), la source $S_{n(t,f)}(t, f)$ a une intensité très supérieure à la contribution des autres sources $n \neq n(t, f)$ ($|S_{n(t,f)}(t, f)| \gg |S_n(t, f)|$), alors compte tenu de la définition d'une direction de l'équation (2.19), le rapport $R_{21}(t, f)$ défini à l'équation (2.12) prend la forme suivante :

$$R_{21}(t, f) \approx \tan \theta_{n(t,f)} \cdot e^{-2i\pi\delta_{n(t,f)}f}.$$

Dans DUET, on calcule les estimations locales du paramètre d'intensité et du délai par les estimateurs suivants [YR04] :

$$\theta(t, f) \triangleq \tan^{-1} |R_{21}(t, f)| \tag{2.20}$$

$$\delta(t, f) \triangleq -\frac{1}{2\pi f} \angle R_{21}(t, f) \tag{2.21}$$

où $\angle z \in]\pi, \pi]$ est la phase du nombre complexe z.

A partir de ces estimées locales, les directions sont obtenues par une méthode de clustering. Dans DUET une fonction potentielle à deux dimensions (une pour la différence d'intensité, l'autre pour le délai) est construite, puis un algorithme de K-means cherche à identifier les pics de la fonction potentielle correspondant aux directions.

Limites des estimateurs de DUET : L'approche DUET est parfaitement valide si la valeur du délai est inférieure à un échantillon et si les gains a_{mn} sont tous positifs.

En effet, si au point (t, f), la source d'indice $n(t, f)$ est la seule source active, l'estimation locale $\theta(t, f)$ de DUET vaudra : $\theta(t, f) \approx |\theta_{n(t,f)}|$, or $\theta_{n(t,f)}$ est négatif si a_{2n} est négatif ($a_{1n} \geq 0$). D'autre part, comme expliqué à la section 2.2.3 ci-dessous, l'estimation du délai par l'équation (2.21) est ambigüe si l'on suppose que la valeur des délais peut être supérieure à un échantillon. Pour se rendre compte de l'implication de cette contrainte, remarquons que si la fréquence d'échantillonnage est de $f_e = 44.1$ kHz (fréquence d'échantillonnage d'un CD audio), alors la distance parcourue par le signal pendant une période d'échantillonnage est de $c/f_e = 7.8$ mm, où $c = 344$ m/s est la vitesse de propagation du son. Ainsi, afin de garantir des délais inférieurs à un échantillon du signal quelle que soit la position des sources, la distance d_{mic} entre microphones doit être au plus $d_{mic} = c/f_e$, qui dans le cas d'un échantillonnage à 44.1 kHz correspond à une distance de 7.8 millimètres.

Remarque sur l'utilisation du diagramme de dispersion réel : Contrairement au cas instantané, les directions étant maintenant des vecteurs complexes, on ne peut plus utiliser les diagrammes de dispersions réels construits avec les points $\mathbf{X}^{\Re\mathrm{e}}(t, f)$ et $\mathbf{X}^{\Im\mathrm{m}}(t, f)$. Pour s'en rendre compte, on peut remarquer que si $S_n(t, f) = \rho_n(t, f)e^{i\phi_n(t,f)}$ est l'unique source active, alors :

$$\left\{ \begin{array}{lcl} R_{21}^{\Re\mathrm{e}}(t,f) & \approx & \tan\theta_n \frac{\cos(\phi_n(t,f)-2i\pi\delta_n f)}{\cos(\phi_n(t,f))} \\ R_{21}^{\Im\mathrm{m}}(t,f) & \approx & \tan\theta_n \frac{\sin(\phi_n(t,f)-2i\pi\delta_n f)}{\sin(\phi_n(t,f))} \end{array} \right. \tag{2.22}$$

Contrairement à l'estimateur $R_{21}(t,f)$ de l'équation (2.12), le rapport $R_{21}^{\mathbb{R}}(t,f)$ dépend donc dans le cas anéchoïque de la phase de la source qui est inconnue.

2.2.3 Ambiguïté de l'estimateur des délais de DUET

A cause de la périodicité de la fonction exponentielle complexe, la phase du rapport $R_{21}(t,f)$ est définie à $2k\pi$ près. Par conséquent, si une seule source d'indice $n(t,f)$ est active,

$$\angle R_{21}(t,f) \approx -2\pi\delta_{n(t,f)}f + 2k\pi \quad \in]-\pi,\pi[.$$

Comme la fréquence réduite $f \leq 1/2$, si la valeur du délai est inférieur à un échantillon, c.a.d. $\delta_{n(t,f)} \in [-1,+1[$, alors $k=0$ et

$$\delta(t,f) \approx \delta_{n(t,f)}.$$

Dans ce cas, l'estimateur de DUET de l'équation (2.21) est tout a fait pertinent.

Par contre, si la valeur du délai est supérieure à un échantillon, c.a.d. $\delta_{n(t,f)} \notin [-1,+1[$, alors :

$$\delta(t,f) \approx \delta_{n(t,f)} + \frac{k}{f},$$

où $\delta_{n(t,f)}$ et k sont inconnus.

Par conséquent, si les délais sont supérieurs à un échantillon, il n'est pas possible de les estimer à partir d'un seul point temps-fréquence.

2.2.4 Autres approches pour l'estimation des délais

Afin de résoudre le problème d'ambiguïté de l'estimation du délai à partir d'un seul point temps-fréquence, d'autres approches cherchent à estimer les délais d'une source à partir de plusieurs points temps-fréquence (à différentes fréquences) supposés appartenir à cette même source.

Puigt et Deville étendent les méthodes TIFROM [PD05] et TIFCORR [PD06] en estimant les délais à partir des points contenus dans une région temps-fréquence :

$$\Omega_{t,f}^F = \{(t, f + k/\mathrm{L}) \qquad | \qquad |k| \leq \mathcal{L}\}. \tag{2.23}$$

En constatant que si une seule source est active dans une région $\Omega_{t,f}^F$, alors les estimations de la phase $\phi_t(f') \stackrel{\triangle}{=} \angle R_{21}(t,f') \approx -2i\pi\delta_{n(t,f)}f' + 2k(t,f')\pi$ pour $(t,f') \in \Omega_{t,f}^F$, sont sur une même droite dont la pente est donnée par $\delta_{n(t,f)}$, à condition que les $k(t,f')$ soient tous identiques dans la région $\Omega_{t,f}^F$. Puigt estime alors cette droite par régression linéaire (au sens des moindres carrés), et en déduit la valeur du délai. Il est connu [Sap90] que la variance de l'estimateur de la pente de la droite de régression linéaire est

inversement proportionnelle à la variance des *variables explicatives* $f \in \Omega^F_{t,f}$. Ainsi les points d'une région $\Omega^F_{t,f}$ ayant des fréquences voisines, la variance des *variables explicatives* est faible, et l'erreur commise sur l'estimation de la pente peut par conséquent être importante, même si l'erreur de régression est faible.

Bofill [Bof03] propose de faire d'abord un clustering des points sur la différence d'intensité, qui a pour but de regrouper grossièrement les points qui sont proches d'une même direction, puis d'estimer les délais dans chacun de ces groupes (clusters).

Bofill remarque que si on corrige la phase de $X_2(t,f)$ par le facteur $2\pi\delta f$, les points (t,f) du cluster n, ont pour valeur approximative :

$$\widetilde{R}_{21}(t,f,\delta) \approx \tan\theta_n \cdot e^{-2i\pi(\delta_n-\delta)f},$$

avec $\widetilde{R}_{21}(t,f,\delta) \triangleq \frac{X_2(t,f)\cdot e^{2i\pi\delta f}}{X_1(t,f)}$. Et dans le cas où $\delta = \delta_n$,

$$\widetilde{R}_{21}(t,f,\delta) \approx \tilde{R}^{\mathbb{R}}_{21}(t,f,\delta) \approx \tan\theta_n,$$

avec : $\widetilde{R}^{\mathbb{R}}_{21}(t,f,\delta) \triangleq \frac{\mathbb{R}X_2(t,f)\cdot e^{2i\pi\delta f}}{\mathbb{R}X_1(t,f)}$.

Ainsi, le rapport $\widetilde{R}^{\mathbb{R}}_{21}(t,f,\delta)$, qui pour une valeur de $\delta \neq \delta_n$ dépend de la phase de la source (comme pour l'équation 2.22), est égal à $\tan\theta_n$ quand $\delta = \delta_n$. Pour estimer le délai δ_n, Bofill cherche alors la valeur de δ pour laquelle les points (t,f) forment un cluster, en utilisant la fonction potentielle de la section 2.1.3, mais avec δ comme variable d'entrée. L'idée est intéressante, mais elle repose toujours sur l'utilisation d'une fonction potentielle, dont les paramètres peuvent s'avérer délicats à régler.

2.3 Conclusion

Nous avons présenté dans ce chapitre les principales familles d'approches exploitant la parcimonie pour estimer les paramètres du mélanges dans les cas instantané et anéchoïque.

Les approches présentées dans ce chapitre exploitent la parcimonie de la représentation temps-fréquence des signaux afin d'estimer les paramètres du mélange par *clustering*. L'approche globale consiste à faire le *clustering* sur l'ensemble des points temps-fréquence, avec ou sans l'utilisation d'une fonction potentielle qui transforme le diagramme de dispersion en une fonction discrète semblable à un histogramme.

Contrairement à l'approche globale qui nécessite que les sources aient des supports quasi-disjoints, l'approche locale quant à elle, se base sur l'hypothèse que pour chaque source il y ait au moins une région temps-fréquence où cette source est la seule à être active. Ces régions sont détectées à l'aide de la variance du rapport $R^{\mathbb{R}}_{21}(t,f)$, et les directions sont estimées dans chaque région temps-fréquence à partir des diagrammes de dispersion locaux.

Dans le cas des mélanges anéchoïques, en plus d'estimer la différence d'intensité des directions, il s'agit d'estimer les délais. Les techniques développées pour le cas instantané qui sont basées sur le rapport R_{21} de DUET, sont facilement généralisables

au cas anéchoïque à condition que les délais soient inférieurs à un échantillon, ce qui n'est malheureusement rarement vérifié en pratique. Pour pouvoir estimer des délais supérieurs à un échantillon, on a besoin, pour chaque source, de la contribution de différents points appartenant à cette même source, à des fréquences différentes. Une piste est, comme le suggère les approches AD-TIFROM [PD05] et AD-TIFCORR [PD06], ainsi que la méthode de Bofill [Bof03], d'exploiter la linéarité de la différence de phase, ou bien la constance du paramètre d'intensité, des directions du modèle anéchoïque.

Une limitation générale des méthodes de clustering de l'état de l'art que nous venons de présenter, est que leurs performances reposent sur le réglage de paramètres, dont la valeur adéquate dépend des données. Ainsi la valeur du facteur de dilatation λ de la fonction potentielle, la grille de discrétisation de la fonction potentielle, la valeur du seuil de la méthode TIFROM à partir duquel deux directions sont considérées comme différentes, le nombre de sources du mélange, sont autant de paramètres dont l'ajustement parfois délicat doit être fait « à la main » en fonction des données traitées. De plus, de même qu'il y a une variabilité entre les mélanges qui explique que le réglage optimal des paramètres soit différent d'un mélange à l'autre, il existe également une variabilité entre les sources du mélange, qui laisse penser qu'un réglage des paramètres qui serait le plus adaptatif possible, serait profitable pour estimer les directions de façon robuste et automatique.

Un autre problème des méthodes de l'état de l'art que nous avons présentées est qu'elles reposent souvent sur des estimateurs asymétriques, ou du moins non invariants par rotation. Ainsi, la précision de ces estimateurs varie selon le paramètre d'intensité des directions du mélange.

Chapitre 3

Exploitation de la parcimonie pour estimer les sources

L'étape d'estimation des sources par la parcimonie exploite les informations spatiales du mélange qui ont été estimées lors de l'étape précédente décrite au chapitre 2, ainsi que l'hypothèse de parcimonie et d'indépendance des sources. Cette approche n'est possible que dans le cas où il y a au moins deux canaux ($M \geq 2$). Par contre, contrairement à l'approche de l'ACI classique (voir section 1.1), elle permet l'estimation des sources dans le cas des mélanges sous-déterminés ($N \geq M$). Selon la détermination du mélange, c.-à-d. selon le rapport entre le nombre de sources N et le nombre de canaux M, et l'hypothèse sur le nombre de sources actives en chaque point temps-fréquence, plusieurs approches sont possibles.

3.1 Estimation linéaire dans le cas (sur-)déterminé

Quand le mélange est (sur-)déterminé et non bruité, il suffit d'inverser la matrice de mélange pour estimer les sources.

Mélanges instantanés : Dans le cas où le mélange est instantané, cette inversion peut se faire dans le domaine temporel

$$\widehat{\mathbf{s}}(t) \stackrel{\triangle}{=} \widehat{\mathbf{A}}^{-1}\mathbf{x}(t),$$

ou bien dans le domaine transformé

$$\widehat{\mathbf{S}}(t,f) \stackrel{\triangle}{=} \widehat{\mathbf{A}}^{-1}\mathbf{X}(t,f).$$

Dans ce dernier cas, il est évidemment nécessaire d'appliquer la transformé inverse afin de restaurer les source dans le domaine temporel.

Dans le cas sur-déterminé, il suffit de remplacer la matrice inverse $\widehat{\mathbf{A}}^{-1}$, par la pseudo-inverse $\widehat{\mathbf{A}}^{\dagger}$. Si comme dans le cas de la TFCT, la représentation du signal dans le domaine transformé est exacte, c.-à-d. que la transformée $\Psi(\mathbf{x})$ du signal $\mathbf{x} = [\mathbf{x}(\tau)]_{\tau=1}^{T}$

suivie de sa reconstruction en permette la reconstruction parfaite $\mathbf{x} = \Psi^{-1}(\Psi(\mathbf{x}))$, alors l'estimation des sources dans le domaine temporel ou dans le domaine transformé est équivalente [Gri07]. Si de plus la matrice de mélange est parfaitement estimée, alors l'estimation des sources sera elle aussi parfaite.

Mélanges anéchoïques : Dans le cas anéchoïque, la matrice le mélange ne peut plus être inversée dans le domaine temporel, mais peut toujours l'être dans le domaine transformé :

$$\widehat{\mathbf{S}}(t,f) \stackrel{\triangle}{=} \widehat{\mathbf{A}}^{-1}(f)\mathbf{X}(t,f).$$

Les sources sont ensuite restaurées dans le domaine temporel en appliquant la transformé inverse.

3.2 Hypothèse d'un nombre de sources actives inférieur au nombre de canaux

Dans le cas sous-déterminé, si l'on suppose que le nombre de sources actives J en chaque point temps-fréquence est inférieur au nombre de canaux M, alors on peut estimer les sources en cherchant la combinaison des sources actives telle que leur estimation minimise l'erreur de reconstruction du mélange $\mathbf{X}(t,f)$. Etant donné que cette approche se base sur l'erreur de reconstruction du mélange, elle ne peut pas fonctionner dans le cas où $J \geq M$, puisqu'il est possible dans ce cas de reconstruire parfaitement le mélange quelle que soit la combinaison des J sources actives.

3.2.1 Cas stéréophonique : Masquage binaire

Dans le cas stéréophonique, on part de l'hypothèse qu'il y a au plus une source active, en chaque point temps-fréquence. On suppose ainsi de façon équivalente que les sources ont un support disjoint dans la représentation temps-fréquence.

L'estimation de la source active $n(t,f)$ est obtenue par projection du mélange sur la direction de cette source, tandis que les autres sources sont considérées comme inactives. On parle ainsi de masquage binaire pour désigner cette approche.

$$\widehat{S}_n(t,f) \stackrel{\triangle}{=} \begin{cases} \mathbf{a}_n^H(f)\mathbf{X}(t,f) & si\ n = \tilde{n}(t,f) \\ 0 & si\ n \neq \tilde{n}(t,f) \end{cases} \tag{3.1}$$

où $\tilde{n}(t,f)$ est l'indice estimé de la source active, qui est choisi de façon à minimiser l'erreur de reconstruction :

$$\tilde{n}(t,f) \stackrel{\triangle}{=} \arg\min_n \|\mathbf{X}(t,f) - \mathbf{a}_n(f)\mathbf{a}_n^H(f)\mathbf{X}(t,f)\|^2 \tag{3.2}$$

$$= \arg\max_n \left|\mathbf{a}_n^H(f)\mathbf{X}(t,f)\right| \tag{3.3}$$

La source $\tilde{n}(t,f)$ qui est considérée comme la seule source active au point temps-fréquence (t,f) est celle qui obtient la plus grande corrélation entre sa direction et le

mélange $\mathbf{X}(t,f)$. Cette méthode est une variante de la méthode DUET [YR04], où au lieu d'estimer la source active par la projection $\mathbf{a}_n^H(f)\mathbf{X}(t,f)$ du mélange sur la direction de cette source, on estime la source par masquage d'un des canaux $X_m(t,f)$ du mélange.

Hypothèse du support quasi-disjoint des sources : Nous savons qu'en pratique le support des sources n'est pas disjoint (voir chapitre 2). L'hypothèse faite est alors celle du support quasi-disjoint des sources [YR04].

Ainsi, en pratique l'approche du masquage binaire cherche à estimer pour chaque point temps-fréquence (t,f) la valeur du coefficient $S_n(t,f)$ de la source S_n qui est la plus active, et attribue une valeur nulle aux coefficients des autres sources qui sont supposées être faiblement actives.

3.2.2 Cas général où $J < M$

On suppose maintenant dans le cas où $M > 2$ que le nombre de sources actives en chaque point temps-fréquence est inférieur au nombre de canaux $J < M$, mais possiblement supérieur à 1. Gribonval a proposé une approche [Gri03] qui généralise le masquage binaire. Dans cette approche l'estimation des $J < M$ sources actives au point temps-fréquence (t,f), est obtenue par inversion de la sous-matrice $\mathbf{A}_{\mathcal{J}(t,f)}(f)$ de $\mathbf{A}$ composée des directions de ces J sources :

$$\left\{ \begin{array}{rcll} \left[\widehat{S}_j(t,f)\right]_{j\in\widetilde{\mathcal{J}}(t,f)} & \triangleq & \mathbf{A}^{\dagger}_{\widetilde{\mathcal{J}}(t,f)}(f)\mathbf{X}(t,f) & \\ \widehat{S}_n(t,f) & \triangleq & 0 & si\ n \notin \widetilde{\mathcal{J}}(t,f) \end{array} \right. \tag{3.4}$$

L'ensemble des sources actives $\mathcal{J}(t,f)$ au point (t,f), appelé *pattern d'activité* [VGP07], est estimé de façon à minimiser l'erreur de reconstruction :

$$\widetilde{\mathcal{J}}(t,f) \triangleq \underset{\mathcal{J}\,\backslash|\,\mathcal{J}\,|\leq J}{\arg\min} \|\mathbf{X}(t,f) - \mathbf{A}_{\mathcal{J}}(f)\mathbf{A}^{\dagger}_{\mathcal{J}}(f)\mathbf{X}(t,f)\|^2 \tag{3.5}$$

$$= \underset{\mathcal{J}\,\backslash|\,\mathcal{J}\,|\leq J}{\arg\max} \|\mathbf{A}_{\mathcal{J}}(f)\mathbf{A}^{\dagger}_{\mathcal{J}}(f)\mathbf{X}(t,f)\|^2 \tag{3.6}$$

où $|\,\mathcal{J}\,|$ est le nombre de sources du pattern d'activité.

3.3 Hypothèse d'une distribution parcimonieuse des sources

Si le nombre de sources supposées actives est supérieur ou égal au nombre de canaux M, il n'est plus possible de déterminer les sources actives en utilisant le critère de l'erreur de reconstruction, puisqu'avec n'importe quelle combinaison de $J \geq M$ sources actives, on peut obtenir une reconstruction parfaite du mélange.

Par contre si l'on spécifie une distribution sur les coefficients des sources, il est possible d'estimer les valeurs de celles-ci au Maximum a Posteriori (critère MAP) sous la contrainte d'une reconstruction exacte du mélange.

3.3.1 Modélisation des sources

Une approche couramment employée consiste à supposer que pour chaque point temps-fréquence, les sources sont indépendantes et ont une distribution parcimonieuse, dont les paramètres sont en général fixes afin de ne pas avoir à les estimer. Nous discuterons de la validité de cette dans la conclusion de ce chapitre ainsi que dans l'introduction du chapitre 4.

La distribution sur les sources est généralement une Laplacienne [WSM05, LLGS99, ZP01], mais d'autres distributions sont parfois utilisées, avec entre autres le modèle bi-gaussien [DM04] avec une Gaussienne ayant une grande variance, l'autre ayant une variance quasi-nulle, ou encore la loi de Student [FG06], dont les paramètres sont apris à l'aide d'un échantillonneur de Gibbs [CG92, GG88].

Comme nous l'avons vu à la section 1.5, la loi gaussienne généralisée, dont la Laplacienne est un cas particulier, est particulièrement adaptée pour modéliser une distribution parcimonieuse. Les coefficients de la TFCT étant complexes, on modélise alors le module des coefficients par une Gaussienne généralisée donnée par l'équation (3.7) de paramètres $p > 0$ et $\beta > 0$. La phase est considérée comme uniformément distribuée.

$$P(S_n(t,f)) = p\frac{\beta^{1/p}}{\Gamma(1/p)}e^{-\beta|S_n(t,f)|^p} \tag{3.7}$$

Les paramètres p et β dirigent respectivement la forme et la variance de la distribution. Dans le cas où $p = 2$, la loi est une Gaussienne, et quand $p = 1$ elle devient Laplacienne.

3.3.2 Critère du Maximum A Posteriori

L'estimation MAP des sources pour un jeu de paramètres (β, p) constant est alors donnée par l'équation (3.8).

$$\widehat{\mathbf{S}}(t,f) = \underset{\mathbf{S}\backslash\mathbf{A}(f)\mathbf{S}=\mathbf{X}(t,f)}{\arg\min} \|\mathbf{S}\|_p \tag{3.8}$$

où $\|\mathbf{S}\|_p$ est la norme l_p de $\mathbf{S}$ définie par $\|\mathbf{S}\|_p \stackrel{\triangle}{=} \sum_{n=1}^{N} |S_n|^p$.

Le critère est considéré comme parcimonieux pour les valeurs de p comprises entre 0 et 1, et la parcimonie du critère augmente quand p diminue. Remarquons au passage que lorsque $p = 2$, les sources sont supposées gaussiennes et identiquement distribuées et l'estimation des sources par le critère (3.8) n'est autre que la pseudo-inversion du mélange.

3.3.3 Méthodes de résolution du critère MAP

La résolution de l'équation (3.8) n'est pas triviale en particulier dans les cas où le paramètre p est inférieur à 1. En effet lorsque $p = 1$, il s'agit d'un problème convexe du second ordre (SOCP) qui peut être résolu de façon efficace [WSM05, Bof03]. Dans

le cas ou $p < 1$, le problème n'est pas convexe, néanmoins Vincent [Vin07] propose un algorithme d'optimisation globale du critère dans le cas où $N = M + 1$.

Si les coefficients de la transformée utilisée sont réels, c'est le cas si l'on utilise une MDCT au lieu de la TFCT par exemple, le problème (3.8) se simplifie. Dans le cas où $p = 1$, le problème se réduit à un problème classique de programmation linéaire. Dans le cas où $0 \leq p < 1$, il existe des algorithmes itératifs comme FOCUSS [GGR95], dont l'utilisation est assez lourde, cependant, il a été prouvé que dans le cas réel, la résolution du problème (3.8) conduisait à une solution où le nombre de sources actives est au plus $J = M$ [KDRE+99]. Par conséquent, dans le cas où le nombre de sources N et le nombre de capteurs M n'est pas trop grand, on peut employer l'approche combinatoire qui consiste à choisir l'ensemble de M sources actives telles que :

$$\widetilde{\mathcal{J}}(t,f) \stackrel{\triangle}{=} \underset{\mathcal{J} \setminus |\mathcal{J}|=M}{\arg\min} \|\mathbf{A}_{\mathcal{J}}(f)^{-1}\mathbf{X}(t,f)\|_p \tag{3.9}$$

Le nombre de combinaisons à tester est $\binom{M}{N}$ dans le cas général, mais n'est que de $\frac{N(N-1)}{2}$ dans le cas stéréophonique.

En pratique, bien que dans le cas complexe le nombre de sources actives puisse être supérieur à M, l'approche combinatoire décrite ci-dessus peut être employée sans qu'il y ait de grosses pertes dans la qualité en séparation [WSM05, Vin07]. Il semble aussi, suivant la conjecture et les vérifications expérimentales de Vincent [Vin07] que lorsque p est inférieur à une certaine valeur p_{crit} qui dépend du nombre de sources, alors le nombre de sources sélectionnées n'excède pas M, et par conséquent l'approche combinatoire se justifie pour des p faibles.

Dans le cas des mélanges instantanés, certaines approches [BZ01] simplifient le problème (3.8) pour des coefficients complexes en divisant l'équation du mélange $\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f)$ en deux mélanges réels $\Re\mathrm{e}\mathbf{X}(t,f) = \mathbf{A}\Re\mathrm{e}\mathbf{S}(t,f)$ et $\Im\mathrm{m}\mathbf{X}(t,f) = \mathbf{A}\Im\mathrm{m}\mathbf{S}(t,f)$, et résolvent le problème (3.8) indépendamment pour la partie réelle et la partie imaginaire du mélange. La valeur complexe de $\widehat{\mathbf{S}}(t,f)$ est ensuite combinée à partir de $\Re\mathrm{e}\widehat{\mathbf{S}}(t,f)$ et de $\Im\mathrm{m}\widehat{\mathbf{S}}(t,f)$. Cette dernière approche suppose implicitement l'indépendance entre la partie réelle et imaginaire de $\mathbf{S}(t,f)$, qui n'a pas de justification expérimentale.

3.4 Conclusion

Nous avons décrit dans ce chapitre différentes approches pour estimer les sources lorsque les directions du mélange sont connues. En particulier, nous avons vu que lorsque le mélange est déterminé ou sur-déterminé, les sources peuvent être estimées par inversion du mélange, sans qu'aucune autre hypothèse ne soit faite. Dans le cas où le nombre de sources est supérieur au nombre de capteurs, le problème est sous-déterminé, et le nombre de solutions du problème est théoriquement infini, cependant l'hypothèse de parcimonie qui peut être exploitée de différentes façons permet d'obtenir une estimée unique des sources. Ainsi si l'on suppose que le nombre de sources actives est inférieur au

nombre de canaux, on peut exploiter l'erreur de reconstruction du mélange comme critère d'estimation des sources, ou bien si l'on suppose que le nombre de sources actives peut atteindre le nombre de capteurs, alors la reconstruction étant parfaite, on peut utiliser un critère de parcimonie sur les sources pour estimer celle-ci de façon univoque.

Les méthodes d'estimation des sources basées sur la parcimonie que nous avons présenté dans ce chapitre sont très utilisées en pratique car elle sont rapides et donnent globalement de bons résultats en séparation. Cependant ces méthodes commettent souvent des erreurs quand plusieurs sources sont actives simultanéments. La raison est que ces méthodes se basent sur un modèle parcimonieux des sources ayant le même jeu de paramètres pour chacune des sources, si bien que seule la *diversité spatiale* est exploitée afin de discriminer les sources. Pour cette raison, comme le constate Xiao [XXF05] les sources qui sont considérées comme actives sont toujours celles dont les directions sont les plus proches de la direction observée du mélange. Autrement dit, dans le cas sous-déterminé, la plupart des combinaisons possibles de sources actives ne sont pas pris en compte par les approches basées sur la parcimonie.

Chapitre 4

Estimation des sources à l'aide d'un MMG spectral

Dans le chapitre précédent, nous avons vu que pour la plupart des méthodes basées sur la parcimonie, le nombre de sources pouvant être estimées ne peut être supérieur au nombre de canaux. Par conséquent il n'est pas possible avec ces méthodes de séparer des sources par la parcimonie dans le cas monophonique où il n'y a qu'un seul capteur ($M = 1$).

D'autres part, les approches basées sur la parcimonie que nous avons présentées supposent que les sources ont toutes la même distribution, alors qu'en pratique les sources ont souvent des propriétés statistiques diverses.

Dans ce chapitre nous faisons le point sur une approche qui se base sur un modèle spectral des sources et qui permet d'estimer les sources quelque soit la détermination du mélange, y compris dans le cas monophonique. Dans cette approche, la distribution d'une source S_n sur une trame t de la TFCT, $S_n(t) \stackrel{\triangle}{=} [S_n(t,f)]_f$, est modélisée par un Modèle de Mélange de Gaussiennes (MMG). Chaque gaussienne de ce MMG spectral peut être interprétée comme une Densité Spectrale de Puissance (DSP) typique d'une source.

L'utilisation des MMG spectraux pour modéliser les sources sonores a été proposée initialement par Ephraim [Eph92] pour le débruitage et par Benaroya [BB03] pour la séparation de source monophonique. Ces modèles sont appris à partir de données d'apprentissage qui sont les signaux sources eux-mêmes (ce qui n'est pas réaliste), ou bien un ensemble de signaux qui doivent avoir des propriétés similaires aux sources du mélange à séparer. Ozerov [Oze06] a étudié le formalisme de Benaroya, par un principe dit d'adaptation, qui permet d'affiner des modèles « généraux » qui ont été appris à partir de données d'apprentissage, à partir de l'observation du mélange. Il a appliqué avec succès ce formalisme pour la séparation voix/musique. Le SDR s'en est trouvé amélioré de 5 dB par rapport aux performances de séparation des modèles non-adaptés.

Dans ce chapitre, nous introduisons le modèle MMG spectral, et nous rappelons la méthode d'estimation des sources connaissant l'ensemble des paramètres des modèles MMG spectraux des sources, telle qu'elle a été utilisée par Benaroya et Ozerov,

c.-à-d. dans le cas monophonique à 2 sources. Nous généralisons ensuite ces formules au cas multicanal à N sources. Cette généralisation constitue une contribution mineure étant donné que ces formules ont déjà été exposées pour des MMG scalaires réels [Att99].

4.1 Le modèle des sources

On reprend ici la formulation du modèle MMG spectral telle que décrite par Ozerov [Oze06]. Les coefficients de la TFCT de la source S_n à la trame t, sont modélisés par un MMG multivarié complexe circulaire, dont la densité de probabilité est donnée par :

$$P(S_n(t)|\,\lambda_n) = \sum_{k_n} \pi_{n,k_n}\, N_c(S_n(t); \bar{0}, \Sigma_{n,k_n}) \tag{4.1}$$

où $N_c(V; \mu, \Sigma)$ est la densité de probabilité du vecteur aléatoire gaussien complexe cirlulaire V de taille d définie par :

$$N_c(V; \mu, \Sigma) = \pi^{-d} \left[\det(\Sigma)\right]^{-1} \exp\left[-(V-\mu)^H \Sigma^{-1} (V-\mu)\right], \tag{4.2}$$

où $\lambda_n \stackrel{\triangle}{=} \{\pi_{n,k_n}, \Sigma_{n,k_n}\}_{k_n}$ est l'ensemble des paramètres du MMG de la source S_n, π_{n,k_n} est le poids de la gaussienne k_n du MMG de la source S_n, et $\Sigma_{n,k_n} \stackrel{\triangle}{=} \mathrm{diag}([\sigma^2_{n,k_n}(f)]_f)$ est la matrice de covariance de la gaussienne k_n du MMG de la source S_n. On suppose que les sources sont localement stationnaires, ce qui signifie que les composantes du vecteur aléatoire $S_n(t)$ sont décorrélées [Pic94], et par conséquent que la matrice de covariance Σ_{n,k_n} est diagonale. Par conséquent, chaque gaussienne $N_c(S_n(t); \bar{0}, \Sigma_{n,k_n})$ du MMG peut s'écrire :

$$N_c(S_n(t); \bar{0}, \Sigma_{n,k_n}) = \prod_f \frac{1}{\pi \cdot \sigma^2_{n,k_n}(f)} \exp\left[-\frac{|S_n(t,f)|^2}{\sigma^2_{n,k_n}(f)}\right]$$

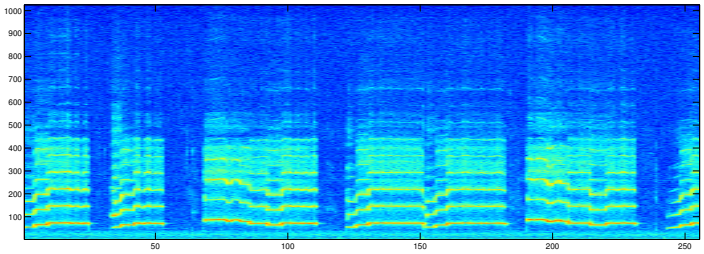
La figure 4.1 donne deux exemples de MMG spectraux à 32 états. On a représenté sur cette figure les DSP $\sigma^2_{n,k_n} = [\sigma^2_{n,k_n}(f)]_f$ pour $1 \leq k_n \leq 32$ des MMG λ_n de deux sources $S_n, n = \{1, 2\}$.

4.2 Estimation des 2 sources d'un mélange monophonique

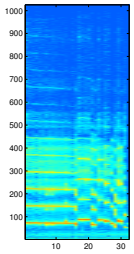
Benaroya et Ozerov considèrent le mélange monophonique $x(t) = s_1(t) + s_2(t)$, qui dans le domaine de la TFCT s'écrit :

$$X(t,f) = S_1(t,f) + S_2(t,f) \tag{4.3}$$

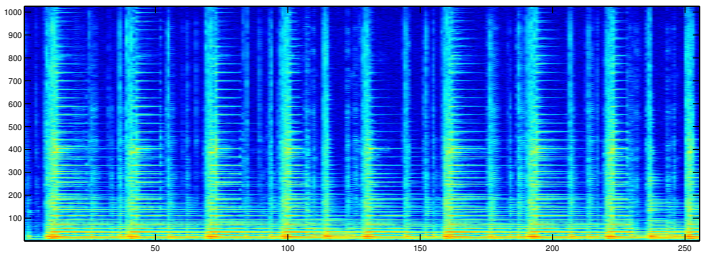
Si maintenant on suppose que l'on connait les modèles $\mathbf{\Lambda} = \{\lambda_1, \lambda_2\}$ des deux sources que l'on souhaite séparer, alors l'estimation de la source S_1 par l'estimateur minimisant



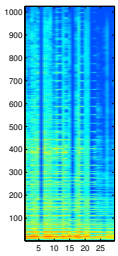
(a) spectrogramme d'un signal de flûte



(b) MMG spectral à 32 états de la source de flûte, chaque état est représenté par sa DSP



(c) spectrogramme d'un signal de guitare



(d) MMG spectral à 32 états de la source de flûte, chaque état est représenté par sa DSP

FIG. 4.1 – exemple de modèles MMG spectraux à 32 états pour modéliser le spectre des sources.

le critère du Minimum de l'Erreur Quadratique Moyenne (MEQM) spectrale défini par :

$$\mathbb{E}\left\{\|\widehat{S}_1 - S_1\|^2\right\} \stackrel{\triangle}{=} \sum_{t,f} \mathbb{E}\left\{|\widehat{S}_1(t,f) - S_1(t,f)|^2\right\} \tag{4.4}$$

est obtenu par le filtre de Wiener pondéré :

$$\begin{aligned}\widehat{S}_1(t,f) &= \mathbb{E}\left\{S_1(t,f)\,|X(t);\mathbf{\Lambda}\right\} \\ &= \sum_{k_1,k_2} \gamma_{k_1,k_2}(t) \frac{\sigma^2_{1,k_1}(f)}{\sigma^2_{1,k_1}(f) + \sigma^2_{2,k_2}(f)} X(t,f)\end{aligned} \tag{4.5}$$

avec $X(t) \stackrel{\triangle}{=} [X(t,f)]_f$, $\sum_{k_1,k_2} \gamma_{k_1,k_2}(t) = 1$ et $\gamma_{k_1,k_2}(t)$ est la probabilité a posteriori que

la source S_1 soit dans l'état k_1 et la source S_2 dans l'état k_2 à la trame t :

$$\begin{aligned}\gamma_{k_1,k_2}(t) &\stackrel{\triangle}{=} P(k_1,k_2|X(t);\mathbf{\Lambda}) \\ &\propto \pi_{1,k_1}\,\pi_{2,k_2}\,N_c(X(t);\bar{0},\Sigma_{1,k_1}+\Sigma_{2,k_2}).\end{aligned} \tag{4.6}$$

L'estimation de la source S_2 est effectuée de manière symétrique en inversant les indices 1 et 2.

Ainsi l'estimation des sources s'effectue pour chaque trame en deux étapes :

1. Calcul des probabilités a posteriori des états $\gamma_{k_1,k_2}(t)$ par l'équation (4.6) ; Cette étape est souvent appelée décodage des états.
2. Estimation des sources par le filtrage de Wiener pondéré de l'équation (4.5).

En pratique, on constate [Oze06] qu'à l'étape d'estimation des sources par l'équation (4.5), il y a peu de différence entre faire la somme pondérée sur l'ensemble des états, et faire l'approximation consistant à ne garder qu'un seul filtre correspondant au couple d'état le plus probable. Cet estimateur approché appelé *estimateur dur* par opposition à l'estimateur de l'équation (4.5) appelé *estimateur doux*, peut permettre une accélération de l'algorithme. Dans le cas où l'on choisit d'utiliser l'*estimateur dur*, l'estimation des sources s'effectue alors pour chaque trame suivant les deux étapes :

1. Calcul du couple d'état le plus probable $(k_1^*(t),k_2^*(t)) = \arg\max_{(k_1,k_2)} \gamma_{k_1,k_2}(t)$;
2. Estimation des sources par le filtre de Wiener : $\widehat{S}_1(t,f) = \frac{\sigma^2_{1,k_1^*(t)}(f)}{\sigma^2_{1,k_1^*(t)}(f)+\sigma^2_{2,k_2^*(t)}(f)} X(t,f)$

4.3 Estimation des N sources d'un mélange à M canaux

Nous présentons dans cette section la généralisation à N sources et M canaux de l'estimation spectrale des sources connaissant les modèles des sources $\mathbf{\Lambda}$, qui a été présenté à la section 4.2. Cette estimation, qui est maintenant spatio-spectrale, nécessite la connaissance de la matrice de mélange $\mathbf{A}$ en plus des modèles des sources $\mathbf{\Lambda}$.

L'équation du filtrage de Wiener pondéré pour estimer le vecteur des sources $\mathbf{S}$ au point temps-fréquence (t,f), qui dans le cas monophonique à deux sources est donné par l'équation (4.5), s'obtient dans le cas d'un mélange à M canaux et N sources, de façon similaire à la dérivation qui a été faite pour les MMG scalaires [Att99, BMC97] :

$$\widehat{\mathbf{S}}(t,f) = \sum_{\mathbf{k}} \gamma_{\mathbf{k}}(t)\mathbf{W}_{\mathbf{k}}(f)\mathbf{X}(t,f) \tag{4.7}$$

où $\mathbf{k} \stackrel{\triangle}{=} [k_1,k_2,\cdots,k_N]^T$ et $\mathbf{\Sigma}_{\mathbf{k}}(f) \stackrel{\triangle}{=} \mathrm{diag}([\sigma^2_{1,\mathbf{k}}(f),\sigma^2_{2,\mathbf{k}}(f),\cdots,\sigma^2_{N,\mathbf{k}}(f)])$.
$\gamma_{\mathbf{k}}(t)$ est la probabilité a posteriori des états à la trame t :

$$\gamma_{\mathbf{k}}(t) \stackrel{\triangle}{=} P(\mathbf{k}|\mathbf{X}(t);\mathbf{A},\mathbf{\Lambda}) \propto \pi_{\mathbf{k}} \prod_f N_c\left(\mathbf{X}(t,f);\bar{0},\mathbf{A}\mathbf{\Sigma}_{\mathbf{k}}(f)\mathbf{A}^T\right) \tag{4.8}$$

où $\mathbf{X}(t) \stackrel{\triangle}{=} [\mathbf{X}(t,f)]_f$ et le filtre de Wiener spatio-spectral est donné par :

$$\mathbf{W_k}(f) \stackrel{\triangle}{=} \mathbf{\Sigma_k}(f)\mathbf{A}^T \left(\mathbf{A}\,\mathbf{\Sigma_k}(f)\mathbf{A}^T\right)^{-1} \tag{4.9}$$

La généralisation à N sources de la formule d'estimation des probabilités a posteriori des états du MMG du mélange, donnée par l'équation (4.8), n'est applicable en pratique que si le nombre de sources N et le nombre d'états $\mathcal{K}$ par source n'est pas trop élevé, car le nombre d'état du mélange, est maintenant de $\mathcal{K}^N$.

L'équation du filtre de Wiener de l'équation (4.9) a été utilisée par Vincent [VR04] qui a évalué notamment l'intérêt d'une séparation spatio-spectrale plutôt que uniquement spatiale ou spectrale, mais avec un modèle des sources spécifiques aux instruments de musique harmoniques.

4.4 Conclusion

Nous avons présenté dans ce chapitre une approche de séparation de sources utilisant des modèles MMG spectraux des sources. Cette approche à été utilisée dans le cas de mélanges monophoniques à deux sources, où les approches spatiales basées sur la parcimonie sont inadaptées. L'extension de cette approche à N sources et M canaux est obtenu en utilisant le formalisme EM pour l'apprentissage de MMG scalaire à partir d'un mélange multicanal [Att99, BMC97], que nous présentons au chapitre 5. Cependant il faut souligner que l'étape de décodage des états ayant une complexité exponentielle en fonction du nombre de sources, cette méthode n'est pas applicable en pratique pour un nombre de sources élevé. Bien qu'il ne semble pas exister de résultats expérimentaux de cette approche spatio-spectrale de séparation de sources, les résultats obtenus par Vincent [VR04] avec un modèle des sources différent, mais basé sur le filtrage de Wiener, montrent que l'utilisation conjointe des informations spatiales et spectrales permet d'améliorer les performances en séparation par rapport à une séparation uniquement spatiale ou uniquement spectrale. Les approches spectrales présentées dans ce chapitre nécessitent néanmoins une étape préalable d'apprentissage des modèles qui fait l'objet du prochain chapitre.

Chapitre 5

Apprentissage de Modèles de Mélange de Gaussiennes (MMG)

Au chapitre précédent, nous avons présenté une approche d'estimation des sources à l'aide d'une modélisation des DSP des sources par des Modèles de Mélange de Gaussiennes (MMG spectraux). Cependant, l'utilisation d'une telle approche nécessite de disposer des paramètres de ces modèles pour chacune des sources. Une étape d'apprentissage des modèles est par conséquent nécessaire à l'utilisation d'une telle approche.

Les approches proposées par Benaroya [BB03] et Ozerov [Oze06] apprennent les paramètres des MMG spectraux des sources par l'algorithme EM, à partir de signaux d'apprentissage qui doivent avoir des propriétés spectrales les plus semblables possibles des sources du mélange. Cette approche n'est par conséquent pas aveugle, car elle nécessite de disposer d'un ensemble d'apprentissage pour chacune des sources du mélange.

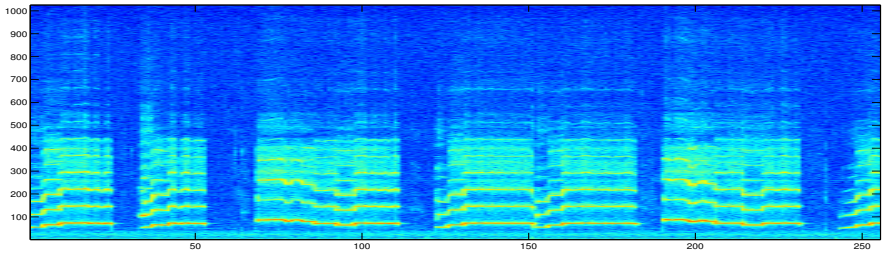
Une approche aveugle [Att99, BMC97] consiste à reprendre le formalisme développé à la section 1.3 pour des sources gaussiennes, et à l'étendre pour des sources MMG. Cette approche a été formalisée dans un cadre général, où les paramètres du mélange et la covariance du bruit sont estimés à chaque itération en plus des paramètres de la distribution des sources. Cependant, rien n'empêche d'utiliser cette approche uniquement pour estimer les paramètres de la distribution des sources si l'on connaît les paramètres du mélange et du bruit. Il suffit alors de ne pas ré-estimer à chaque itération de l'algorithme EM les paramètres que l'on connait.

5.1 Apprentissage hors-ligne

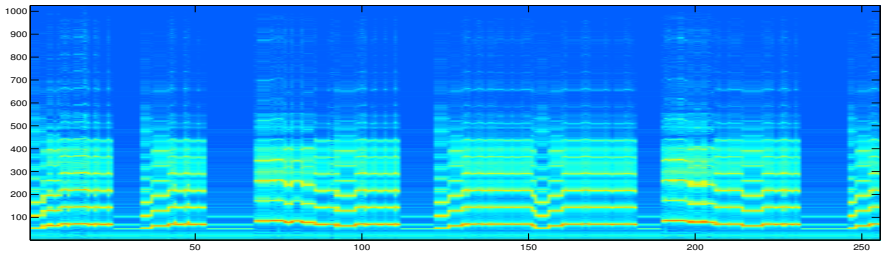
Les approches proposées par Benaroya [BB03] et Ozerov [Oze06] apprennent les paramètres des modèles MMG par un algorithme EM qui cherche à maximiser les vraisemblances $P(\bar{S}_1|\,\lambda_1)$ et $P(\bar{S}_2|\,\lambda_2)$, où $\bar{S}_1$ et $\bar{S}_2$ sont les signaux d'apprentissage. Le détail de l'algorithme EM donné par Benaroya [BB03] et rappelé par Ozerov [Oze06] est résumé en annexe par l'algorithme 1 à la section B.1.

La figure 5.1 illustre les étapes d'apprentissage à l'aide de l'algorithme 1 de l'annexe B.1, et de décodage par l'estimation dure, sur la source (oracle) S. Nous parlons

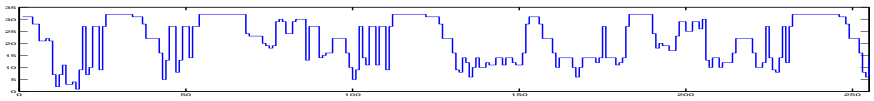
d'apprentissage oracle (respectivement le décodage oracle), quand l'apprentissage (respectivement le décodage) est effectué à partir de la source S du mélange qui n'est évidemment pas disponible dans les cas pratiques.



(a) spectrogramme d'un signal de flûte



(b) spectrogramme représentant les variances des Gaussiennes du MMG décodé



(c) indices des états décodés du MMG

FIG. 5.1 – Illustration de l'apprentissage et du décodage oracle (sur la source) d'un MMG à 32 états par l'estimation dure. La ressemblance entre le spectrogramme du signal de flûte et les variances des Gaussiennes du MMG décodé par l'estimation dure illustre l'adéquation du modèle MMG aux données. En particulier on peut remarquer que certains états du MMG semblent correspondre à une note en particulier, ou à un silence.

Limites de l'approche : L'apprentissage hors-ligne nécessite de disposer des sources d'apprentissage $\bar{S}_n$ correspondant aux sources S_n du mélange. Dans le cas multi-canal, il est aussi nécessaire de faire correspondre les modèles MMG aux directions correspondantes. Dans la pratique, il est difficile de disposer de tels ensembles d'apprentissage, qui doivent avoir des propriétés statistiques les plus proches possibles des sources du mélange, pour que la séparation de sources soit de bonne qualité [Oze06]. Pour éviter d'utiliser des modèles génériques appris hors-ligne, Ozerov [Oze06] a développé un for-

malisme d'adaptation qui permet d'ajuster le modèle d'une source à partir du mélange. Le principe de l'approche est de maximiser non pas le critère MV, mais le critère MAP, en supposant une loi *a priori* sur les paramètres du modèle. L'utilisation d'*a priori* permet alors à l'algorithme EM de converger plus facilement vers des solutions pertinentes, mais il est pour cela nécessaire d'avoir des *a priori* qui soient adaptés au problème. Dans le cadre de sa thèse consistant à séparer une source de voix mélangée à de la musique, Ozerov a pris comme *a priori* une loi uniforme pour la source de voix et une distribution de Dirac pour la source de musique, ce qui est équivalent à apprendre les modèles de la source de voix au MV en fixant les paramètres du modèle de musique. Les paramètres du modèle de musique ont été obtenus par apprentissage au MV sur les parties du signal de mélange où uniquement la musique était présente. Les résultats expérimentaux ont montré qu'une telle approche permet d'améliorer les performances en séparation de plusieurs dB en SDR, cependant elle nécessite des *a priori* sur les sources du mélange.

Dans le cadre d'une problématique de séparation aveugle des sources dans laquelle s'inscrit ma thèse, nous ne disposons pas de ces signaux d'apprentissage. Par contre nous avons vu aux sections 1.2 et 1.3 qu'il existe une approche d'apprentissage de modèles des sources à partir du mélange.

5.2 Apprentissage à partir du mélange

Dans cette section nous présentons l'approche d'apprentissage de MMG à partir du mélange tel que proposée par Attias et Bermond [Att99, BMC97]. Le modèle MMG qui est utilisé par Attias et Bermond est appliqué à des variables aléatoires réelles, et par conséquent n'est pas identique au modèle MMG spectral qui a été présenté au chapitre 4, où le modèle est appliqué à des vecteurs aléatoires complexes, cependant la transposition des formules d'un modèle à l'autre ne pose pas de difficulté. Nous développons dans cette section les formules dans le cas d'un MMG scalaire réel, et nous donnons à l'annexe B.3 l'application de ces résultats pour l'apprentissage de MMG spectraux.

5.2.1 Le Modèle MMG scalaire

Nous reprenons les même notations qu'à la section 1.2, où les colonnes $\mathbf{s}(t), 1 \leq t \leq T$ de la matrice des sources $\mathbf{S} = [\mathbf{s}(1), \cdots, \mathbf{s}(T)]$ sont les réalisations iid du vecteur aléatoire $\mathbf{s}$, où les composantes $s_n, 1 \leq n \leq N$ de $\mathbf{s} = [s_1, \cdots, s_N]^T$ sont des variables aléatoires indépendantes qui suivent une distribution de probabilité mélanges de Gaussiennes de moyenne μ_{n,k_n} et de variance σ^2_{n,k_n} :

$$P(s_n; \xi_{s_n}) = \sum_{k_n} \pi_{n,k_n} \mathcal{N}\left(s_n; \mu_{n,k_n}, \sigma^2_{n,k_n}\right) \tag{5.1}$$

avec $\xi_{s_n} = \left\{\pi_{n,k_n}, \mu_{n,k_n}, \sigma^2_{n,k_n}\right\}_{k_n}$, et $\sum_{k_n} \pi_{n,k_n} = 1$ sont les poids des gaussiennes du MMG de la source n. On peut considérer la distribution a priori des sources comme un processus en 2 étapes : La première étape consiste à tirer au hasard un indice j

avec la probabilité $P(k_n = j) = \pi_{nj}$, puis tirer une valeur suivant la loi gaussienne $P(s_n|k_n = j; \xi_{s_n}) = \mathcal{N}\left(s_n; \mu_{n,j}, \sigma^2_{n,j}\right)$. Il est alors assez naturel de considérer que le vecteur $\mathbf{k} = [k_1, k_2, \ldots, k_N]^T$ soit un vecteur aléatoire cachée (non observé).

En utilisant la définition de la probabilité conditionnelle, on a : $P(\mathbf{s}, \mathbf{k}|\mathbf{x}; \xi) = P(\mathbf{s}|\mathbf{x}, \mathbf{k}; \xi)P(\mathbf{k}|\mathbf{x}; \xi)$. Par conséquent, la distribution des sources connaissant le mélange, est un mélange de Gaussiennes dont la densité de probabilité est :

$$P(\mathbf{s}|\mathbf{x}; \xi) = \sum_{\mathbf{k}} P(\mathbf{s}, \mathbf{k}|\mathbf{x}; \xi) = \sum_{\mathbf{k}} \gamma_{\mathbf{k}} \mathcal{N}\left(\mathbf{s}; \boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x}), \mathbf{C}_{\mathbf{k}}\right) \tag{5.2}$$

où :

$$\begin{aligned}
\gamma_{\mathbf{k}} &= P(\mathbf{k}|\mathbf{x}; \xi) && (5.3)\\
\boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x}) &= \boldsymbol{\mu}_{\mathbf{k}} + \mathbf{W}_{\mathbf{k}}\left(\mathbf{x} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{k}}\right) && (5.4)\\
\mathbf{W}_{\mathbf{k}} &= \boldsymbol{\Sigma}_{\mathbf{k}}\mathbf{A}^T\mathbf{B}_{\mathbf{k}} && (5.5)\\
\mathbf{C}_{\mathbf{k}} &= \boldsymbol{\Sigma}_{\mathbf{k}} - \boldsymbol{\Sigma}_{\mathbf{k}}\mathbf{A}^T\mathbf{B}_{\mathbf{k}}\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{k}} && (5.6)\\
\mathbf{B}_{\mathbf{k}} &= \left(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{k}}\mathbf{A}^T + \mathbf{R}_b\right)^{-1} && (5.7)\\
\boldsymbol{\mu}_{\mathbf{k}} &= [\mu_{1,k_1}, \mu_{2,k_2}, \ldots, \mu_{N,k_N}]^T && (5.8)\\
\boldsymbol{\Sigma}_{\mathbf{k}} &= \mathrm{diag}\left(\left[\sigma^2_{1,k_1}, \sigma^2_{2,k_2}, \ldots, \sigma^2_{N,k_N}\right]\right) && (5.9)
\end{aligned}$$

Comme dans le cas d'une distribution gaussienne des sources (voir les équations (A.26),(A.27) et (A.29),(A.30)), il y a deux façons d'exprimer les equations (5.5) et (5.6). Nous ne donnons ici que la forme correspondant aux équations (A.26) et(A.27). Il reste à exprimer la distribution à posteriori de $\mathbf{k}$, qui s'obtient en utilisant la loi de Bayes :

$$\begin{aligned}
\gamma_{\mathbf{k}} = P(\mathbf{k}|\mathbf{x}; \xi) &= \frac{P(\mathbf{x}|\mathbf{k}; \xi)\pi_{\mathbf{k}}}{P(\mathbf{x}; \xi)} \\
&\propto \pi_{\mathbf{k}}\mathcal{N}\left(\mathbf{x}; \mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{k}}\mathbf{A}^T + \mathbf{R}_b\right)
\end{aligned} \tag{5.10}$$

5.2.2 Algorithme EM pour des sources mélanges de Gaussiennes

La dérivée de la fonctionnelle EM est donnée par l'équation (1.10), où les variables cachées sont les sources, mais aussi les variables $\mathbf{k}$ décrivant quelle Gaussienne du mélange de Gaussiennes a été "sélectionnée" $\mathbf{Z} = (\mathbf{S}, \mathbf{K})$. La distribution conjointe des variables $(\mathbf{X}, \mathbf{S}, \mathbf{K})$ se factorise simplement :

$$\log P(\mathbf{X}, \mathbf{S}, \mathbf{K}; \xi) = \log P(\mathbf{X}|\mathbf{S}, \mathbf{K}; \xi_x) + \log P(\mathbf{S}|\mathbf{K}; \xi_s) + \log P(\mathbf{K}; \xi_k) \tag{5.11}$$

Ainsi l'optimisation de la fonctionnelle EM (donnée par l'équation (1.8)) se sépare en trois termes que l'on peut optimiser séparément car ils dépendent de paramètres différents. Le premier terme de (5.11) permet l'estimation de la matrice de mélange et de la covariance du bruit. Les formules de ré-estimation, qui ont été établies par les équations (A.15) et (A.16), et ne dépendent pas du modèle des sources, du moment

que l'on connait les espérances conditionnelles $\mathbb{E}\{\mathbf{s}|\mathbf{x};\xi\}$ et $\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\}$. Ces dernières s'obtiennent à partir des formules (5.2), (5.10) :

$$\begin{aligned}
\mathbb{E}\{\mathbf{s}|\mathbf{x};\xi\} &= \sum_{\mathbf{k}} P(\mathbf{k}|\mathbf{x};\xi)\mathbb{E}\{\mathbf{s}|\mathbf{x},\mathbf{k};\xi\} \\
&= \sum_{\mathbf{k}} \gamma_{\mathbf{k}}\mathbb{E}\{\mathbf{s}|\mathbf{x},\mathbf{k};\xi\} \qquad (5.12) \\
\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\} &= \sum_{\mathbf{k}} P(\mathbf{k}|\mathbf{x};\xi)\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x},\mathbf{k};\xi\} \\
&= \sum_{\mathbf{k}} \gamma_{\mathbf{k}}\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x},\mathbf{k};\xi\} \qquad (5.13) \\
\mathbb{E}\{\mathbf{s}|\mathbf{x},\mathbf{k};\xi\} &= \boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x}) \qquad (5.14) \\
\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x},\mathbf{k};\xi\} &= \boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x})\boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x})^T + \mathbf{C}_{\mathbf{k}} \qquad (5.15)
\end{aligned}$$

Si l'on ne connait pas la matrice de mélange, alors on peut toujours utiliser la formule (A.21) pour estimer la covariance du bruit.

Pour estimer les paramètres des sources, on cherche à maximiser la fonctionnelle EM du second terme de l'équation (5.11) [BMC97, Att99] :

$$\mathbb{E}\{\log P(\mathbf{S}|\mathbf{K};\xi_s)|\mathbf{X},\xi'\} = \sum_t \sum_n \sum_{k_n} \int \log P(s_n|k_n;\xi_{s_n}) P(s_n,k_n|\mathbf{x}(t);\xi') d_{s_n} \qquad (5.16)$$

En annulant la dérivée par rapport aux paramètres $\xi_s = \left\{\mu_{n,k_n}, \sigma^2_{n,k_n}\right\}$, on obtient les formules de ré-estimation [Att99] :

$$\begin{aligned}
\mu^{(l+1)}_{n,k_n} &= \frac{\sum_t \sum_{\{k_j\}j\neq n} \gamma^{(l)}_{\mathbf{k}}(t) \cdot \mathbb{E}\left\{ s_n \left| \mathbf{k}, \mathbf{x}(t); \xi^{(l)} \right.\right\}}{\sum_t \sum_{\{k_j\}j\neq n} \gamma^{(l)}_{\mathbf{k}}(t)} \qquad (5.17) \\
\sigma^{2,(l+1)}_{n,k_n} &= \frac{\sum_t \sum_{\{k_j\}j\neq n} \gamma^{(l)}_{\mathbf{k}}(t) \cdot \mathbb{E}\left\{ s^2_n \left| \mathbf{k}, \mathbf{x}(t); \xi^{(l)} \right.\right\}}{\sum_t \sum_{\{k_j\}j\neq n} \gamma^{(l)}_{\mathbf{k}}(t)} - \mu^{(l+1),2}_{n,k_n} \qquad (5.18)
\end{aligned}$$

où $\gamma^{(l)}_{\mathbf{k}}(t) = P\left(\mathbf{k}\left|\mathbf{x}(t);\xi^{(l)}\right.\right)$

Pour estimer les poids $\xi_k = \{\pi_{n,k_n}\}$ des Gaussiennes, on cherche à maximiser la fonctionnelle EM du troisième terme $P(\mathbf{K};\xi_k)$ de la probabilité complète :

$$\mathbb{E}\{\log P(\mathbf{K};\xi_k)|\mathbf{X},\xi'\} = \sum_t \sum_n \sum_{k_n} \log P(k_n;\xi_{k_n}) P(k_n|\mathbf{x}(t);\xi') \qquad (5.19)$$

En annulant la dérivée par rapport au paramètre $\xi_k = \{\pi_{n,k_n}\}$, on obtient la formule de ré-estimation [Att99] :

$$\pi^{(l+1)}_{n,k_n} = \frac{1}{T} \sum_t \sum_{\{k_j\}j\neq n} \gamma^{(l)}_{\mathbf{k}}(t) \qquad (5.20)$$

Pour résumer, l'algorithme EM pour les mélange de Gaussiennes s'effectue en 2 étapes :

- étape E :
  calcul des statistiques suffisantes $\mathbb{E}\{\mathbf{s}|\mathbf{x},\mathbf{k};\xi^{(l)}\}$ et $\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x},\mathbf{k};\xi^{(l)}\}$ (équations (5.12) et (5.13)) ainsi que les probabilités des états $\gamma_{\mathbf{k}}^{(l)}(t)$ (voir équation (5.10)) ;
- étape M :
  ré-estimation des paramètres $\mathbf{A}^{(l+1)}, \mathbf{R}_b^{(l+1)}$ (avec les équations (1.11) et (1.12)) et des paramètres $\left\{\pi_{n,k_n}, \mu_{n,k_n}, \sigma^2_{n,k_n}\right\}^{(l+1)}$ à l'aide des grandeurs calculées à l'étape E (équations (5.20), (5.17), (5.18)).

5.2.3 Estimation des sources

L'estimation des sources par la MEQM est donnée par l'espérance conditionnelle des sources connaissant les observations :

$$\widehat{\mathbf{s}} = \mathbb{E}\{\mathbf{s}|\mathbf{x};\xi\}$$

Dans le cas du MMG, la probabilité des sources connaissant les observations est un mélange de Gaussiennes (voir équation (5.2)). Par conséquent, l'estimateur MEQM du MMG est :

$$\widehat{\mathbf{s}} = \sum_{\mathbf{k}} \gamma_{\mathbf{k}} \boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x}) \tag{5.21}$$

où $\gamma_{\mathbf{k}}$ est défini par l'équation (5.3), et $\boldsymbol{\rho}_{\mathbf{k}}(\mathbf{x})$ est défini par l'équation (5.4).

5.2.4 Applications

L'algorithme EM pour des MMG a été utilisé par Attias [Att99], Bermond [BMC97], Moulines [MCG97] pour estimer les sources conjointement à la matrice de mélange, dans le cas sur-déterminé. Dans le cas sous-déterminé, Bermond [BC99a, Ber00] a comparé la méthode EM avec des méthodes basées sur les cumulants du troisième et du quatrième ordre [Com98, Car91]. Ces dernières, plus simples que la méthode EM, ne nécessitent aucune information a priori sur la distribution des sources, mais bien que fonctionnant dans le cas sous-déterminé, possèdent une borne sur le nombre maximum de sources identifiables. Les expériences de Bermond sur des signaux synthétiques, ont montré que la méthode EM donne de meilleurs résultats pour l'estimation de la matrice de mélange, que pour les méthodes basées sur les cumulants.

Des expériences ont été menées par Snoussi [SPMP+01] et Cardoso [CSDP02] dans le domaine spectral, pour faire de la séparation d'images astronomiques.

Dans le domaine audio, Davis [DM04], utilise un modèle MMG dans le cas bruité sous-déterminé. Son modèle est très simple car il n'y a que 2 états (2 Gaussiennes) par source, un état dit "off" représentant l'inactivité de la source, avec une variance fixée à zéro, et un état "on" représentant l'activité de la source avec une variance fixée à 1. Dans les deux cas la moyenne est nulle. Les expériences sont menées sur des signaux audio, dans le cas sous-déterminé (2 canaux, 3 sources), et les résultats sont comparés avec

la méthode de Lewicki [LS00b]. Les experiences montrent que la méthode de Lewicki obtient de meilleurs résultats en séparation que le MMG dans le cas non bruité, mais dans le cas bruité, le MMG obtient les meilleurs résultats. Ces résultats s'expliquent surtout par le fait que le modèle de Lewicki, contrairement au MMG ne prend pas en compte le bruit. D'autres part la modélisation des sources par des MMG ayant seulement deux états est comme le mentionne M. Davis assez « grossière ».

5.2.5 Limites de l'approche

L'algorithme EM pour des MMG permet d'estimer conjointement la matrice de mélange et les sources y compris dans le cas sous déterminé, ainsi que les paramètres *de nuisance* (covariance du bruit et des sources). L'algorithme EM est robuste au bruit additif gaussien, et le nombre de sources identifiables n'est théoriquement pas limité [BC99a]. Cependant sa mise en oeuvre peut poser quelques problèmes.

Complexité algorithmique : Une des limites principales de l'approche EM pour les MMG, est que le nombre de Gaussiennes (c.a.d. le nombre d'états) de la distribution du mélange croît exponentiellement en fonction du nombre de sources. Si il y a N sources et que chaque source est modélisée par $\mathcal{K}$ Gaussiennes, alors le nombre de Gaussiennes du mélange est $\mathcal{K}^N$. Ainsi la complexité calculatoire de l'algorithme EM peut vite devenir rédhibitoire.

Afin de réduire la complexité, Attias [Att99] propose de remplacer la distribution des variables cachées conditionnellement aux observations, par une forme plus simple, factorisable (mais dépendant d'un paramètre supplémentaire), dite approximation variationnelle [JGJS99].

Olshausen [OM99], propose lui de ne considérer uniquement les états les plus probables, et d'estimer la distribution de ces états conditionnellement aux observations par un échantillonneur de Gibbs.

Davis [DM04] propose lui aussi de réduire le nombre d'états en faisiant des hypothèses simplificatrices sur la distribution a priori des sources et en faisant une hypothèse de parcimonie pour supprimer des combinaisons d'états de la distribution du mélange observé. Ainsi, il propose d'une part de ne retenir que 2 états par sources (l'un « actif », l'autre « inactif »), et d'autre part d'interdire que 2 sources soient actives en même temps. Ces deux contraintes permettent d'écrire une approximation de la vraisemblance qui possède une complexité linéaire en fonction du nombre de sources.

Initialisation et convergence : Comme nous l'avons déjà fait remarqué à la section 1.3, un problème important de l'approche EM est sa convergence, qui non seulement peut être très lente, mais en plus dépend énormément de son initialisation. Ainsi l'optimum dans lequel l'algorithme va converger dépend fortement de son initialisation. Par conséquent, si l'on part d'une initialisation aléatoire comme cela est fait dans l'état de l'art, il y a de grandes chances que l'algorithme converge dans un optimum local. La quantité d'optimums locaux, ainsi que la vitesse de convergence dépend de la proportion de données manquantes [JK97]. Dans notre modélisation, il s'agit des sources, ainsi que

des états du MMG du mélange. Ces derniers croissent exponentiellement en fonction du nombre de sources. Pour ces raisons ces modèles sont souvent utilisés pour un nombre très faible d'états (2 ou 3) par sources [Car07, Att99, DM04].

L'algorithme EM a un comportement critique pour l'estimation de la matrice de mélange dans les cas limites où le bruit tend vers zéro. Nous avons vu à la section 1.3.3 que dans le cas d'un bruit nul dans la modélisation, la matrice de mélange n'évolue pas, et pour un bruit faible, la convergence de la matrice de mélange est très lente [BC99b] Cependant ce problème disparait si l'on a estimé la matrice de mélange auparavant, par exemple à l'aide d'une méthode basée sur la parcimonie, et que l'on fixe les paramètres correspondant dans la procédure EM.

5.3 Conclusion

Dans ce chapitre, nous avons présenté deux approches de l'état de l'art pour l'apprentissage des paramètres des MMG des sources. La première approche consiste à faire un apprentissage hors-ligne pour chacune des sources. Cette approche est non-aveugle car elle suppose que l'on ait pour chacune des sources du mélange un ensemble d'apprentissage ayant des propriétés statistiques similaires aux sources du mélange.

Une autre approche consiste à apprendre les paramètres des MMG des sources, directement à partir du mélange. Cette approche aveugle comporte malheureusement deux incovénients majeurs : Premièrement la complexité calculatoire exponentielle de l'algorithme ne permet de traiter des modèles MMG n'ayant que quelques états. Deuxièmement il se pose le problème de l'initialisation des paramètres du modèle, qui si elle est aléatoire a peu de chances de converger vers une solution satisfaisante (proche de l'optimum global).

L'une des contributions principales de cette thèse qui sera développée au chapitre 8 est de proposer une nouvelle approche d'apprentissage aveugle de MMG spectraux, qui n'a pas les problèmes de convergence et de complexité calculatoire de l'approche que nous avons décrit dans ce chapitre.

Deuxième partie

Contributions

Chapitre 6

Estimation robuste des paramètres du mélange

L'estimation des paramètres du mélange est une tâche majeure de la séparation de sources. Elle est la première étape de l'architecture en 2 étapes des approches de l'analyse en composantes parcimonieuses (voir section 1.5), et elle est l'unique tâche non-triviale à accomplir dans le cas des mélanges instantanés (sur-)déterminés. En effet, dans ce dernier cas, l'étape d'estimation des sources s'effectue par simple (pseudo-)inversion de la matrice de mélange.

Nous avons vu au chapitre 2 que la principale hypothèse utilisée par l'état de l'art pour estimer les paramètres du mélange était la parcimonie de la représentation. Grâce à la parcimonie des sources dans le domaine temps-fréquence, il est possible d'extraire des informations locales sur les directions du mélange. Ces informations locales peuvent être combinées par une méthode de clustering afin d'estimer les directions du mélange.

L'hypothèse de parcimonie utilisée par les méthodes comme DUET est très forte car elle suppose que le support des sources est quasi-disjoint dans le plan temps-fréquence. Dans la réalité les sources se recouvrent assez souvent, et ceci d'autant plus que le nombre de sources est grand. La méthode TIFROM [AD03] se base sur une hypothèse de parcimonie beaucoup plus souple que celle de DUET, puisqu'il est supposé que pour chaque source, il y ait au moins une région temps-fréquence où cette source est la seule à être active. Dans ces régions, la différence d'intensité entre canaux fournit alors une information suffisante pour déterminer, dans le cas des mélanges instantanés, l'une des directions du mélange.

Limites des approches existantes : Une des limites principales des méthodes de l'état de l'art que nous avons présentées au chapitre 2 est leur manque de robustesse qui est lié au fait que des paramètres qui dépendent fortement des données à traiter sont des valeurs fixes, non-adaptatives. Le réglage de ces paramètres doit en pratique être fait « à la main » à partir de connaissances *a priori* sur le nombre de sources et la distance minimale entre les directions du mélanges.

Une autre limitation des méthodes de l'état de l'art est qu'elles se basent sur des

estimateurs non-invariants par rotation, et la plupart du temps asymétriques. Autrement dit, ces estimateurs n'accordent pas la même sensibilité aux différentes directions du mélange.

Nous avons vu au chapitre 2 que l'une des principales difficultés de l'estimation des directions d'un mélange anéchoïque est l'estimation des grands délais. En effet, peu de méthodes permettent d'estimer de grands délais. En particulier, l'estimation des délais par les méthode de type DUET [YR04] n'est valide uniquement pour des délais qui ne dépassent pas 1 échantillon du signal. Nous avons présenté au chapitre 2 quelques méthodes [Bof03, PD06, PD05] permettant de dépasser cette limitation.

Hypothèse de travail : L'approche que nous proposons pour estimer les directions d'un mélange instantané repose sur les mêmes hypothèses que TIFROM. Dans le cas anéchoïque, si l'on suppose que les délais peuvent être supérieurs à 1 échantillon, l'hypothèse doit être un peu plus forte que dans le cas instantané. Pour chaque source il doit y avoir plusieurs régions à différentes fréquences où cette source est l'unique source active. Le paramètre d'intensité et la différence de phase fournissent alors une information locale relative à l'une des directions du mélange. Ces estimations locales doivent être combinées avec d'autres estimations locales de la même direction afin de pouvoir estimer cette dernière.

Aperçu de l'approche : L'approche présentée dans ce chapitre repose sur la proposition et la définition d'un nouvel algorithme de *clustering* appelé DEMIX (Direction Estimation of Mixing matrIX), qui estime à la fois le nombre et les directions des sources. Cet algorithme de clustering combine les informations des différences de phase et des paramètres d'intensité en donnant plus de poids aux régions temps-fréquence où une source domine significativement les autres. De telles régions sont *fiables* dans la mesure où les valeurs des différences de phase et des paramètres d'intensité calculées dans ces régions sont des estimées des directions du mélange ayant une faible variance d'erreur.

Une des contributions importantes de cette thèse est une mesure dite de *fiabilité*, dont la valeur est associée à une région temps-fréquence, et qui caractérise la dominance d'une des sources par rapport aux autres. Cette mesure de fiabilité a un rôle similaire, (à une inversion près) à la variance $\mathrm{Var}\left(R_{21}\left(\Omega_{t,f}\right)\right)$ de TIFROM définie par l'équation (2.13), et à la valeur absolue du coefficient de corrélation inter-canal $\left|\widehat{C}_{X_1X_2}\left(\Omega_{t,f}\right)\right|$ de la méthode TIFCORR définie par l'équation (2.15). Cependant, la mesure de fiabilité contrairement aux mesures de TIFROM et TIFCORR, est invariante par rotation. L'utilisation de la fiabilité dans l'étape de clustering, permet de combiner les estimations locales des différentes régions temps-fréquence, en donnant plus de poids aux points qui ont une grande fiabilité.

Résultats : Les expériences montrent que l'algorithme de clustering DEMIX que nous présentons dans ce chapitre est alors plus robuste que les méthodes de l'état de

l'art que nous avons évaluées [1], car il permet d'estimer plus précisément les directions des sources, y compris dans des situations difficiles. C'est à dire quand :

- les sources sont nombreuses relativement au nombre de capteurs ;
- les sources sont proches les unes des autres ;
- les délais entre capteurs sont grands.

La **première étape** de notre approche est une étape d'*extraction des caractéristiques* (*feature extraction* en anglais) qui consiste à définir les estimateurs qui seront utilisés lors l'étape de clustering. Parmi ces estimateurs, on définit :

- l'*estimateur local de direction* qui est semblable aux estimateurs classiques des paramètres d'intensité et de différences de phase ;
- la *fiabilité* associée à l'estimateur local de direction qui permet de facilement discriminer les régions temps-fréquence où une seule source est active, des régions où aucune ou plusieurs sources sont actives, c.-à-d. de discriminer les régions temps-fréquence pour lesquelles l'estimateur local de direction apporte une information fiable sur l'une des directions du mélange, des régions pour lesquelles l'estimateur local de direction est très difficilement exploitable.

La **seconde étape** de notre approche consiste à combiner les estimations locales « extraites » lors de la première étape, à travers notre algorithme de clustering DEMIX qui va regrouper les estimations locales qui décrivent les mêmes directions, afin de les fusionner pour estimer les directions du mélange et leur nombre.

6.1 1re étape : Extraction et sélection des caractéristiques

Notre approche est basée sur une représentation temps-fréquence, obtenue en calculant la TFCT du mélange, et comme pour l'approche TIFROM, sur la notion de régions temps-fréquence $\Omega_{t,f}$, au voisinage de chaque point temps-fréquence (t, f).

Pour chaque région temps-fréquence, on définit deux valeurs :

1. La direction locale $\hat{\mathbf{u}}(\Omega_{t,f})$ de la source dominante ;
2. La mesure de fiabilité locale, notée $\widehat{\mathcal{T}}(\Omega_{t,f})$, qui prend une grande valeur quand les vecteurs $\mathbf{X}(\tau,\omega)/(\tau,\omega) \in \Omega_{t,f}$ du diagramme de dispersion de la région $\Omega_{t,f}$ sont concentrés le long de la direction $\hat{\mathbf{u}}(\Omega_{t,f})$, indiquant qu'une seule source est significativement active.

Le rôle de la mesure de fiabilité $\widehat{\mathcal{T}}(\Omega_{t,f})$ étant de discriminer les situations où essentiellement une seule source est active des situations où aucune ou plusieurs sources sont actives, elle discrimine de façon équivalente les situations où la direction $\hat{\mathbf{u}}(\Omega_{t,f})$ correspond à l'une des directions du mélange, des situations où la direction $\hat{\mathbf{u}}(\Omega_{t,f})$ a très peu de chance de correspondre à l'une des directions du mélange.

Afin d'estimer les directions $\hat{\mathbf{u}}(\Omega_{t,f})$ et les mesures de fiabilité qui leurs sont associées, nous faisons une Analyse en Composantes Principales (ACP) [Jol02, Ben73] à partir des

[1] il serait cependant intéressant d'évaluer, dans le cadre de campagnes d'évaluation, DEMIX avec d'autres méthodes de l'état de l'art comme TIFROM et TIFCORR, en particulier pour l'estimation de grands délais.

vecteurs $\mathbf{X}(\tau, \omega)$ de la région $\Omega_{t,f}$, et définissons le vecteur $\hat{\mathbf{u}}(\Omega_{t,f})$ comme la direction principale de l'ACP. La mesure de fiabilité $\widehat{\mathcal{T}}(\Omega_{t,f})$ qui sera définie à l'équation (6.3) à partir des valeurs propres de l'ACP est une mesure de l'alignement des points $\mathbf{X}(\tau, \omega)$ le long de la direction $\hat{\mathbf{u}}(\Omega_{t,f})$.

6.1.1 Les régions temps-fréquence

Nous considérons deux types de régions temps-fréquences, qui sont définies de la même façon que dans la méthode TIFROM [PD06] que nous avons présenté à la section 2.1.4. Nous distinguons les régions dites temporelles et les régions dites fréquentielles. Chaque point temps-fréquence (t, f), possède alors deux voisinages : son voisinage temporel $\Omega^T_{t,f}$, composé des points temps-fréquence qui sont à la même fréquence f que le point (t, f), mais sur des trames τ voisines de t, et son voisinage fréquentiel $\Omega^F_{t,f}$, composé des points temps-fréquence qui sont sur la même trame t que le point (t, f), mais sur des fréquences ω voisines de f.

Les points $\mathbf{X}(t, f)$ sont calculés sur une grille temps-fréquence discrète d'indices : $t = k\,\mathrm{L}/2, k \in \mathbb{Z}$ et $f = l/\mathrm{L}, 0 \leq l \leq \mathrm{L}/2$. Le voisinage temporel (respectivement fréquentiel) d'un point temps-fréquence (t, f), est défini par :

$$\Omega^T_{t,f} = \{(t + k\,\mathrm{L}/2, f) \quad | \quad |k| \leq \mathcal{L}\} \tag{6.1}$$

$$\Omega^F_{t,f} = \{(t, f + k/\mathrm{L}) \quad | \quad |k| \leq \mathcal{L}\}. \tag{6.2}$$

Approche multi-échelle : Afin d'exploiter les caractéristiques temporelles et fréquentielles des représentations de la TFCT à différentes échelles (pour différentes tailles de fenêtre), on peut considérer les régions temps-fréquence calculées à différentes échelles de la TFCT. Les régions temps-fréquence à des échelles différentes auront alors des formes différentes dans le plan temps-fréquence, ce qui peut être intéressant pour s'adapter automatiquement aux formes des composantes des signaux sources.

6.1.2 Diagramme de dispersion local

Nous avons vu que le diagramme de dispersion est un outil de visualisation intéressant, car il permet de se rendre compte de la structure des points temps-fréquence, et en particulier de voir si des points sont alignés dans des directions du mélange. Si les points du diagramme de dispersion sont disposés de façon non structurée (comme sur la figure 4(c)), alors il y a peu de chance qu'un quelconque algorithme de clustering puisse trouver les directions du mélange. A contrario, si les points sont tous alignés sur des directions du mélange (comme sur le figure 2.2(b)), alors il sera très facile d'estimer les directions du mélange par un algorithme de clustering.

Plutôt que d'observer le diagramme de dispersion global (sur l'ensemble des points temps-fréquence, voir un exemple figure 2.1(b)), nous pouvons visualiser les points temps-fréquence qui appartiennent à un même voisinage. Nous appelons cet outil de visualisation qui a été introduit à la section 2.1.4, et dont deux exemples sont présentés à la figure 6.1, le diagramme de dispersion local.

Les régions temps-fréquence qui ont été définies à la section 6.1.1 contiennent $|\Omega| = 2\mathcal{L}+1$ points complexes $\mathbf{X}(\tau,\omega), (\tau,\omega) \in \Omega$. Nous construisons la matrice $\mathbf{X}(\Omega)$ de taille $M \times |\Omega|$, dont les colonnes sont les points $\mathbf{X}(\tau,\omega)$. On parlera alors du diagramme de dispersion de $\mathbf{X}(\Omega)$ pour désigner le diagramme de dispersion complexe des points $\mathbf{X}(\tau,\omega)$. Visualiser des vecteurs complexes n'est pas aisé, cependant dans le cas des mélanges instantanés, puisque les directions des sources $\mathbf{a}_n(f)$ sont à valeur réelle, nous pouvont utiliser un diagramme de dispersion (réel) de $\mathbf{X}^{\mathbb{R}}(\Omega)$. La matrice $\mathbf{X}^{\mathbb{R}}(\Omega)$, de taille $M \times 2|\Omega|$ est alors constituée des vecteurs colonnes $\Re\mathrm{e}\mathbf{X}(\tau,\omega)$ et $\Im\mathrm{m}\mathbf{X}(\tau,\omega), (\tau,\omega) \in \Omega$.

6.1.3 Analyse en Composantes Principales et mesure de fiabilité

Si l'on suppose qu'une seule source est active dans un voisinage Ω, alors les points du diagramme de dispersion local sont alignés le long de la direction du mélange correspondant à cette unique source active. Ceci est illustré par la figure 6.1(b). Dans le cas contraire, si aucune source n'est active, ou si plusieurs sources sont actives en même temps, alors les points du diagramme de dispersion local sont dispersés dans l'espace défini par les sources actives, qui par définition n'est pas de dimension 1. Ces points n'ont donc aucune raison d'être alignés dans une direction particulière. Ce cas est illustré par la figure 6.1(a).

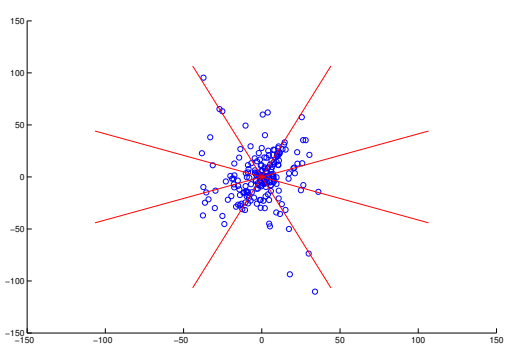
Par conséquent, si une seule source est active dans un voisinage Ω, alors l'analyse en composantes principales (ACP) calculée sur $\mathbf{X}(\Omega)$ (repectivement sur $\mathbf{X}^{\mathbb{R}}(\Omega)$) fournit la direction principale $\hat{\mathbf{u}}(\Omega) \in \mathbb{C}^M$ (resp. $\hat{\mathbf{u}}(\Omega) \in \mathbb{R}^M$) correspondant à l'unique source active. De plus, l'ACP fournit les valeurs propres $\hat{\lambda}_1(\Omega) \geq \ldots \geq \hat{\lambda}_M(\Omega) \geq 0$ de la matrice de covariance $\widehat{\mathbf{R}}_x(\Omega) \stackrel{\triangle}{=} \frac{1}{|\Omega|}\mathbf{X}(\Omega)\mathbf{X}^H(\Omega)$ (respectivement $\widehat{\mathbf{R}}_x^{\mathbb{R}}(\Omega) \stackrel{\triangle}{=} \frac{1}{|\Omega|}\mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$). La matrice $\widehat{\mathbf{R}}_x(\Omega)$ (resp. $\widehat{\mathbf{R}}_x^{\mathbb{R}}(\Omega)$) est une matrice complexe hermitienne définie nonnégative (DNN) (resp. réelle symétrique DNN) de taille $M \times M$.

Nous proposons de définir la mesure (empirique) de fiabilité par :

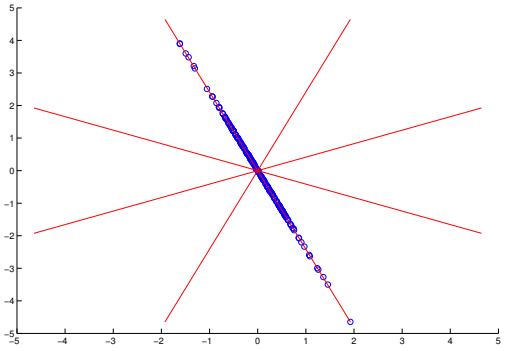
$$\widehat{\mathcal{T}}(\Omega) \stackrel{\triangle}{=} \widehat{\lambda}_1(\Omega) \Big/ \frac{1}{M-1}\sum_{m=2}^{M} \widehat{\lambda}_m(\Omega) \; . \tag{6.3}$$

Nous verrons à la section 6.2 que cette mesure de fiabilité peut être interprétée comme un rapport signal à bruit (RSB) entre la source dominante et la contribution des autres sources plus éventuellement un bruit (si l'on considère un modèle de mélange avec un bruit additif). Il est alors assez commode d'exprimer la fiabilité en décibel (dB) : $10\log_{10}(\widehat{\mathcal{T}}(\Omega))$.

La figure 6.1 (qui a déjà été présentée à la figure 2.4 mais sans les valeurs de fiabilité associées) montre le diagramme de dispersion local de $\mathbf{X}^{\mathbb{R}}(\Omega)$ dans deux régions temps-fréquence différentes : Une où plusieurs sources sont actives simultanément, et une autre où il y a essentiellement une seule source active. On peut remarquer que la mesure de fiabilité est élevée dans le second cas de figure où essentiellement une source est active, et faible dans le cas opposé.



(a) Région temps-fréquence où plusieurs sources sont active. La mesure de fiabilité est faible (4.7 dB)



(b) Région temps-fréquence où une essentiellement une seule source est active. La mesure de fiabilité est élevée (50.7 dB)

FIG. 6.1 – Diagrammes de dispersion local de deux régions temps-fréquence. Les droites indiquent la vraie position des sources du mélange. La fenêtre de la TFCT est de taille $\mathrm{L} = 4096$, et la taille du voisinage est de $|\Omega| = 99$.

6.1.4 Comparaison de l'approche locale avec l'approche globale

Afin de comparer l'approche locale utilisant l'ACP qui est décrite dans ce chapitre avec l'approche globale de l'état de l'art, nous proposons une extension de la définition de diagramme de dispersion.

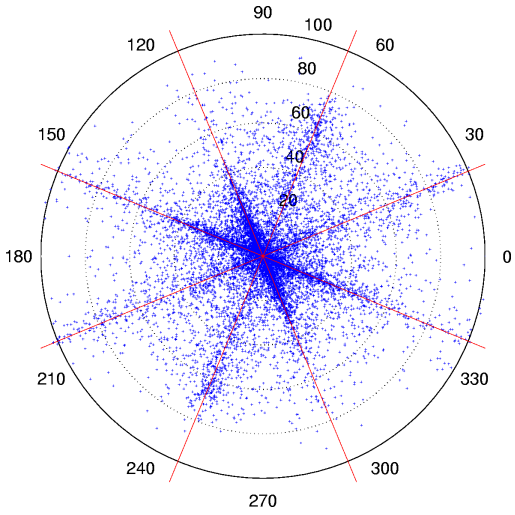
Diagramme de dispersion pondéré : Le diagramme de dispersion d'un ensemble de points réels $\mathbf{x}(t) = [x_1(t), x_2(t)]^T$ a été défini comme la représentation graphique des points $(x_1(t), x_2(t))$. l'équation de ces points peut se ré-écrire en coordonnées polaires :

$$(x_1(t), x_2(t)) = (\rho(t)\cos\theta(t), \rho(t)\sin\theta(t)),$$

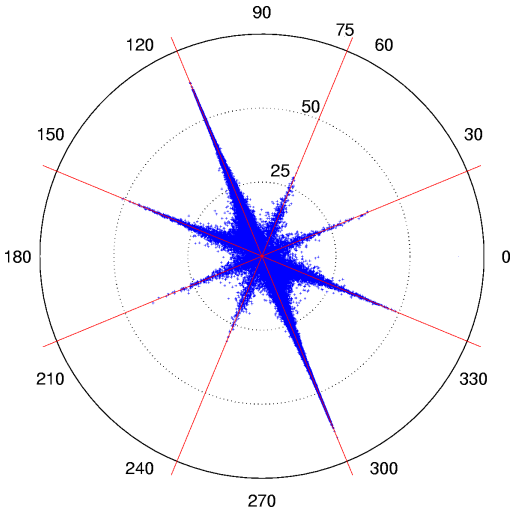
où $\rho(t) \triangleq \sqrt{x_1^2(t) + x_2^2(t)}$, et $\theta(t) \triangleq \tan^{-1}(x_2(t)/x_1(t))$.

On définit le *diagramme de dispersion pondéré* de $\mathbf{x}(t)$, par la représentation graphique des points $(\rho'(t)\cos\theta(t), \rho'(t)\sin\theta(t))$, $\theta(t) = \tan^{-1}(x_2(t)/x_1(t))$, mais où où $\rho'(t)$ est la valeur de pondération qui peut être différente de $\sqrt{x_1^2(t) + x_2^2(t)}$.

Sur la figure 6.2(a), on peut observer les diagrammes de dispersion des points temps-fréquence $\mathbf{X}^{\mathbb{R}}(t, f)$ pondérés par leur énergie $\rho^2(t, f) = \|\mathbf{X}^{\mathbb{R}}(t, f)\|^2$, qui sont utilisés par l'approche standard. Sur la figure 6.2(b), on peut observer les points $\hat{\mathbf{u}}(\Omega_{t,f})$ obtenus par ACP sur l'ensemble des régions temps-fréquence, pondérés par la valeur en dB de leur fiabilité $10\log_{10}\widehat{\mathcal{T}}(\Omega_{t,f})$. On peut observer que les points de la figure 6.2(b) sont plus concentrés le long des directions du mélange que ceux de la figure 6.2(a). Par conséquent il parait évident que les points obtenus par ACP sont de meilleurs candidats que ceux de l'approche globale, à l'estimation des directions via un algorithme de clustering. Ceci sera confirmé par des expériences menées à la section 7.



(a) Diagramme de dispersion des points $\mathbf{X}^{\mathbb{R}}(t,f)$ pondérés par leur énergie $\rho^2(t,f) = \|\mathbf{X}^{\mathbb{R}}(t,f)\|^2$.



(b) Diagramme de dispersion des points $\hat{\mathbf{u}}(\Omega_{t,f})$ pondérés par leur mesure de fiabilité $10\log_{10}\widehat{\mathcal{T}}(\Omega_{t,f})$. La taille des voisinage est : $|\Omega_{t,f}| = 9$.

FIG. 6.2 – Comparaison des diagrammes de dispersion pondérés des points utilisés par l'approche globale standard et des points utilisés par l'approche locale utilisant l'ACP. Les segments de droite indiquent les positions des vraies directions qui sont au nombre de 4. La taille des fenêtres de la TFCT est L = 4096.

6.2 Le modèle de mélange local gaussien de DEMIX

Dans cette section, nous proposons un modèle du mélange dans une région temps-fréquence afin de pouvoir interpréter les estimateurs de direction et de fiabilité empirique que nous avons introduit à la section précédente. Les résultats obtenus dans cette section sont des résultats asymptotiques sur la qualité de ces estimateurs qui servent à l'étape de *clustering* de la section suivante.

Dans le cas d'un mélange instantané ($\delta_n = 0$), la matrice de mélange $\mathbf{A}(f)$ est une matrice à valeurs réelles qui ne dépend pas de la fréquence ($\mathbf{A}(f) = \mathbf{A}$). Le mélange des sources (bruité) qui est à valeur complexe s'écrit : $\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f) + \mathbf{B}(t,f)$, mais il est possible de le décomposer de façon équivalente en un mélange à valeurs réelles en prenant les parties réelles et imaginaires du mélange : $[\Re\mathbf{X}(t,f), \Im\mathbf{X}(t,f)] = \mathbf{A}[\Re\mathbf{S}(t,f), \Im\mathbf{S}(t,f)] + [\Re\mathbf{B}(t,f), \Im\mathbf{B}(t,f)]$.

Dans notre modèle de mélange local, nous supposons que la source S_n, ayant pour direction $\mathbf{a}_n$, est la source la plus active dans la région Ω. Les valeurs des parties réelles et imaginaires de la TFCT de cette source dans la région temps-fréquence Ω sont modélisées par des variables aléatoires indépendantes, gaussiennes de moyenne nulle et de variance σ_n^2. Autrement dit, la valeur de la TFCT de la source est modélisée par une variable aléatoire gaussienne complexe circulaire centrée et de variance $2\sigma_n^2$. Les contributions des autres sources, incluant éventuellement le bruit $\mathbf{B}$, sont modélisées par une

distribution gaussienne centrée de dimension M ayant comme matrice de covariance : $\sigma_b^2 \mathbf{I}_M$.

Par conséquent, les entrées $\Re\mathbf{X}(\tau,\omega), \Im\mathbf{X}(\tau,\omega)$, $(\tau,\omega) \in \Omega$ du diagramme de dispersion de $\mathbf{X}^{\mathbb{R}}(\Omega) = \mathbf{a}_n \cdot S_n^{\mathbb{R}}(\Omega) + \mathbf{B}^{\mathbb{R}}(\Omega)$ suivent une distribution gaussienne isotrope centrée $\mathcal{N}\left(0, \mathbf{R}_x^{\mathbb{R}}\right)$ avec :

$$\mathbf{R}_x^{\mathbb{R}} = \sigma_n^2 \mathbf{a}_n \mathbf{a}_n^T + \sigma_b^2 \mathbf{I}_M. \tag{6.4}$$

La matrice de covariance $\mathbf{R}_x^{\mathbb{R}}$ est une matrice à valeurs réelles symétrique. La plus grande valeur propre de $\mathbf{R}_x^{\mathbb{R}}$ est $\lambda_1 = \sigma_n^2 + \sigma_b^2$ et est associée à la direction principale $\mathbf{a}_n$. Les valeurs propres restantes sont : $\lambda_2 = \ldots = \lambda_M = \sigma_b^2$. Selon cette modélisation, la direction principale calculée à partir de la matrice de covariance $\mathbf{R}_x^{\mathbb{R}}$ est la « vraie » direction $\mathbf{a}_n$. Nous appelons cette direction, la « vraie » direction dans la mesure où elle est calculée à partir de la matrice de covariance $\mathbf{R}_x^{\mathbb{R}}$ du modèle. Dans la pratique nous n'avons pas accès à cette matrice de covariance, mais seulement à une estimation de celle-ci. De la même manière, la « vraie » mesure de fiabilité est définie par :

$$\mathcal{T} \triangleq \frac{\lambda_1}{\frac{1}{M-1}\sum_{m=2}^{M} \lambda_m} = \sigma_n^2/\sigma_b^2 + 1 \tag{6.5}$$

et peut être interprétée comme un rapport signal à bruit (RSB) entre la source dominante, et la contribution des autres sources (et du bruit si l'on considère un mélange bruité).

Si les observations du diagramme de dispersion de $\mathbf{X}^{\mathbb{R}}(\Omega)$ étaient suffisantes pour obtenir une estimation parfaite de la matrice de covariance $\mathbf{R}_x^{\mathbb{R}}$, notre analyse s'arrêterait là. Cependant, dans la pratique la direction principale $\hat{\mathbf{u}}(\Omega)$ et la mesure de fiabilité $\widehat{\mathcal{T}}(\Omega)$ sont calculées par ACP à partir d'un échantillon de seulement $|\Omega|$ points. Par conséquent la direction $\hat{\mathbf{u}}(\Omega)$ et la mesure de fiabilité $\widehat{\mathcal{T}}(\Omega)$ ne sont que des estimations de respectivement la « vraie » direction $\mathbf{a}_n$ et la « vraie » mesure de fiabilité $\mathcal{T}$. Cependant il est possible en se basant sur la théorie des matrices aléatoires de quantifier la précision de ces estimateurs en fonction de la taille $|\Omega|$ de l'échantillon.

6.2.1 Distributions asymptotiques

Dans cette section, nous nous basons sur la théorie des matrices aléatoires afin de développer des résultats asymptotiques sur l'estimateur de direction $\hat{\mathbf{u}}(\Omega)$, ainsi que sur la mesure de fiabilité empirique $\widehat{\mathcal{T}}(\Omega)$.

D'après le théorème [HS03, Theorem 5.7], la matrice de covariance empirique $\widehat{\mathbf{R}}_x^{\mathbb{R}} \triangleq |\Omega|^{-1} \mathbf{X}^{\mathbb{R}}(\Omega)(\mathbf{X}^{\mathbb{R}}(\Omega))^T$ suit une distribution de Wishart $|\Omega|^{-1} \mathcal{W}_M(\mathbf{R}_x^{\mathbb{R}}, |\Omega| - 1)$ de dimension M. D'après le théorème [HS03, Theorem 9.4], si la décomposition spectrale de $\mathbf{R}_x^{\mathbb{R}}$ est $\mathbf{R}_x^{\mathbb{R}} = \mathbf{U}\Lambda\mathbf{U}^T$, et si les valeurs propres Λ sont positives et distinctes deux à deux, alors la décomposition spectrale de $\widehat{\mathbf{R}}_x^{\mathbb{R}} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}^T$ converge en loi, quand la taille de l'échantillon $|\Omega|$ est grande, vers une distribution gaussienne :

$$\sqrt{|\Omega| - 1} \cdot (\hat{\Lambda} - \Lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\Lambda^2) \tag{6.6}$$

$$\sqrt{|\Omega| - 1} \cdot (\hat{\mathbf{u}}_1 - \mathbf{u}_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{V}) \tag{6.7}$$

où $\xrightarrow{\mathcal{L}}$ signifie la convergence en loi. La matrice $M \times M$ de covariance $\mathbf{V}$ est donnée par :

$$\begin{aligned}
\mathbf{V} &= \lambda_1 \sum_{m \geq 2} \frac{\lambda_m}{(\lambda_m - \lambda_1)^2} \mathbf{u}_m \mathbf{u}_m^T \\
&= \left(\frac{\sigma_n^2}{\sigma_b^2} + 1 \right) \cdot \left(\frac{\sigma_b^2}{\sigma_n^2} \right)^2 (\mathbf{I}_M - \mathbf{u}_1 \mathbf{u}_1^T) \\
&= \frac{\mathcal{T}}{(\mathcal{T} - 1)^2} \cdot (\mathbf{I}_M - \mathbf{u}_1 \mathbf{u}_1^T). \qquad (6.8)
\end{aligned}$$

A partir de l'équation (6.8), on peut constater que la « vraie » mesure de fiabilité $\mathcal{T}$ est directement liée à la matrice de covariance $\mathbf{V}$ de l'estimateur $\hat{\mathbf{u}}(\Omega) = \hat{\mathbf{u}}_1$ de la vraie direction $\mathbf{a}_n = \mathbf{u}_1$. Cependant, en pratique, la « vraie » mesure de fiabilité n'est pas observée. Nous disposons uniquement de la mesure de fiabilité empirique $\widehat{\mathcal{T}}$. Par conséquent la relation (6.7) ne peut être utilisée directement.

6.2.2 Fiabilité empirique robuste

Afin de pouvoir évaluer la qualité de l'estimateur de direction en exploitant la relation asymptotique de l'équation (6.7), nous définissons une mesure dite de *fiabilité empirique robuste*, dont la valeur est une borne au dessus de laquelle on espère que se trouve la « vraie » fiabilité.

Définition 6.1 (Fiabilité empirique robuste) *À partir de la fiabilité empirique $\widehat{\mathcal{T}}(\Omega)$, on définit la fiabilité empirique robuste $\tilde{\mathcal{T}}(\Omega)$ de niveau $1 - \alpha$ par :*

$$\tilde{\mathcal{T}}(\Omega) \stackrel{\triangle}{=} \widehat{\mathcal{T}}(\Omega) e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}} \qquad (6.9)$$

où $q(\alpha)$ est le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$.

Le quantile $q(\alpha) = F^{-1}(1-\alpha)$ est défini par la fonction réciproque F^{-1} de la fonction de répartition $F(q) = \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ de la loi $\mathcal{N}(0, 1)$.

D'après le théorème 6.2.1, pour une taille $|\Omega|$ de l'échantillon suffisamment grande, la valeur de la « vraie » fiabilité $\mathcal{T}(\Omega)$ est garantie d'être supérieure à celle de la fiabilité empirique robuste $\tilde{\mathcal{T}}(\Omega)$ avec une probablilité $1 - \alpha$.

Dans nos experiences, nous avons choisi la valeur $1 - \alpha = 99\%$, qui correspond au quantile $q(\alpha) = 2.33$.

Theorème 6.2.1 (Intervalle de confiance de la fiabilité) *Soit $\Lambda = diag([\lambda_1, \cdots, \lambda_M])$ et $\hat{\Lambda} = diag([\hat{\lambda}_1, \cdots, \hat{\lambda}_M])$ deux matrices diagonales aléatoires qui convergent en loi lorsque $|\Omega|$ tend vers $+\infty$ suivant l'équation (6.6), et soit $\mathcal{T}$ défini par l'équation (6.5) et $\widehat{\mathcal{T}}$ défini par $\widehat{\mathcal{T}} \stackrel{\triangle}{=} \frac{\hat{\lambda}_1}{\frac{1}{M-1}\sum_{m=2}^{M} \hat{\lambda}_m}$. Alors si $|\Omega|$ est suffisamment grand :*

$$P\left(\mathcal{T} \geq \widehat{\mathcal{T}} b(\alpha) \right) = 1 - \alpha \qquad (6.10)$$

est l'intervalle de confiance $[\widehat{\mathcal{T}}b(\alpha), +\infty[$ *de niveau* $1-\alpha$ *pour le paramètre* $\mathcal{T}$*, avec :*

$$b(\alpha) = e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}} \tag{6.11}$$

$q(\alpha)$ *étant le quantile d'ordre* $1-\alpha$ *de la loi* $\mathcal{N}(0,1)$*.*

Preuve En utilisant la distribution asymptotique de $\hat{\Lambda}$ donnée par l'équation (6.6), et en posant $\hat{\mu} \triangleq (\hat{\lambda}_1, \frac{1}{M-1}\sum_{m=2}^{M}\hat{\lambda}_m)$ et $\mu \triangleq (\sigma_n^2+\sigma_b^2, \sigma_b^2)$, nous avons :

$$\sqrt{|\Omega|-1}\cdot(\hat{\mu}-\mu) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 2\cdot\mathrm{diag}\left(\mu_1^2, \frac{\mu_2^2}{M-1}\right)\right).$$

En posant $\frac{1}{2}\ln\widehat{\mathcal{T}} = f(\hat{\mu})$ avec $f(x_1,x_2) = \frac{1}{2}\ln x_1 - \frac{1}{2}\ln x_2$, et en utilisant le théorème [HS03, Theorem 4.11], nous avons en posant $\mathbf{d} = (\frac{\partial f}{\partial x_i}\Big|_{\mu})_{i=1,2}$:

$$\sqrt{|\Omega|-1}\left(\frac{1}{2}\ln\widehat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T}\right) \xrightarrow{\mathcal{L}}$$
$$\mathcal{N}\left(0, 2\cdot\mathbf{d}^T\mathrm{diag}\left(\mu_1^2, \mu_2^2/(M-1)\right)\mathbf{d}\right).$$

On peut facilement vérifier que $\mathbf{d}^T\mathrm{diag}\left(\mu_1^2, \mu_2^2/(M-1)\right)\mathbf{d} = \frac{M}{4(M-1)}$, et on obtient alors :

$$\sqrt{\frac{2(M-1)}{M}}\cdot\sqrt{|\Omega|-1}\left(\frac{1}{2}\ln\widehat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Ainsi, pour un *grand échantillon* de taille $|\Omega|$, nous avons :

$$P\left(\mathcal{T} \geq \widehat{\mathcal{T}}e^{-q\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}}\right) =$$
$$P\left(\frac{1}{2}\ln\widehat{\mathcal{T}} - \frac{1}{2}\ln\mathcal{T} \leq q\sqrt{\frac{M}{2(|\Omega|-1)(M-1)}}\right) = F(q).$$

où $F(q)$ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Par conséquent, il y a un quantile $q(\alpha) = F^{-1}(1-\alpha)$ tel que $\mathcal{T} \geq \widehat{\mathcal{T}}e^{-q(\alpha)\sqrt{\frac{2M}{(|\Omega|-1)(M-1)}}}$ avec une probabilité $1-\alpha$.

6.2.3 Précision de l'estimation de direction

Nous pouvons maintenant revenir à la relation entre $\tilde{\mathcal{T}}(\Omega)$ et la matrice de covariance $\mathbf{V}$ de la distribution asymptotique de $\hat{\mathbf{u}}(\Omega)$ autour de la vraie direction $\mathbf{a}_n$.

Proposition 6.2.2 *Si le mélange est gaussien centré avec comme matrice de covariance, la matrice* $\mathbf{R}_x^{\mathbb{R}}$ *définie par l'équation* (6.4)*, et si* $\hat{\mathbf{u}}_1$ *est le vecteur propre correspondant à la plus grande valeur propre de la matrice de covariance empirique* $\widehat{\mathbf{R}}_x^{\mathbb{R}}$ *de* $\mathbf{R}_x^{\mathbb{R}}$*, calculée à partir d'un échantillon de taille* $|\Omega|$*. Alors la distribution de* $\hat{\mathbf{u}}_1$ *converge en loi, quand la taille de l'échantillon* $|\Omega|$ *est grande, vers :*

$$\hat{\mathbf{u}}_1 \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{a}_n, \sigma^2(\mathcal{T}(\Omega)) \cdot \mathbf{R}) \tag{6.12}$$

avec

$$\sigma^2(\mathcal{T}) \triangleq \frac{\mathcal{T}}{(|\Omega| - 1) \cdot (\mathcal{T} - 1)^2} \tag{6.13}$$

$$\mathbf{R} \triangleq \mathbf{I}_M - \mathbf{a}_n \mathbf{a}_n^T. \tag{6.14}$$

Preuve D'après l'équation (6.7), la distribution asymptotique de $\hat{\mathbf{u}}_1$ est :

$$\hat{\mathbf{u}}_1 \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{a}_n, (|\Omega| - 1)^{-1} \mathbf{V})$$

Proposition 6.2.3 *A partir des conditions de la Proposition 6.2.2, l'espérance asymptotique du carré de la distance entre* $\hat{\mathbf{u}}_1$ *et* $\mathbf{a}_n$ *est :*

$$\mathbb{E}\left\{\|\hat{\mathbf{u}}_1 - \mathbf{a}_n\|^2\right\} = (M - 1) \cdot \sigma^2(\mathcal{T}).$$

Preuve D'après l'équation (6.12), la distance asymptotique au carré entre $\hat{\mathbf{u}}_1$ et $\mathbf{a}_n$ est :

$$\|\hat{\mathbf{u}}_1 - \mathbf{a}_n\|^2 = \sigma^2(\mathcal{T}(\Omega)) \cdot \Xi$$

où la variable aléatoire $\Xi \sim \chi^2(M-1)$ est distribuée selon une distribution du χ^2 avec $M - 1$ degrés de liberté.

Test d'homogénéité du χ^2 : On suppose deux variables aléatoires indépendantes $\hat{\mathbf{u}}(\Omega_1)$, $\hat{\mathbf{u}}(\Omega_2)$ distribuées selon les lois :

$$\hat{\mathbf{u}}(\Omega_1) \sim \mathcal{N}(\mathbf{a}_n, \sigma^2(\mathcal{T}(\Omega_1)) \cdot \mathbf{R}) \tag{6.15}$$

$$\hat{\mathbf{u}}(\Omega_2) \sim \mathcal{N}(\mathbf{a}_{n'}, \sigma^2(\mathcal{T}(\Omega_2)) \cdot \mathbf{R}) \tag{6.16}$$

On souhaite élaborer le test de l'hypothèse $H_0 = \{\mathbf{a}_n = \mathbf{a}_{n'}\}$ contre $H_1 = \{\mathbf{a}_n \neq \mathbf{a}_{n'}\}$.

Si $\mathbf{a}_n = \mathbf{a}_{n'}$, la distance au carré entre $\hat{\mathbf{u}}(\Omega_1)$ et $\hat{\mathbf{u}}(\Omega_2)$ est donnée par :

$$\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2 = \left(\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))\right) \Xi \tag{6.17}$$

où $\Xi \sim \chi^2(M-1)$. Par conséquent :

$$P\left(\frac{\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2}{\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))} \geq q_{\chi^2_{M-1}}(\alpha)\right) = \alpha$$

où $q_{\chi^2_{M-1}}(\alpha)$ est le quantile de niveau α de la loi χ^2 avec $M-1$ degrés de liberté.

Par conséquent le test d'homogénéité du χ^2 de niveau α consiste à rejetter l'hypothèse H_0 si :

$$\|\hat{\mathbf{u}}(\Omega_1) - \hat{\mathbf{u}}(\Omega_2)\|^2 \geq q_{\chi^2_{M-1}}(\alpha) \cdot \left(\sigma^2(\mathcal{T}(\Omega_1)) + \sigma^2(\mathcal{T}(\Omega_2))\right) \tag{6.18}$$

Dans nos experiences, nous avons choisi la valeur $1-\alpha = 99\%$, qui correspond dans le cas stéréophonique ($M=2$) au quantile $q_{\chi^2_{M-1}(\alpha)} = 6.63$.

Conclusion : Les résultats asymptotiques que nous avons obtenus dans cette section, qui relient l'estimateur de direction local à la mesure de fiabilité (également locale), vont être exploités dans l'étape de *clustering* que nous développons dans la section suivante.

6.3 La mesure de proximité de DEMIX

Quelle que soit la méthode de *clustering* utilisée pour estimer les directions du mélange, nous devons définir une *mesure de proximité* dont le rôle est de quantifier le degrée de similitude entre les points $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$ estimés lors de l'étape d'extraction et de sélection des caractéristiques. Ces points étant indexés par une région Ω, par la suite nous employons parfois le terme de région pour parler de ces points et vis versa. Les points $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$ étant définis par une direction locale estimée $\hat{\mathbf{u}}(\Omega)$ et une fiabilité empirique $\widehat{\mathcal{T}}(\Omega)$, nous définissons dans un premier temps la distance entre deux directions locales, puis en s'appuyant sur les résultats asymptotiques de la section 6.2, nous définissons la distance entres deux points.

Distance entre directions : Du fait de l'indétermination de signe dans la définition d'une direction, la distance entre deux vecteurs unitaires génériques $\mathbf{u}_1, \mathbf{u}_2$, qui représentent des directions locales de sources, doit être définie avec attention. Deux directions locales sont proches l'une de l'autre si l'angle entre ces deux directions est faible, c.-à-d. quand $|\langle \mathbf{u}_1, \mathbf{u}_2 \rangle|$ est proche de 1. Par conséquent, on peut définir le carré de la distance d'une direction par :

$$d^2(\mathbf{u}_1, \mathbf{u}_2) \stackrel{\triangle}{=} \min_{|z|=1, z\in\mathbb{C}} \|\mathbf{u}_1 - z\mathbf{u}_2\|^2 = 2\left(1 - |\langle \mathbf{u}_1, \mathbf{u}_2 \rangle|\right). \tag{6.19}$$

Distance entre deux points : Selon le modèle développé à la section 6.2, deux points Ω_1 et Ω_2 doivent appartenir au même cluster, si les distributions asymptotiques gaussiennes de leur direction $\hat{\mathbf{u}}(\Omega_1)$ et $\hat{\mathbf{u}}(\Omega_2)$ ont la même moyenne, c.-à-d. $\mathbf{u}(\Omega_1) = \mathbf{u}(\Omega_2)$.

Le test d'homogénéité du χ^2 exprimé par l'équation (6.18) permet alors de tester cette hypothèse. Si cette hypothèse est acceptée, on dit alors que les deux points sont

suffisamment proches. D'après l'équation (6.18), la probabilité que les deux points soient *suffisamment proches* dépend uniquement de la distance :

$$\mathrm{D}_{\mathcal{T}}\left((\hat{\mathbf{u}}(\Omega_1), \mathcal{T}(\Omega_1)), (\hat{\mathbf{u}}(\Omega_2), \mathcal{T}(\Omega_2))\right)$$

où $\mathrm{D}_{\mathcal{T}}$ est définie par :

$$\mathrm{D}_{\mathcal{T}}\left((\mathbf{u}_1, \mathcal{T}_1), (\mathbf{u}_2, \mathcal{T}_2)\right) \triangleq \frac{d(\mathbf{u}_1, \mathbf{u}_2)}{\sqrt{\sigma^2(\mathcal{T}_1) + \sigma^2(\mathcal{T}_2)}} \tag{6.20}$$

Cette mesure n'est pas sans rappeler le contraste de Fisher [DH73] qui est utilisé notamment pour l'analyse discriminante de données, afin de maximiser la variance inter-classe tout en minimisant les variances intra-classes. Il faut aussi noter que les valeurs $\mathcal{T}(\Omega_1)$ et $\mathcal{T}(\Omega_2)$ ne sont pas accessibles dans la pratique. Par conséquent nous les substituons par les valeurs de leur estimée empirique $\widehat{\mathcal{T}}(\Omega_1)$ et $\widehat{\mathcal{T}}(\Omega_2)$, ou bien par une quantité plus pessimiste donnée par la valeur de la fiabilité empirique robuste $\tilde{\mathcal{T}}(\Omega) < \widehat{\mathcal{T}}(\Omega)$ définie à l'équation (6.9), et dépendant d'un autre quantile $q(\alpha)$.

6.4 2$^{\mathrm{e}}$ étape : Estimation des paramètres du mélange par classification

Pour estimer les directions du mélanges, nous combinons les estimations locales de la première étape à l'aide d'un algorithme de clustering. Il existe trois grandes familles d'algorithmes de clustering [TK03] :

1. Les algorithmes itératifs basés sur une fonction de coût ;
2. Les algorithmes hierarchiques ;
3. Les algorithmes séquentiels ;
4. L'approche Bayésienne ;

les algorithmes itératifs basés sur une fonction de coût : Nous pourrions faire le clustering avec un algorithme standard comme le K-means [HW79] qui cherche à minimiser la variance intra-cluster par une procédure itérative, cependant cette approche comporte de nombreux défauts [XI05] dont les deux principaux sont résumés ci-dessous :

1. Le minimum (local) dans lequel convergence l'algorithme K-means dépend fortement de son initialisation qui est aléatoire. Ainsi, dans la pratique, comme utilisé par la méthode DUET, le clustering est généralement exécuté I fois avec I initialisations différérentes et la solution qui minimise la variance intra-cluster est retenue. Une autre solution pour améliorer l'algorithme K-Means [HW79], qui est adoptée par la méthode *enhanced-LBG* (ELBG) [PR00, PR01], consiste à utiliser un mécanisme semblable aux algorithmes génétiques afin d'éviter de converger dans un minimum local. Nous évaluerons les performances de cet algorithme à la partie expérimentale développée au chapitre 7.

2. Le nombre N de clusters doit être fixé à l'avance. Il existe des heuristiques pour évaluer automatiquement le nombre de clusters, qui consistent à exécuter l'algorithme de clustering pour différentes valeurs de N, puis à estimer la valeur N suivant un critère heuristique. Beaucoup de critères ont été proposés, mais leurs performances dépendent fortement des données [XI05].

Les algorithmes hierarchiques : Les algorithmes de clustering hiérarchiques [TK03] cherchent à représenter les clusters sous forme d'une hiérarchie, qui est généralement un dendogramme. Ces méthodes permettent d'estimer automatiquement le nombre de clusters par des heuristiques, cependant ils sont généralement d'une complexité calculatoire en $O(n^2)$ voir $O(n^3)$ en fonction du nombre n de points qui dans le cas des signaux audio peut être très important. Par exemple, pour un signal de durée $d = 10$ s, de fréquence d'échantillonnage $fe = 10$ kHz, avec un taux de recouvrement de la TFCT de $r = 2$, et un nombre d'échelles d'analyse de la TFCT de $nbE = 10$, le nombre de points à traiter est : $n = d \times fe \times r \times nbE = 2\ 000\ 000$ de points.

Les algorithmes séquentiels : Le principe des algorithmes de clustering séquentiels est que les données d'entrées sont présentées une seule fois (ou seulement quelques fois) à l'algorithme et dans un certain ordre. Ces algorithmes ont l'avantage d'avoir une faible complexité calculatoire et de ne dépendent pas de l'initialisation. Il est également possile d'estimer automatiquement le nombre de clusters à l'aide d'heuristiques.

L'approche bayésienne : Il est possible de se placer dans une approche bayésienne afin d'estimer les directions du mélange, y compris pour estimer le nombre de sources du mélange. En effet les modèles de mélange par processus de Dirichlet (DPM) [Fer73, Ant74, Ras00, DT07] permettent d'estimer automatiquement le nombre de composantes du mélange. Cette approche n'a pas été développée dans le cadre de cette thèse mais semble particulièrement intéressante. Cependant il faut noter que l'inférence dans ce modèle est basée sur l'échantillonnage de Gibbs et par conséquent a de grandes chances de nécessiter un temps de calcul important.

Les algorithmes DEMIX : Les algorithmes DEMIX (Direction Estimation of the Mixing matrIX) que nous proposons et que nous définissons dans cette section appartiennent à la famille des algorithmes séquentiels. Nous allons dans un premier temps présenter DEMIX-Instantané, qui comme son nom l'indique est adapté aux mélanges instantanés, puis nous présenterons ensuite DEMIX-Anéchoïque pour l'estimation des directions d'un mélange anéchoïque. Les algorithmes DEMIX sont assez proches de l'algorithme BSAS (Basic Sequential Algorithmic Scheme) [TK03]. L'idée de base de l'algorithme BSAS est la suivante : à chaque fois qu'un point est présenté à l'algorithme, ce point est :

- soit assigné à un cluster existant si la distance du point à l'un des clusters est inférieur à un certain seuil dit de dissemblance ;
- soit assigné à un nouveau cluster créé pour l'occasion.

Le comportement de l'algorithme BSAS dépend de trois paramètres importants :

1. La mesure de distance $D(\cdot,\cdot)$ entre un point et un cluster ; Nous avons défini à l'équation (6.20) la mesure de distance $\mathrm{D}_{\mathcal{T}}(\cdot,\cdot)$ entre deux points. Dans DEMIX, la distance entre un point Ω et un cluster C_{k} est défini par la distance $\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(\Omega),\widehat{\mathcal{T}}(\Omega)\right),\left(\hat{\mathbf{u}}(\Omega_{\mathrm{k}}),\widehat{\mathcal{T}}(\Omega_{\mathrm{k}})\right)\right)$ entre ce point et un point particulier du cluster appelé centroïde.
2. La valeur du seuil de dissemblance ζ qui est utilisé pour décider si un point est *suffisamment proche* d'un cluster pour pouvoir appartenir à celui-ci ; Dans DEMIX un point Ω est *suffisamment proche* d'un cluster si le test d'homogénéité du χ^2 entre ce point et le centroïde Ω_k du cluster est accepté, c.-à-d. si :

$$\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(\Omega),\widehat{\mathcal{T}}(\Omega)\right),\left(\hat{\mathbf{u}}(\Omega_{\mathrm{k}}),\widehat{\mathcal{T}}(\Omega_{\mathrm{k}})\right)\right)\leq\zeta \tag{6.21}$$

   avec $\zeta=\sqrt{q_{\chi^2_{M-1}}(\alpha)}$.
3. L'ordre dans lequel les points sont présentés à l'algorithme. Dans le cas de DEMIX, les points Ω sont présentés à l'algorithme selon l'ordre décroissant de leur fiabilité $\widehat{\mathcal{T}}(\Omega)$. Contrairement à l'algorithme BSAS, qui considère les points de la séquence les uns après les autres, dans DEMIX, quand un cluster est créé, tous les points de la séquence initiale qui sont considérés comme *suffisamment proches* de ce cluster, sont ajoutés à ce cluster. Par conséquent, un point peut appartenir à plusieurs clusters à la fois.

Contrairement à l'algorithme BSAS, une étape supplémentaire est rajoutée à la fin de l'algorithme DEMIX dans le but d'éliminer les clusters non significatifs.

6.4.1 DEMIX-Instantané

La première étape de l'algorithme consiste à créer les clusters de façon itérative en sélectionant le point Ω_{k} ayant la plus grande fiabilité $\widehat{\mathcal{T}}(\Omega_{\mathrm{k}})$, et en agrégeant autour de lui les points *suffisamment proches* de $\hat{\mathbf{u}}(\Omega_{\mathrm{k}})$.

Le nombre K de clusters créés est déterminé automatiquement par l'algorithme et dépend de la structure du diagramme de dispersion des points $\left(\hat{\mathbf{u}}(\Omega),\widehat{\mathcal{T}}(\Omega)\right)$.

La seconde étape de l'algorithme est d'estimer le centroïde $\hat{\mathbf{u}}^c_{\mathrm{k}}$ de chaque cluster en sélectionant les points du cluster ayant les plus grandes fiabilités (voir section 6.4.1.2) puis en faisant une somme pondérée de ces points selon leur fiabilité. Enfin, nous faisons un test statistique pour éliminer les clusters qui sont considérés comme non suffisamment fiables. Le nombre de clusters $\widehat{N}\leq\mathrm{K}$ restant à la fin de l'algorithme fournit l'estimation du nombre de sources, et les centroides de ces clusters sont les estimations des directions de la matrice de mélange. Nous détaillons maintenant les différentes étapes de l'algorithme.

6.4.1.1 Création des clusters

La première étape de l'algorithme consiste à créer itérativement K clusters $C_{\mathrm{k}} \subset P$ où P est l'ensemble des régions Ω du diagramme de dispersion.

1.1) initialiser : $\mathrm{K} = 0$, $P_{\mathrm{K}} = P_0 = P$;

1.2) sélectionner la région $\Omega_{\mathrm{K}} \in P_{\mathrm{K}}$ ayant la plus grande fiabilité :

$$\Omega_{\mathrm{K}} \stackrel{\triangle}{=} \arg \max_{\Omega \in P_{\mathrm{K}}} \widehat{\mathcal{T}}(\Omega);$$

1.3) créer le cluster C_{K} avec l'ensemble des régions $\Omega \in P$ telles que $\hat{\mathbf{u}}(\Omega)$ soit *suffisamment proche* de $\hat{\mathbf{u}}(\Omega_{\mathrm{k}})$;

1.4) mettre à jour $P_{\mathrm{K}+1} = P_{\mathrm{K}} \setminus C_{\mathrm{K}}$ en retirant de P_{K}, les points du cluster C_{K} qui sont encore dans P_{K} ;

1.5) arrêter si $P_{\mathrm{K}} = \emptyset$, sinon incrémenter $\mathrm{K} \leftarrow \mathrm{K}+1$ et retourner à l'étape 1.2.

Remarquons qu'à l'étape 1.3, le cluster nouvellement créé peut contenir des régions déjà présentes dans un cluster précédemment créé.

6.4.1.2 Estimation des directions

Après avoir créé K clusters $\{C_{\mathrm{k}}\}_{\mathrm{k}=1}^{\mathrm{K}}$, nous en estimons les centroïdes $\mathbf{u}(C_{\mathrm{k}})$.

Comme on peut l'observer sur le diagramme de dispersion pondéré de la figure 6.3, la distribution des points autour d'une direction du mélange est symétrique uniquement à partir d'une certaine valeur de la fiabilité qui dépend de la distance de cette direction à ces directions voisines. Par conséquent, pour que l'estimation des directions ne soit pas biaisée, il est préférable de les estimer à partir d'une sélection $C'_{\mathrm{k}} \subset C_{\mathrm{k}}$ des « meilleurs » points des clusters.

Pour cela nous allons sélectionner les points pour lesquels la fiabilité est supérieure à un seuil adaptatif. Pour définir ce seuil, nous posons comme condition qu'un point ne puisse pas être sélectionné pour estimer les centroides de deux clusters différents. Afin d'avoir néanmoins un nombre de points le plus grand possible pour l'estimation du centroide, nous posons comme critère de choisir comme seuil la fiabilité la plus faible possible sous la contrainte qu'aucun point n'appartenant à deux clusters différents ne puisse être sélectionné :

$$\eta_{\mathrm{k}} = \min \eta \quad \text{s.c. } \nexists \Omega \in C_{\mathrm{k}} \cap [\cup_{j \neq \mathrm{k}} C_j] \diagdown \mathcal{T}(\Omega) > \eta \tag{6.22}$$

Nous faisons alors l'estimation des centroides suivant les trois étapes :

2.1) déterminer le seuil de fiabilité :

$$\eta_{\mathrm{k}} \stackrel{\triangle}{=} \max_{\Omega \in C_{\mathrm{k}} \cap \left[\cup_{j \neq \mathrm{k}} C_j\right]} \widehat{\mathcal{T}}(\Omega) \tag{6.23}$$

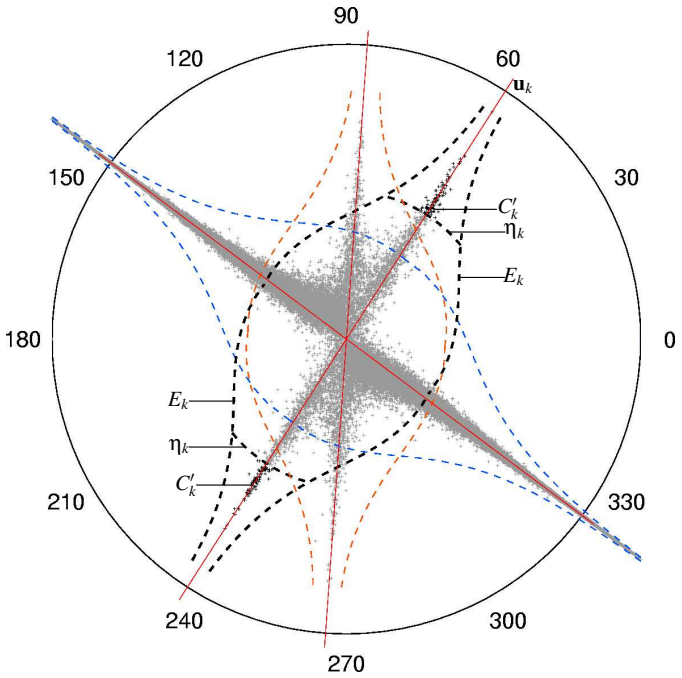


FIG. 6.3 – Illustration de la façon, d'obtenir le cluster C'_{k} par seuillage adaptatif η_{k} du cluster C_{k}, puis d'estimer la direction $\mathbf{u}_{\mathrm{k}}$. Le diagramme de dispersion est celui des points $\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right)$ construit de la même façon que pour la figure 6.2(b), mais pour un mélange différent. Le contour E_{k} représente les limites du cluster C_{k} définies par les points $(\mathbf{u}, \mathcal{T})$ tels que $\mathrm{D}_{\mathcal{T}}\left((\mathbf{u}, \mathcal{T}), \left(\hat{\mathbf{u}}(\Omega_{\mathrm{k}}), \widehat{\mathcal{T}}(\Omega_{\mathrm{k}})\right)\right) = \zeta$

2.2) sélectionner uniquement les points qui ont une fiabilité empirique suffisante :

$$C'_{\mathrm{k}} \stackrel{\triangle}{=} \left\{ \Omega \in C_{\mathrm{k}} \,\middle|\, \widehat{\mathcal{T}}(\Omega) \geq \eta_{\mathrm{k}} \right\}.$$

La figure 6.3 illustre la façon d'obtenir le cluster C'_{k}.

2.3) estimer le centroide $\mathbf{u}(C_{\mathrm{k}})$ à partir des points C'_{k}.

A la lumière du modèle statistique développé à la section 6.2, Eq. (6.12)-(6.14), chaque direction locale $\hat{\mathbf{u}}(\Omega)$ du cluster seuillé C'_{k} est distribuée selon $\mathcal{N}(\mathbf{u}_{\mathrm{k}}, \sigma^2(\mathcal{T}) \cdot \mathbf{R})$. L'estimateur non biaisé à variance minimum de la « vraie » direction $\mathbf{u}_{\mathrm{k}}$ est donné par :

$$\mathbf{v}_{\mathrm{k}} \stackrel{\triangle}{=} \frac{\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\mathcal{T}(\Omega)) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\mathcal{T}(\Omega))} \tag{6.24}$$

En pratique, puisque les vecteurs de direction $\hat{\mathbf{u}}(\Omega)$ sont seulement définis au signe près, nous multiplions chaque direction par le signe $\varepsilon(\Omega)$ tel que la corrélation $\langle \varepsilon(\Omega) \cdot \hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_{\mathrm{k}}) \rangle$ avec la direction du centroide $\hat{\mathbf{u}}(\Omega_{\mathrm{k}})$ soit positive. Aussi, les valeurs des fiabilités doivent être remplacées par leur estimée :

$$\tilde{\mathbf{v}}_{\mathrm{k}} \triangleq \frac{\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega)) \cdot \mathtt{sign}\left(\langle \hat{\mathbf{u}}(\Omega), \hat{\mathbf{u}}(\Omega_{\mathrm{k}}) \rangle\right) \cdot \hat{\mathbf{u}}(\Omega)}{\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\widehat{\mathcal{T}}(\Omega))}. \tag{6.25}$$

Finalement la formule d'estimation de la direction correspondant au cluster C_{k} est après normalisation :

$$\mathbf{u}(C_{\mathrm{k}}) \triangleq \tilde{\mathbf{v}}_{\mathrm{k}} / \|\tilde{\mathbf{v}}_{\mathrm{k}}\| \tag{6.26}$$

6.4.1.3 Elimination des clusters non fiables

La dernière étape de l'algorithme consiste à éliminer des clusters non fiables. Il peut arriver que deux clusters soient très proches et qu'il soit désirable de les fusionner. Cependant il se peut aussi que deux directions soient très proches l'une de l'autre, formant ainsi deux clusters très proches que l'on ne souhaite pas fusionner. Cependant, à condition de pouvoir définir la fiabilité d'un cluster, on peut faire l'hypothèse que deux clusters qui sont proches l'un de l'autre correspondent à deux directions distinctes du mélange, si et seulement si ces deux clusters ont chacun une grande fiabilité.

D'après l'équation (6.12) de la section 6.2, chaque direction $\mathbf{v}_{\mathrm{k}}$ est distribué selon la loi $\mathcal{N}(\mathbf{u}_{\mathrm{k}}, \sigma^2(\mathcal{T}(\Omega)) \cdot \mathbf{R})$, avec $\mathbf{R} = \mathbf{I}_M - \mathbf{a}_n \mathbf{a}_n^T$, et l'estimateur non biaisé à variance minimum, défini par l'équation (6.24), est distribué selon :

$$\mathbf{v}_{\mathrm{k}} \sim \mathcal{N}\left(\mathbf{u}_{\mathrm{k}}, \Big(\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\mathcal{T})\Big)^{-1} \mathbf{R}\right) \tag{6.27}$$

D'après la proposition 6.2.3 , Nous pouvons alors caractériser l'erreur d'estimation $\mathbf{v}_{\mathrm{k}}$ de la direction $\mathbf{u}_{\mathrm{k}}$ par :

$$\sigma^2(C_{\mathrm{k}}) \triangleq \mathbb{E}\left\{\|\tilde{\mathbf{v}}_{\mathrm{k}} - \mathbf{u}_{\mathrm{k}}\|^2\right\} = (M-1) \cdot \Big(\sum_{\Omega \in C'_{\mathrm{k}}} \sigma^{-2}(\mathcal{T}(\Omega))\Big)^{-1} \tag{6.28}$$

Dans la pratique, comme nous ne disposons pas de la fiabilité $\mathcal{T}(\Omega)$, nous la remplaçons par la valeur de la fiabilité empirique robuste $\tilde{\mathcal{T}}(\Omega)$. La mesure de variance $\sigma^2(C_{\mathrm{k}})$ peut être convertie en la fiabilité $\mathcal{T}(C_{\mathrm{k}})$ par la fonction réciproque de la bijection donnée par l'équation (6.13).

La procédure d'élimination des clusters non fiables consite à ré-itérer l'étape de *Création des clusters* de la section 6.4.1.1 mais en prenant comme données d'entrée, les centroides $\mathbf{u}(C_{\mathrm{k}})$ des clusters avec leurs fiabilité associées $\mathcal{T}(C_{\mathrm{k}})$ au lieu des points $\Big(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\Big)$. De cette façon, le cluster C_j fusionnera avec le cluster $C_o \neq C_j$ si :

$$\mathrm{D}_{\mathcal{T}}\left(\Big(\hat{\mathbf{u}}(C_j), \widehat{\mathcal{T}}(C_j)\Big), \Big(\hat{\mathbf{u}}(C_o), \widehat{\mathcal{T}}(C_o)\Big)\right) \leq \zeta_c \tag{6.29}$$

où ζ_c est le seuil de l'équation (6.21) que nous utilisons à la place de ζ.

Le modèle développé à la section 6.2 n'est pas valable pour les points $\Big(\hat{\mathbf{u}}(C_{\mathrm{k}}), \widehat{\mathcal{T}}(C_{\mathrm{k}})\Big)$, si bien que la valeur du seuil ζ_c ne correspond pas à un quantile de la loi du χ^2. Dans nos experiences, nous avons utilisé une valeur de $\zeta_c = 2.33\,\zeta$.

Remarquons aussi que si nous connaissons le nombre N de directions du mélange, cette étape d'élimination des clusters consiste simplement à garder les N clusters qui ont les plus grandes valeurs de fiabilité.

6.5 Estimation des délais intercanaux

Nous présentons dans cette section notre contribution à l'estimation des délais $\Delta_n = [\delta_{1n}, \delta_{2n}, \ldots, \delta_{Mn}]^T \in \mathbb{R}^M, \delta_{1n} = 0$ d'une direction anéchoïque :

$$\mathbf{a}_n(f) = [a_{1n}, a_{2n}e^{-2i\pi\delta_{2n}f}, \ldots, a_{Mn}e^{-2i\pi\delta_{Mn}f}]^T \in \mathbb{C}^M.$$

Comme mentionné à la section 2.2.3, si le délai entre deux canaux est supérieur à un échantillon du signal, il n'est pas possible d'estimer ce délai uniquement à partir d'un point temps-fréquence. Il est nécessaire de rassembler différents points temps-fréquence, à différentes fréquences, pour lesquels la même source est active. Nous présentons dans un premier temps la méthode dans le cas stéréophonique, pour lequel un seul délai par source doit être estimé, puis nous étendons la méthode pour un nombre de canaux supérieur à deux. Nous verrons ensuite comment on insère l'étape d'estimation des délais dans l'algorithme DEMIX-Anéchoïque.

6.5.1 Principe de la méthode

Cas où une seule source est active : Afin d'expliquer l'idée de base de la méthode, nous supposons dans un premier temps qu'une seule source n est active à la trame t. Dans ce cas, à cette trame, pour toutes les fréquences, le rapport de DUET satisfait : $R_{21}(t,f) = \tan(\widehat{\theta}(t,f))e^{i\widehat{\phi}(t,f)} \approx \frac{a_{2n}}{a_{1n}}e^{-2i\pi f\delta_n}$, et par conséquent, la Transformée de Fourier Inverse (TFI) de $R_{21}(t,f)/|R_{21}(t,f)| \approx e^{-2i\pi f\delta_n}$ est un Dirac à l'instant δ_n :

$$r_{21}(\tau) \stackrel{\triangle}{=} \int \frac{R_{21}(t,f)}{|R_{21}(t,f)|}e^{i2\pi f\tau}df \approx \delta(\tau - \delta_n) \tag{6.30}$$

La méthode GCC-PHAT [KC76] consiste à détecter le pic de la fonction $r_{21}(\tau)$ définie à l'équation (6.30) avec l'estimateur suivant :

$$\widehat{\delta}_n \stackrel{\triangle}{=} \arg\max_{\tau} r_{21}(\tau) \tag{6.31}$$

Cas où plusieurs sources sont actives : En pratique, on peut rarement observer une trame entière où une seule source est active.

Nous supposons que pour estimer le délai δ_n, nous avons identifié le point Ω_n de plus grande fiabilité et en particulier son profil d'intensité $\text{abs}(\hat{\mathbf{u}}(\Omega_n))$, de la même façon qu'à l'étape 1.2 de la section 6.4.1.1. On rappel que $\text{abs}(\cdot)$ est la fonction de $\mathbb{C}^M$ dans $\mathbb{R}^M$ qui à tout élément du vecteur de $\mathbb{C}^M$ fait correspondre le module de cet élément.

Le principe de notre approche consiste de façon semblable à la méthode de Bofill [Bof03] à :

1. Créer un cluster (temporaire) $\tilde{C}_n$, obtenu à partir du profil d'intensité, abs($\hat{\mathbf{u}}(\Omega_n)$), et qui a pour but de regrouper grossièrement les points qui sont proches de la direction n, pour lesquelles la source S_n est essentiellement l'unique source active. Plus formellement, le cluster temporaire $\widetilde{C}_n$ est constitué de l'ensemble des points Ω tels que abs($\hat{\mathbf{u}}(\Omega)$) soit *suffisamment proche* de abs($\hat{\mathbf{u}}(\Omega_n)$), c.-à-d. les points Ω tels que :

$$d(\text{abs}(\hat{\mathbf{u}}(\Omega)), \text{abs}(\hat{\mathbf{u}}(\Omega_n))) \leq \zeta_2 \cdot \sigma(\tilde{\mathcal{T}}(\Omega_n)) \tag{6.32}$$

   où $\tilde{\mathcal{T}}(\Omega_n)$ est la fiabilité empirique robuste définie par l'équation (6.9) et ζ_2 est un seuil qui a été reglé à la valeur $\zeta_2 = \zeta$ dans nos experiences.

2. Estimer le délai δ_n à partir des points du cluster $\tilde{C}_n$.

Dans chacune des régions $\Omega \in \tilde{C}_n$, la phase de la direction complexe $\hat{\mathbf{u}}(\Omega)$ est $e^{i\phi(\Omega)} \approx e^{-2i\pi\delta_n f}$, où $f = f(\Omega)$ est la « fréquence centrale » de la région temps-fréquence. La qualité de cette approximation est liée à la valeur de $\sigma^2(\widehat{\mathcal{T}}(\Omega))$ donnée par l'équation (6.13). En pondérant par leur précision $1/\sigma^2(\widehat{\mathcal{T}}(\Omega))$ toutes les estimations correspondant aux régions temps-fréquence $\Omega \in \tilde{C}_n$ ayant f comme fréquence centrale, on peut s'attendre à obtenir une meilleur estimée de la phase pour la source n, à la fréquence f. Nous proposons alors l'estimateur suivant :

$$R_{\tilde{C}_n}(f) \stackrel{\triangle}{=} \frac{\sum_\Omega w_f(\Omega) e^{i\widehat{\phi}(\Omega)}}{\sum_\Omega w_f(\Omega)} \approx e^{-2i\pi\delta_n f} \tag{6.33}$$

avec

$$w_f(\Omega) \stackrel{\triangle}{=} \begin{cases} 1/\sigma^2(\widehat{\mathcal{T}}(\Omega)) & \text{si } \Omega \in \widetilde{C}_n \text{ et } f = f(\Omega) \\ 0 & \text{sinon} \end{cases} . \tag{6.34}$$

Finalement, la TFI de $R_{\tilde{C}_n}(f)$ (voir équation (6.35)) fournit une approximation, dont la qualité dépend de celle de l'équation (6.33), d'un Dirac à l'instant δ_n.

$$r_{\tilde{C}_n}(\tau) \stackrel{\triangle}{=} \int R_{\tilde{C}_n}(f) e^{i2\pi f \tau} df \approx \delta(\tau - \delta_n) \tag{6.35}$$

Le plus haut pic de cette fonction est l'estimateur du délai de la source n :

$$\widehat{\delta}_n \stackrel{\triangle}{=} \arg\max_\tau r_{\tilde{C}_n}(\tau) \tag{6.36}$$

En pratique, on considère qu'un pic est *correctement identifié* si l'amplitude du pic $r_{\tilde{C}_n}(\widehat{\delta}_n)$ est supérieure d'au moins 3 dB à tout autre pic de $r_{\tilde{C}_n}$. Cette notion de pic *correctement identifié* sera utilisée lorsque nous expliquerons comment insérer l'étape d'estimation des délais dans DEMIX-Anéchoïque.

Remarquons que notre estimateur de délai étend GCC-PHAT au cas où il y a plusieurs sources. Il est aussi dans le principe assez proche de celui de Bofill que nous avons décrit à la section 2.2.4.

6.5.2 Estimation des délais dans le cas où il y a plus de 2 canaux

Dans le cas où le nombre M de canaux est supérieur à 2, il y a $M-1>1$ délais intercanaux à estimer, dont les valeurs dépendent du canal servant de référence.

Remarquons que même pour un point temps-fréquence où il y a essentiellement une seule source active, si l'intensité sur le canal m est proche de zéro (c.-à-d. $\Re\mathrm{e}X_m(t,f)\approx 0$ et $\Im\mathrm{m}X_m(t,f)\approx 0$), alors l'estimation de la phase $\angle X_m(t,f)=\tan^{-1}\left(\frac{\Im\mathrm{m}X_m(t,f)}{\Re\mathrm{e}X_m(t,f)}\right)$ sur ce canal sera instable, de même que la différence de phase entre un canal $k\neq m$ et le canal m.

Afin d'éviter ces instabilités de phase, nous proposons de choisir comme référence, pour le cluster C_K, le canal qui a la plus grande intensité du profil d'intensité $\mathrm{abs}(\hat{\mathbf{u}}(\Omega_\mathrm{K}))=(u_m)_{m=1}^M$ de la région temps-fréquence Ω_K. A partir du canal de référence qui est par conséquent le canal $m_\mathrm{K}\stackrel{\triangle}{=}\arg\max_m|u_m|$, nous estimons les délais intercanaux $\widehat{\delta}'_{m\,\mathrm{K}}$ entre les canaux $m\neq m_\mathrm{K}$ et le canal de référence m_K, en utilisant la méthode d'estimation des délais du cas stéréophonique qui a été présentée à la section 6.5.1. Nous considérons que Δ_K est *correctement identifié* si tous les délais de Δ_K sont *correctement identifiés* par la méthode de la section 6.5.1, c.-à-d. si l'amplitude du plus grand pic est supérieure d'au moins 3 dB à tous les autres pics.

Discrétisation de la méthode : En pratique, la représentation temps-fréquence est calculée sur une grille temps-fréquence discrète. En conséquence, les estimateurs définis aux équations (6.35) et (6.36) fournissent seulement des valeurs discrètes des délais. Si la TFI de l'équation (6.35) est calculée avec les mêmes fréquences que celles de la TFCT $\mathbf{X}(t,f)$, alors la résolution temporelle de l'estimateur du délai donné à l'équation (6.36) est de 1 échantillon du signal. Il est néanmoins possible d'augmenter cette résolution en faisant du zéro-padding ou du « zoom spectral » [HS77] sur la fonction de l'équation (6.33).

6.5.3 DEMIX Anéchoïque

Nous détaillons maintenant l'algorithme DEMIX-Anéchoïque, qui étend DEMIX-Instantané au cas des mélanges anéchoïques. DEMIX-Anéchoïque est structuré avec les trois mêmes étapes que DEMIX-Instantané :

1. Création des clusters ;
2. Estimation des directions ;
3. Elimination des clusters non fiables.

La différence principale avec l'algorithme DEMIX-Instantané est dans l'étape de création des clusters, parce que chaque direction du mélange $\mathbf{a}_\mathrm{k}(f)$, en plus d'être caractérisée par un *profil d'intensité*, est aussi caractérisé par son/ses délai(s) intercanaux Δ_k. Par conséquent, la définition du centroide $\mathbf{u}_{C_\mathrm{k}}(f)\in\mathbb{C}^M$ doit être différente de celle du cas instantané. Celui-ci dépend maintenant de la fréquence f et est caractérisé par :

- un profil d'intensité $\mathrm{abs}(\mathbf{u}_{C_\mathrm{k}}(f))$ qui est **indépendant de la fréquence**.
- la phase des différents canaux qui **dépend de la fréquence** et des délais Δ_k.

Les deux principales différences par rapport à la méthode instantanée sont d'une part l'incorporation de l'étape d'estimation des délais que nous venons de présenter, et d'autre part une nouvelle façon de décider qu'une direction locale $\hat{\mathbf{u}}(\Omega)$ est *suffisamment proche* d'un cluster en tenant compte des phases et des délais.

6.5.3.1 Création des clusters et estimation des délais

Cette étape suit la même procédure itérative (décrite à la section 6.4.1.1) que DEMIX-Instantané, excepté pour l'étape 1.3 de création des clusters qui se divise maintenant en deux étapes :

1.3.a) Estimer les délais intercanaux Δ_{K} à partir du cluster temporaire $\widetilde{C}_K$ constitué des points $\Omega \in P_{\mathrm{K}}$ *suffisamment proches* de $\mathrm{abs}(\hat{\mathbf{u}}(\Omega_{\mathrm{K}}))$. Cette étape à été détaillé à la section 6.5.

1.3.b) **si** Δ_{K} est considéré comme *correctement identifié* (cf section 6.5) : définir le centroide $\mathbf{u}_{C_{\mathrm{K}}}(f)$ en utilisant le profil d'intensité $\mathrm{abs}(\hat{\mathbf{u}}(\Omega_{\mathrm{K}}))$ et les délais Δ_{K} ; créer le cluster C_{K} avec toutes les régions $\Omega \in P$ *suffisamment proches* de $\mathbf{u}_{C_{\mathrm{K}}}(f)$;
sinon : rejeter le cluster $C_{\mathrm{K}} \stackrel{\triangle}{=} \tilde{C}_{\mathrm{K}}$;

A l'étape 1.3.a, nous calculons les distances entre profils d'intensité. En contraste, à l'étape 1.3.b, nous devons calculer les distances entre les directions locales complexes $\hat{\mathbf{u}}(\Omega)$ et le centroide $\mathbf{u}_{C_{\mathrm{K}}}(f)$, dont la valeur dépend de la fréquence.

Pour illustrer ceci, prenons comme exemple le cas particulier où $M = 2$. La direction complexe d'une région Ω est alors donnée par : $\hat{\mathbf{u}} = [\cos\widehat{\theta},\ \sin\widehat{\theta} \cdot e^{i\widehat{\phi}}]^T$ tandis que la direction du centroide est donnée par : $\mathbf{u}_{C_{\mathrm{k}}}(f) = [\cos\widehat{\theta}_{\mathrm{k}},\ \sin\widehat{\theta}_{\mathrm{k}} \cdot e^{-i2\pi\widehat{\delta}_{\mathrm{k}} f}]^T$. Afin de calculer la distance entre la direction locale $\hat{\mathbf{u}}(\Omega)$ et le centroide $\mathbf{u}_{C_{\mathrm{K}}}(f)$, nous calculons la distance $\mathrm{D}_{\mathcal{T}}(.,.)$ pour la fréquence $f = f(\Omega)$, qui est la fréquence centrale de la région Ω. Rappelons que chaque région Ω est indexée par les indices (t, f) que nous avons omis pour facilité la lecture. Par conséquent, à l'étape 1.3.b, nous considérons comme *suffisamment proches* les régions $\Omega \in P$ telles que :

$$\mathrm{D}_{\mathcal{T}}\left(\left(\hat{\mathbf{u}}(\Omega), \widehat{\mathcal{T}}(\Omega)\right), \left(\mathbf{u}_{C_{\mathrm{K}}}(f(\Omega)), \widehat{\mathcal{T}}(\Omega_{\mathrm{K}})\right)\right) \leq \zeta, \tag{6.37}$$

Autrement dit, dans DEMIX-Anéchoïque, la distance $\mathrm{D}_{\mathcal{T}}$ et le seuil ζ relatifs à l'algorithme BSAS, sont les même que pour DEMIX-Instantané (voir équation (6.21)), mais dans DEMIX-Anéchoïque, cette distance dépend de la fréquence $f(\Omega)$ de la région Ω.

6.5.3.2 Estimation des directions

Après avoir créé les K clusters C_{k}, le profil d'intensité des centroides $\mathbf{u}_{C_{\mathrm{k}}}(f)$ pour lesquels les délais ont été *correctement identifiés* (cf étape 1.3.b), est ré-estimé de la même façon que pour l'étape d'estimation des directions du cas instantané (cf section 6.4.1.2). Les délais Δ_{k} du centroide obtenus à l'étape 1.3.b sont conservés tel quels.

6.5.3.3 Elimination des clusters non fiables

L'étape d'élimination des clusters non fiables est identique à celle du cas instantané de la section 6.4.1.3, à part pour l'équation (6.29), où la mesure de distance d qui est utilisée dans la définition de la distance $\mathrm{D}_{\mathcal{T}}$ (voir équation (6.20)) doit être changée par la distance d_c définie ci-dessous, pour prendre en compte le fait que dans le cas anéchoïque, les centroïdes dépendent de la fréquence f.

$$d_c(\mathbf{u}_{C_i}, \mathbf{u}_{C_j}) = \int d(\mathbf{u}_{C_i}(f), \mathbf{u}_{C_j}(f))df. \tag{6.38}$$

6.6 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche pour estimer les directions spatiales, ainsi que le nombre de sources, d'un mélange multicanal instantané ou anéchoïque.

La méthode DEMIX que nous proposons, comme pour la méthode TIFROM, se base sur une hypothèse de parcimonie plus faible que celle de DUET, puisque contrairement à DUET qui suppose que le support des sources est quasi-disjoint dans le plan temps-fréquence, DEMIX et TIFROM supposent que pour chaque source, il existe des régions temps-fréquence (une seule suffit dans le cas instantané) où cette source domine significativement les autres. L'estimation des directions est alors effectuée à partir des diagrammes de dispersion locaux correspondant à ces régions.

L'information spatiale contenue dans ces diagrammes de dispersions locaux, et plus particulièrement l'information spatiale des matrices de covariances locales correspondant à ces diagrammes de dispersion, est exploitée à l'aide d'un modèle statistique simple où l'on modélise la source dominante par une Gaussienne de dimension 1 et la contribution des autres sources par une Gausienne isotrope. L'exploitation de ce modèle permet d'obtenir :

1. une estimation locale de la direction de la source qui domine dans la région considérée et qui a la propriété d'invariance par rotation ;
2. une valeur de fiabilité qui rend compte de la confiance que l'on peut accorder à cette estimation locale de direction.

Ces estimées locales sont ensuite fusionnées dans un algorithme de clustering pour estimer les directions du mélange. Bien qu'un algorithme de clustering classique comme K-means où ELBG puisse être utilisé pour cette étape, nous proposons un algorithme ad-hoc qui permet d'estimer le nombre de sources du mélange, ainsi que d'estimer des délais supérieurs à un échantillon par une méthode qui étend GCC-PHAT au cas multi-sources.

Nous évaluons au chapitre suivant la robustesse de l'approche proposée par rapport à l'approche globale de DUET.

Chapitre 7

Évaluation des méthodes d'estimation des paramètres du mélange

Dans le but d'évaluer les algorithmes DEMIX-Instantané et DEMIX-Anéchoïque, nous proposons dans cette section trois expériences qui ont pour objectif de :

- comparer l'algorithme DEMIX-Instantané, avec différentes approches classiques de clustering, en particulier l'algorithme ELBG [PR00, PR01] (qui rappelons le est une version améliorée, pour éviter de converger dans des minimas locaux, de l'algorithme de clustering K-Means [HW79]) et des variantes de celui-ci ;
- évaluer les limites de DEMIX-Instantané sur des mélanges anéchoïques, en faisant varier progressivement le délai entre directions, d'une valeur très faible jusqu'à une grande valeur, dans le but de passer progressivement des conditions d'un mélange instantané, à un mélange anéchoïque ;
- comparer les performances de DEMIX-Anéchoïque et de DUET pour l'estimation des directions de mélanges anéchoïques obtenus par simulation de chambres anéchoïques.

Toutes les expériences sont effectuées dans le cas stéréophonique ($M = 2$).

7.1 Méthodes évaluées

Nous décrivons brièvement les différents algorithmes d'estimation des paramètres du mélange que nous évaluons au cours de ce chapitre.

7.1.1 Expériences sur des mélanges instantanés

Intention expérimentale :
Nous souhaitons :

1. Comparer l'algorithme DEMIX-Instantané avec l'état de l'art des algorithmes de clustering. Pour cela nous évaluons l'algorithme ELBG.

2. Evaluer l'intérêt de :
   - l'effet de « lissage » de l'estimation des directions locales par l'ACP au lieu de l'estimation ponctuelle de la méthode standard ;
   - l'effet de la pondération des directions locales par la mesure de fiabilité plutôt que par l'énergie.

Les quatre variantes de l'ELBG : Nous considérons 4 variantes de l'algorithme ELBG :

- ELBG sur les angles $\theta(t,f)$ bruts obtenus à partir des points temps-fréquences $\mathbf{X}(t,f)$. Autrement dit, il s'agit de la méthode standard où le clustering est fait à partir des points du diagramme de dispersion global ;
- WELBG (version *pondérée* de l'ELBG [ARB05]) sur les angles bruts $\theta(t,f)$ obtenus à partir des points temps-fréquences $\mathbf{X}(t,f)$ en utilisant les amplitudes $\rho(t,f) = \|\mathbf{X}(t,f)\|$ comme poids ;
- ELBG sur les angles $\widehat{\theta}(t,f)$ obtenus à partir de $\hat{\mathbf{u}}(t,f)$ après l'ACP.
- WELBG sur les angles $\widehat{\theta}(t,f)$ obtenus à partir de $\hat{\mathbf{u}}(t,f)$ après l'ACP, mais en exploitant la fiabilité en utilisant la valeur de pondération : $1 \Big/ \sigma^2\left(\widehat{\mathcal{T}}(t,f)\right)$ ($\sigma^2(\mathcal{T})$ étant défini à l'équation (6.13)).

Les différents algorithmes testés sont présentés sur le schéma de la figure 7.1.

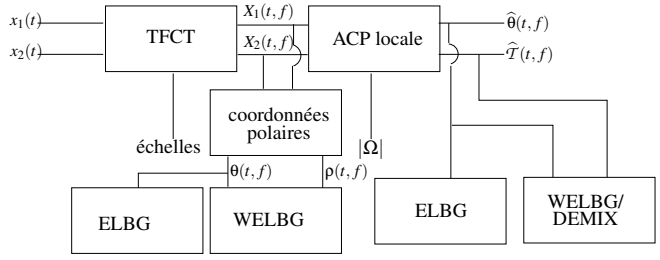


FIG. 7.1 – schéma bloc montrant les différents algorithmes de clustering testés à partir du flux de données

7.1.2 Expériences sur des mélanges anéchoïques

Nous comparons notre méthode DEMIX-Anéchoïque pour les mélanges anéchoïques avec l'estimation des paramètres du mélanges faite par DUET, ainsi qu'avec la méthode DEMIX-Instantané afin d'en évaluer les limites.

7.2 Conditions communes à l'ensemble des expériences

Nous décrivons dans cette section les mesures de performance utilisées pour nos expériences ainsi que les valeurs des paramètres utilisées par les méthodes évaluées.

7.2.1 Mesures de performance

Comme les deux méthodes DEMIX peuvent estimer non seulement les directions des sources, mais aussi le nombre de sources du mélange, nous proposons deux mesures pour évaluer les performances de chacune de ces caractéristiques.

Estimation du nombre de sources : Une première mesure de performance est le taux de succès dans l'estimation du nombre de sources. Cette mesure est uniquement appliquée à l'algorithme DEMIX, parce que les algorithmes DUET (basé sur l'algorithme de clustering K-means), ainsi que ELBG et ses variantes, n'estiment pas le nombre de sources.

Estimation précise des directions : Nous proposons une mesure de l'erreur moyenne d'estimation des directions appelée MDE (pour *mean direction error*), qui est la distance moyenne entre les vraies directions et les directions estimées. L'appariement des directions estimées avec les vraies directions est déterminé de façon à minimiser l'erreur moyenne d'estimation des directions.

Plus formellement, soit un mélange linéaire instantané, pour lequel les directions sont $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N]$ et soient les directions estimées : $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2, \dots, \widehat{\mathbf{a}}_N]$, le MDE est alors définit par :

$$\text{MDE } (\mathbf{A}, \hat{\mathbf{A}}) \stackrel{\triangle}{=} \min_{\pi \in S_N} \frac{1}{N} \sum_{n=1}^{N} d(\mathbf{a}_n, \hat{\mathbf{a}}_{\pi(n)}) \tag{7.1}$$

où S_N est le groupe des permutations de taille N. Remarquons qu'on aurait aussi pu choisir l'erreur maximale plutôt que l'erreur moyenne :

$$\min_{\pi \in S_N} \max_{1 \leq n \leq N} d(\mathbf{a}_n, \hat{\mathbf{a}}_{\pi(n)}),$$

cependant les résultats (non présentés ici) obtenus avec cette mesure étaient quasi-identiques à ceux obtenus avec le MDE. Afin de mesurer également l'erreur relativement à la distance entre les vraies directions, nous utilisons la mesure relative RMDE qui est le MDE divisé par la distance minimale entre les vraies directions :

$$\text{RMDE } (\mathbf{A}, \hat{\mathbf{A}}) \stackrel{\triangle}{=} \frac{\text{MDE } (\mathbf{A}, \hat{\mathbf{A}})}{\min_{n \neq n'} d(\mathbf{a}_n, \mathbf{a}_{n'})}. \tag{7.2}$$

Le RMDE est nul si et seulement si l'estimation est parfaite, et si le RMDE est proche de 1, cela signifie que l'erreur d'estimation des directions est du même ordre de grandeur que la distance entre les vraies directions, ce qui correspond évidemment à des performances médiocres.

Nous définissons des mesures de performance similaires pour le cas des mélanges anéchoïques en remplaçant dans les équations (7.1) et (7.2) la distance $d(.,.)$ par la distance $d_c(.,.)$ qui est définie à l'équation (6.38).

Performances en séparation : Afin de pouvoir comparer entre eux des algorithmes définis pour des mélanges instantanés et d'autres pour des mélanges anéchoïques, et qui par conséquent ne fournissent pas le même type de donnée en sortie, nous proposons d'évaluer ces algorithmes sur la tâche de séparation de sources, en utilisant un algorithme d'estimation des sources ayant comme donnée d'entrée l'estimation des directions. La mesure de performance en séparation que nous utilisons est la mesure du rapport signal à distortion (SDR) [GBVF03].

Pour l'estimation des sources, nous utilisons la méthode classique du masquage binaire qui a été présentée à la section 3.2.1.

7.2.2 Signaux d'évaluation

Les signaux sources : Nous évaluons les algorithmes de clustering sur des signaux de parole composés de 200 extraits de 5 secondes, de voix polonaises, échantillonnées à $fe = 4$ kHz. Ces signaux sont les signaux de référence qui ont été utilisés au cours du workshop : 2005 IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING et sont disponibles à l'URL `http://mlsp2005.conwiz.dk/`.

Nombre de mélanges testés : Pour chaque configuration des paramètres du mélange, que nous détaillerons pour chacunes des expériences que nous ferons, nous choisissons par tirage aléatoire $T = 20$ différentes configurations des sources parmi les 200 signaux sources disponibles. Ainsi, pour chaque algorithme testé, nous retiendrons comme résultat la moyenne du RMDE sur ces T mélanges.

7.2.3 Paramétrage des méthodes

Paramètres de la TFCT : Pour tous les algorithmes d'estimation des paramètres du mélange instantané, ainsi que nos méthodes DEMIX-Instantané et DEMIX-Anéchoïque, qui peuvent être utilisés en combinant les coefficients de la TFCT $\mathbf{X}(t, f)$ à différentes échelles, nous avons utilisé ensemble des tailles de fenêtre allant de 64 échantillons à 32768 par pas géométrique de 2.

Pour la méthode DUET, que nous utilisons pour estimer des mélanges anéchoïques, nous utilisons la version qui a été implémentée par ses auteurs [YR04], que nous remercions par la même occasion. La taille optimale de la fenêtre de la TFCT utilisée par la méthode est calculée en fonction de la fréquence d'échantillonnage, et est dans le cas de nos expériences de 256 échantilons, ce qui correspond à une taille de 63 ms. Cette taille de fenêtre est celle qui est utilisée par la méthode d'estimation des sources connaissant les paramètres du mélange, que nous utilisons pour évaluer les performances en séparation.

Pour l'ensemble des méthodes, des fenêtres de Hanning sont utilisées avec un recouvrement de 1/2.

Paramètres de DEMIX : Nous avons utilisé les mêmes valeurs des paramètres pour les méthodes DEMIX-Instantané et DEMIX-Anéchoïque, à part pour le paramètre ζ_2

qui n'est utilisé que par DEMIX-Anéchoïque.

La taille des voisinages pour calculer les ACP locales des méthodes DEMIX est de $|\Omega| = 10$ points.

Les valeurs des paramètres de seuillage sont les suivant :

- le seuil ζ réglant le test de χ^2 d'appartenance d'un point à un cluster lors de l'étape de création des clusters : $\zeta = 2.33$;
- le seuil ζ_c réglant le test de fusion des clusters lors de l'étape d'élimination des clusters : $\zeta_c = 5.43$;
- (Pour DEMIX-Anéchoïque seulement) le seuil ζ_2 réglant le test d'appartenance d'un point à un cluster temporaire basé sur le paramètre d'intensité, lors de l'étape de création des clusters : $\zeta_2 = 2.33$.

Initialisation des algorithmes de clustering : Puisque l'ELBG et ses variantes sont initialisés de façon aléatoire, et que la convergence de ces algorithmes est dépendante de l'initialisation, nous exécutons $I = 10$ fois chacun des tests que nous effectuons sur ces algorithmes, et nous gardons uniquement le résultat de l'instance ayant obtenu le meilleur résultat, c.-à-d. ayant obtenu la plus petite erreur MDE. Par conséquent les résultats obtenus pour ces algorithmes sont plutôt optimistes en comparaison de la méthode DEMIX qui ne nécessite pas d'initialisation.

7.3 Évaluation sur les mélanges instantanés

7.3.1 Protocole expérimental

Intention expérimentale : Nous souhaitons évaluer la robustesse des méthodes d'estimation des paramètres d'un mélange instantané en faisant varier le nombre de sources. Ainsi, il est évident qu'en augmentant le nombre de sources dans le mélange, la parcimonie de la représentation diminue et les directions des sources sont de plus en plus proches.

Afin d'évaluer le comportement des méthodes quand les directions se rapprochent sans pour autant que la parcimonie du mélange en soit affectée, nous proposons une deuxième expérience, où l'on fait uniquement varier la distance entre les directions.

1re expérience : N sources à espacement égal Pour la première experience, nous construisons des mélanges linéaires instantanés sans bruit, pour lesquels les directions des sources sont également espacées comme dans l'article [BZ01], avec un nombre de sources variant de $N = 2$ à $N = 15$. Pour l'algorithme DEMIX-Instantané, qui est le seul des algorithmes qui sache compter le nombre de sources, nous évaluons le pourcentage de réussite (parmi les T exemples) à compter les sources pour $N = 2$ jusqu'à $N = 11$.

2e expérience : 3 sources se rapprochant de plus en plus Pour la seconde expérience, nous plaçons 3 sources avec les angles suivants : $\theta_{l+2} = \frac{\pi}{4} + l\varepsilon\pi$, avec $l \in \{-1, 0, 1\}$. Dans cette expérience, nous faisons uniquement varier la distance angulaire

nombre de sources	2	3	4	5	6	7	8	9	10	11
DEMIX Inst	95	100	100	95	95	90	75	70	15	0

TAB. 7.1 – pourcentage d'estimation correcte du nombre de sources

$\varepsilon\pi$ dans le but d'évaluer la robustesse de l'algorithme quand les sources sont proches les unes des autres (ε faible).

7.3.2 Résultats

Nous observons (tableau 7.1) que jusqu'à $N = 7$ sources, DEMIX estime correctement le nombre de directions dans plus de neuf cas sur dix, mais quand $N > 10$ il échoue systématiquement. Ces résultats indiquent certainement que plus le nombre de sources est important, moins l'hypothèse qu'il y ait pour chaque source au moins une région temps-fréquence pour laquelle cette source est la seule à être active, est valide.

Comme on peut le voir sur les figures 7.2 et 7.3, DEMIX obtient les meilleurs performances en terme de RMDE, même en étant comparé avec les meilleurs instances (parmi les $I = 10$ correspondant à des initialisations différentes) des variantes d'ELBG. Notons qu'étant donné le fait que DEMIX n'estime pas toujours correctement le nombre de source du mélange, les résultats affichés pour DEMIX à la figure 7.2 sont le RMDE moyenné sur les exemples pour lesquelles DEMIX a correctement estimé le nombre de sources. Par conséquent le RMDE de DEMIX n'a pas été estimé pour $N > 10$, car il n'y avait plus d'exemple pour lequel le nombre de sources ait été correctement estimé.

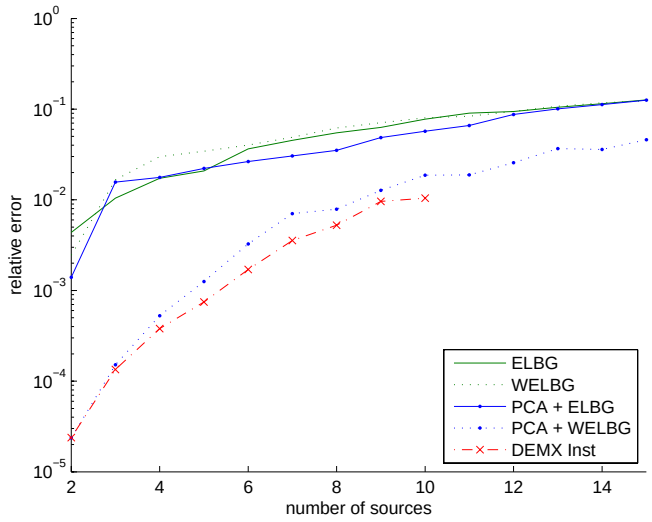


FIG. 7.2 – Erreur d'estimation des directions (en terme de RMDE) en fonction du nombre de sources pour les algorithmes DEMIX-Instantané (DEMIX-Inst) et les meilleurs instances (parmi les $I = 10$ initialisations) des quatre variantes de la méthode ELBG

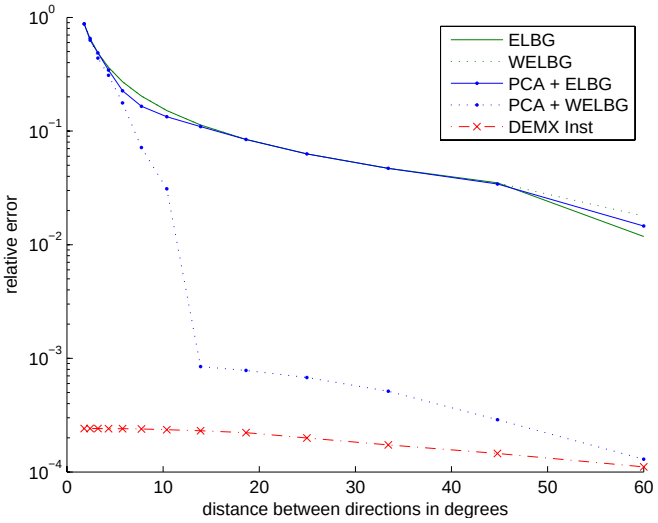


FIG. 7.3 – Erreur d'estimation des directions (en terme de RMDE) en fonction de la distance angulaire entre 3 sources, pour DEMIX-Instantané (DEMIX-Inst) et les meilleurs instances (parmi les $I = 10$ initialisations) des quatre variantes de la méthode ELBG

Le comportement de DEMIX est particulièrement remarquable dans la seconde expérience, pour laquelle les mélanges sont constitués d'uniquement 3 sources, qui peuvent être très proches les unes des autres. Le RMDE des quatre variantes d'ELBG approchent 1 quand la distance entre les vraies directions se rapproche de zéro (voir figure 7.3), ce qui indique que l'erreur d'estimation est presqu'aussi élevée que la distance entre les directions. Autrement dit les variantes d'ELBG confondent les directions.

A l'opposé, DEMIX-Instantané reste très robuste quand les directions sont très proches : Comme on peut le voir sur la figure 7.3, le RMDE de DEMIX-Instantané (DEMIX-Inst) reste inférieur à $3 \cdot 10^{-4}$ quelle que soit la distance entre les directions. Des expériences non reportées dans ce manuscrit montrent que la méthode DEMIX commence à « décrocher » à partir d'une distance entre directions inférieur à 10^{-3} degrés.

Nous remarquons, par l'observation des résultats des quatre variantes de ELBG, que le remplacement des estimations de directions locales ponctuelles (méthode standard) par celles obtenues par ACP locale, n'améliore pas significativement les résultats. Par contre on constate que :

- l'**utilisation de la fiabilité** , exploitant davantage les meilleurs estimées locales, a un **impacte déterminant** sur les résultats ;
- quand un nombre limité de sources est présent, mais que ces sources sont proches les unes des autres, le choix de l'algorithme de clustering qui effectue la classification (par opposition à l'étape de sélection des caractéristiques) est important, et DEMIX-Instantané obtient des performances bien meilleures que ses concurents.

7.4 Évaluation sur des mélanges anéchoïques synthétiques

Intention expérimentale :
Nous souhaitons :

1. évaluer la robustesse de DEMIX-Instantané sur des mélange anéchoïques et faiblement anéchoïque (ayant de faibles délais).
2. comparer les performances de DEMIX-Anéchoïque avec la méthode DUET de l'état de l'art, et évaluer la robustesse de ces méthodes pour estimer de grands délais.

7.4.1 Protocole expérimental

Nous proposons une expérience sur des mélanges anéchoïques, où l'on fait varier progressivement le délai entre les canaux. Le but étant de passer progressivement des conditions d'un mélange instantané pour lequel les délais sont nuls, à des mélanges anéchoïques ayant des délais de plus en plus grand.

De façon similaire à la seconde expérience menée à la section 7.3.1, nous utilisons des mélanges stéréophoniques de 3 sources ayant les différences d'intensité $\theta_n \in \{\pi/12, \pi/4, 5\pi/12\}$ et les délais $\delta_n \in \{-\delta, 0, +\delta\}$. La valeur des délais $\delta \geq 0$ représente le degré d'« anéchoïsme » qui varie.

Nous évaluons les résultats en séparation de sources (en estimant les sources par masquage binaire) sur la figure 7.4 afin de pouvoir comparer les methodes anéchoïques avec DEMIX-Intantané, et nous donnons les valeurs d'erreurs RMDE d'estimation des directions pour les méthodes DUET et DEMIX-Anéchoïque sur la figure 7.5.

7.4.2 Résultats

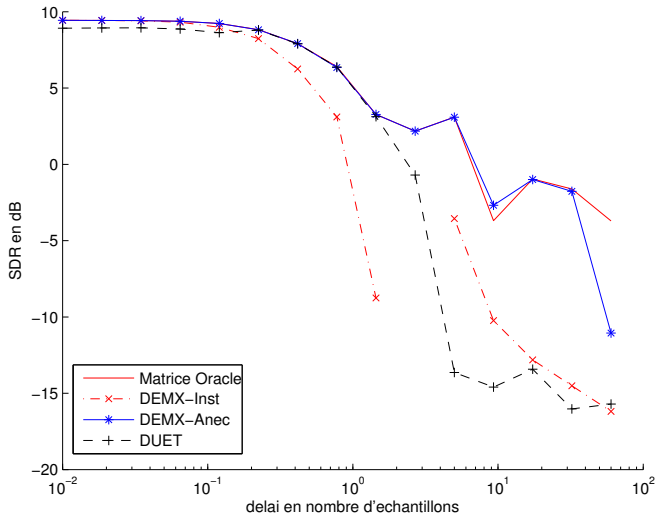


FIG. 7.4 – SDR en fonction de la valeur absolue du délai δ des deux sources latérales

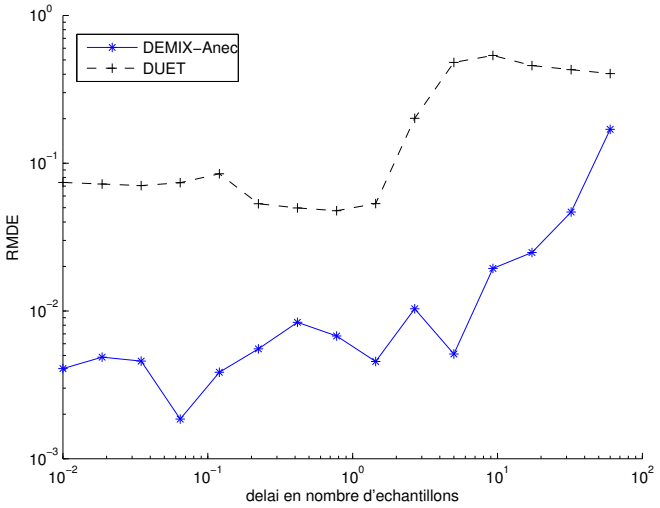


FIG. 7.5 – RMDE en fonction de la valeur absolue du délai δ des deux sources latérales

Performances en séparation à partir des directions oracles : Comme en témoigne les résultats, affichés sur la figure 7.4, obtenus en séparation à partir des directions oracles, l'estimation des sources par la technique du masquage binaire est peu performante pour estimer les sources quand le délai est grand. En effet pour des délais supérieurs égals à 10 échantillons, le SDR obtenu à partir des directions oracles est inférieur à 0 dB. Par conséquent il préférable, pour évaluer les performances des méthodes anéchoïques pour les grands délais, d'examiner les performances en RMDE, affichées sur la figure 7.5.

Performances de DEMIX-Anéchoïque : Sur la figure 7.4, on remarque que : quelle que soit la valeur du délai δ, les performances de DEMIX-Anéchoïque sont excellentes, puisque celles-ci sont sont pratiquement égales à celles obtenues à partir de la vraie matrice de mélange (matrice oracle), jusqu'à un délai de 5 échantillons, et très proches ou meilleures de celles-ci pour pour des délais compris entre 5 et 30 échantillons (75 millisecondes à 4 kHz). Puis pour un délai de 60 échantillons, le SDR de DEMIX-Anechoïque devient inférieur de 7 dB par rapport à celui obtenu avec la matrice oracle. La figure 7.5 vient confirmer que les performances d'estimation des directions anéchoïques par la méthode DEMIX-Anéchoïque sont très bonnes pour des délais inférieurs à 60 échantillons, mais deviennent supérieures à 0,1 RMDE à partir de cette valeur.

Robustesse de DEMIX-Instantané : Pour des délais très courts, DEMIX-Instantané et DEMIX-Anéchoïque obtiennent des résultats similaires en séparation (voir la figure 7.4), équivalents à ceux de l'oracle, et de 0.5 dB supérieurs à ceux de DUET. Pour un délai inférieur à 0.12 échantillon, DEMIX-Instantané est à peine de 0.26 dB en dessous des résultats de l'oracle, mais pour des délais supérieur à 1 échantillon, les

performances de DEMIX-Instantané plongent à plus de 10 dB en dessous de DEMIX-Anéchoïque. En effet, pour des délais compris entre 1.5 et 5 échantillons, les performances de DEMIX-Instantané n'ont même pas pu être calculées car l'algorithme échoue systématiquement à l'étape d'estimation du nombre de sources.

Performances de DUET : En ce qui concerne DUET, on observe que les performances, quelle soient évaluées en séparation (figure 7.4) ou en RMDE (figure 7.5), s'effondrent quand le délai excède 2 échantillons, ce qui confirme les limites théoriques de DUET pour l'estimation de délais supérieurs à 1 échantillon.

Comparaisons des performances de DEMIX-Anéchoïque et de DUET : Dans tous les cas, les performances de DUET, que ce soit en séparation ou en RMDE, restent inférieures à celles de DEMIX-Anéchoïque.

Même pour de faibles delais compris entre 10^{-2} et 10^{-1} échantillons, le SDR de DUET est de 0.5 dB inférieur à celui de DEMIX-Anéchoïque.

La figure 7.5 vient confirmer que les performances d'estimation des directions anéchoïques par la méthode DEMIX-Anéchoïque sont largement meilleures (d'un facteur supérieur à 10 en RMDE) à DUET pour des délais inférieurs à 60 échantillons, puis les performances se resserrent pour le délai de 60 échantillons.

Remarquons cependant, que les performances en RMDE de DEMIX-Anéchoïque pour un délai de 60 échantillons, ne sont que 2 à 3 fois supérieures à celles de DUET pour des délais inférieurs à 2 échantillons.

7.5 Évaluation sur des mélanges anéchoïques obtenus par simulation de scène sonore

Notre troisième expérience est une comparaison des performances de DEMIX-Anéchoïque avec DUET sur des mélanges anéchoïques obtenus par simulation d'une « salle » anéchoïque à l'aide de la toolbox RoomSim [DCB05] de MATLAB.

7.5.1 Protocole expérimental

Dans cette simulation, deux microphones cardioïdes sont placés au centre de la salle, à 20 cm l'un de l'autre, et leurs directions se croisent avec un angle de 90 degrés. Les sources sont placées sur un cercle, dont le centre est lui aussi situé au milieu de la salle, de manière à être aussi espacées que possible les unes des autres, et de façon symétrique par rapport au plan médiateur du bipoint formé par la position des deux microphones. Les sources et les microphones sont sur le même plan. Nous illustrons ceci avec l'exemple d'un mélange de 7 sources : La positions des sources et des microphones dans la salle est celle de la figure 7.6. Les directions anéchoïques correspondant à cette disposition sont présentées dans le tableau 7.7.

L'expérience consiste à estimer les performances des algorithmes en fonction d'un nombre de sources allant de $N = 2$ à $N = 7$.

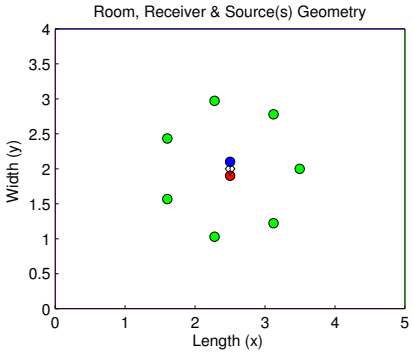


FIG. 7.6 – Configuration de la salle pour $N = 7$ sources

n	1	2	3	4	5	6	7
θ_n	0.12	0.13	0.56	0.78	1.01	1.44	1.45
δ_n	-1	-2.2	-1.8	0	+1.8	+2.2	1

FIG. 7.7 – Directions des 7 sources correspondant à la configuration de salle illustrée par la figure 7.6. La première ligne du tableau représente la différence d'intensité θ (en radians), et la seconde ligne les délais δ (en échantillons)

7.5.2 Résultats

On observe que DEMIX-Anéchoïque arrive à estimer le nombre de sources jusqu'à $N = 4$ dans plus de 9 cas sur 10 (voir tableau 7.2).

La figure 7.8 montre que, l'erreur relative d'estimation des directions en terme de RMDE moyen, de DEMIX-Anéchoïque est systématiquement inférieure à celle de DUET d'un facteur de 10, et ce quelle que soit le nombre de source du mélange.

Deux hypotèses majeures peuvent expliquer le fait que DEMIX-Anéchoïque obtienne des résultats bien meilleurs que DUET :

Premièrement, les délais supérieurs à 1 échantillon produisent des estimations ambiguës à cause du phénomène de repliement de phase décrit à la section 2.2.3. Phénomène qui disparait dans DEMIX-Anéchoïque, grâce à une méthode pour estimer les délais similaire à GCC-PHAT.

Deuxièmement, comme expliqué à la section 2.1.4.2, l'estimateur $R_{21}(t,f) = X_2(t,f)/X_1(t,f)$ de DUET et TIFROM est très instable pour les directions des sources, pour lesquelles l'intensité est présente presque uniquement sur un seul des canaux, (c.a.d. quand θ est proche de 0 ou de $\pi/2$). Contrairement à DUET, DEMIX-Anéchoïque repose sur des estimateurs de direction locale qui gardent la même

nb de sources	2	3	4	5	6	7
DEMIX-Anec	90	100	95	65	5	0

TAB. 7.2 – Pourcentage d'estimation correcte du nombre de sources

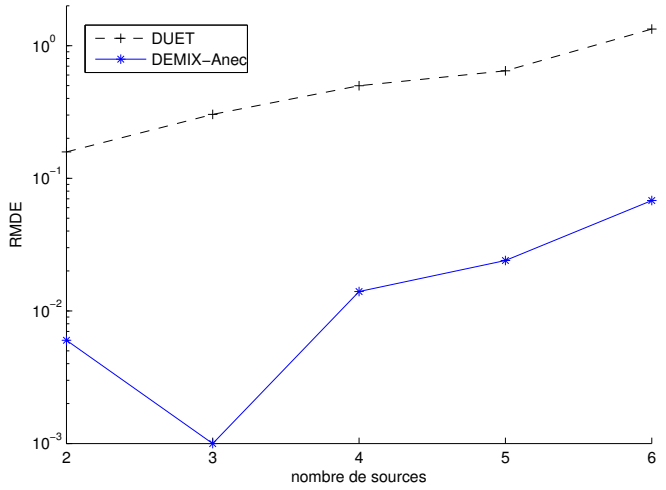


FIG. 7.8 – Erreur d'estimation des directions (en termes de RMDE) en fonction du nombre de sources pour les algorithmes DEMIX-Anéchoïque (DEMIX-Anec) et DUET

robustesse quelque soit le profil d'intensité de la direction.

7.6 Conclusion

Les expériences menées dans le cas stéréophonique montrent que DEMIX est capable de :

- compter correctement (avec un taux de succès supérieur à 90%) le nombre de sources d'un mélange anéchoïque optenu par simulation, jusqu'à 4 sources ;
- estimer avec précision (avec un $RMDE < 3 \cdot 10^{-4}$) des directions très proches, pour lesquelles les algorithmes classiques de clustering comme K-means et ELBG échouent systématiquement ($RMDE \approx 1$) ;
- estimer des directions anéchoïques ayant des délais pouvant atteindre des valeurs de 60 échantillons (pour une taille de fenêtre de 512 échantillons), alors que les performances de la méthode DUET s'effondrent pour des délais supérieurs à 2.

Les résultats de la campagne d'évaluation SISEC 2008 montrent que la méthode DEMIX est au niveau de l'état de l'art en ce qui concerne la précision de l'estimation des directions. Le point qui est sans doute le plus innovant est la capacité de la méthode DEMIX à estimer des directions proches. Il serait cependant intéressant d'évaluer sur des mélanges anéchoïques, dans le cadre de campagnes d'évaluation, DEMIX avec d'autres méthodes de l'état de l'art comme TIFROM [PD05] et TIFCORR [PD06] qui sont également capables d'estimer de grands délais.

Dans le chapitre suivant, nous allons voir que l'information spatiale des matrices de covariances locales, qui a été exploitée pour estimer les directions du mélange, peut être mise à profit conjointement à la connaissance des directions du mélange, pour

estimer localement les paramètres de modèles probabilistes de chacune des sources. Ce qui permettra d'établir une approche d'estimation des sources alternative aux approches par la parcimonie.

Chapitre 8

L'apprentissage aveugle de modèles de sources

L'estimation des sources par la parcimonie dans le plan temps-fréquence (voir chapitre 3), se fait de façon ponctuelle et *spatiale* ; ponctuelle dans le sens où chaque point temps-fréquence est traité indépendamment des autres, *spatiale* dans le sens où les contributions des sources en un point temps-fréquence sont estimées uniquement à partir de l'observation du mélange et des directions des sources.

Limites de l'information spatiale : Pour les points temps-fréquence où le nombre de sources actives est inférieur au nombre de canaux, les sources peuvent être estimées parfaitement par une approche *spatiale*, car les sources actives sont alors dans un sous-espace du mélange qui nous permet de les identifier et de les séparer par inversion de la sous-matrice correspondante. Cependant dans le cas sous-déterminé, quand le nombre de sources actives est supérieur ou égal au nombre de canaux, l'information *spatiale* n'est pas suffisante pour déterminer la solution de façon unique. Dans ce dernier cas, un critère de parcimonie permet d'estimer les sources de façon unique, cependant son utilisation n'est pas pertinente dans la mesure où l'hypothèse de parcimonie est en contradiction avec l'hypothèse que le nombre de sources actives est supérieur au nombre de canaux. Ainsi, si comme dans l'exemple d'un mélange stéréophonique à trois sources illustré par la figure 11, où les deux sources S_1 et S_3 qui sont aux extremités du mélange ont une énergie semblable, tandis qu'une troisième source S_2 placée au centre n'est pas active, alors le coefficient du mélange pointera dans une direction proche de celle de S_2, et le critère de parcimonie attribuera préférentiellement l'énergie à la source S_2.

Exploitation de l'information spectrale : Pour résoudre le problème de l'estimation des sources lorsque le nombre de sources actives est supérieur ou égal au nombre de canaux, il est aussi possible d'exploiter en plus de l'information *spatiale*, l'information *spectrale* des sources. L'information spectrale a été utilisée avec succès pour la séparation de sources audio dans le cas monophonique [Ben03, Oze06] (voir chapitre 4) où l'information spatiale est indisponible. Cependant ces approches nécessitent d'avoir

des connaissances a priori sur les sources. Or nous nous plaçons dans un cadre aveugle réaliste où nous ne disposons d'aucune information a priori sur les sources.

L'approche MGL : Nous disposons cependant d'une information supplémentaire que l'on peut mettre à profit afin d'estimer les modèles de source en aveugle. Nous proposons dans chapitre d'exploiter à nouveau l'information contenue dans les matrices de covariances locales des observations $\widehat{\mathbf{R}}_x$ définies à la section 6.1.3, qui a été utilisée par la méthode DEMIX pour estimer la mesure de fiabilité. La matrice de covariance locale des observations, conjointement à la matrice de mélange nous donne alors une information sur la distribution locale des sources. En effet, si l'on suppose que les sources sont localement (dans un voisinage temps-fréquence) gaussiennes stationnaires et indépendantes, et dans le cas d'un mélange non bruité $\mathbf{X}(t,f) = \mathbf{AS}(t,f)$, alors $\mathbf{R}_x(t,f) = \mathbf{A}\mathbf{\Sigma}(t,f)\mathbf{A}^T$, où $\mathbf{R}_x(t,f)$ est la matrice de covariance du mélange au point temps-fréquence (t,f), tandis que $\mathbf{\Sigma}(t,f)$ est la matrice de covariance des sources au point temps-fréquence (t,f). Par conséquent la matrice de covariance empirique du mélange $\widehat{\mathbf{R}}_x(t,f) = |\Omega_{t,f}|^{-1} \sum_{(\tau,\omega)\in\Omega_{t,f}} \mathbf{X}(\tau,\omega)\mathbf{X}^H(\tau,\omega)$ calculée au voisinage $\Omega_{t,f}$ du point temps-fréquence (t,f) vaut approximativement :

$$\widehat{\mathbf{R}}_x(t,f) \approx \mathbf{A}\mathbf{\Sigma}(t,f)\mathbf{A}^T \tag{8.1}$$

Sous certaines conditions que nous allons discuter à la section 8.1 il est possible d'estimer les variances locales $\mathbf{\Sigma}(t,f)$ des sources par résolution du système linéaire (8.1), et finalement d'estimer les sources par filtrage de Wiener (voir figure 8.1).

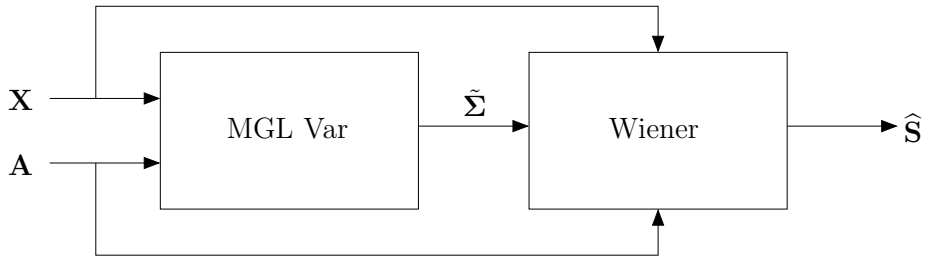


FIG. 8.1 – Schéma bloc de l'approche MGL, où à partir du mélange $\mathbf{X}$ et de la matrice de mélange $\mathbf{A}$, les variances des sources sont estimées en chaque point temps fréquence (on représente par le symbole $\tilde{\mathbf{\Sigma}}$ la valeur de ces estimées), puis les sources sont estimées par filtrage de Wiener.

Nous évaluons cette méthode à la section 8.1, que nous appelons : Modèle Gaussien Local et que nous notons `MGL`. Le modèle MGL tient compte de la structure localement stationnaire des sources, mais ne tient pas compte de la structure fréquentielle des sources, c.-à-d. du fait que les spectres des sources ont tendance à se répéter au cours du temps.

Exploitation de la structure répétitive des spectres des sources : Il peut être alors intéressant de prendre en compte la structure spectrale des sources par l'utilisation

d'un modèle MMG spectral des sources. Les méthodes de séparation de sources audio de Benaroya et Ozerov [Ben03, Oze06] s'appuient sur un modèle MMG spectral des sources qui suppose que chaque source est composée d'un ensemble de formes spectrales (voir chapitre 4). Cette modélisation se justifie amplement pour les signaux audio, au regard des résultats obtenus en séparation monophonique à partir des modèles MMG spectraux oracles (appris par un algorithme EM sur les sources originales) [Ben03, Oze06].

À propos des méthodes de séparation de sources basées sur des MMG, nous distinguons dans l'état de l'art deux approches :

1. L'utilisation de modèles MMG spectraux « connus » pour la séparation monophonique :

   L'utilisation de modèles MMG spectraux oracles dans le cas audio monophonique a été étudié par Benaroya [Ben03]. Pour chacune des sources, un algorithme EM est utilisé pour apprendre un MMG à partir de la source originale. Afin d'obtenir une estimation des variances $\sigma_n^2(t,f)$ pour chaque source et chaque point temps-fréquence, Benaroya effectue alors un décodage des états qui correspond à calculer, pour chaque trame, à partir du mélange, la combinaison d'états des MMG des sources la plus probable.

2. L'apprentissage aveugle de MMG (ponctuels) à partir du mélange : Il est théoriquement possible d'apprendre les MMG par un algorithme EM directement à partir des observations du mélange. Cependant cette approche proposée par Moulines et Attias et rappelée à la section 5.2 est probématique à cause :
   - de la complexité algorithmique qui est exponentielle en fonction du nombre de sources ;
   - des nombreux minima locaux, qui imposent d'avoir une bonne initialisation des paramètres du modèle, sous peine d'obtenir des estimations aberrantes des paramètres du modèle.

Exploitation de la connaissance sur la matrice de mélange : Si l'on dispose des paramètres du mélange $\mathbf{A}$, comme c'est le cas dans une architecture en deux étapes où l'on estime d'abord les paramètres du mélange puis les sources, deux approches sont alors possibles :

1. On peut alors utiliser l'algorithme EM de Moulines et Attias en fixant les paramètres du mélange, ce qui a pour effet de réduire le nombre de données manquantes et aussi le nombre de minima locaux, ainsi que les problèmes d'estimation des paramètres du mélange qui se posent quand le bruit est faible (voir 1.3.3). Cependant la complexité de l'algorithme reste toujours exponentielle en fonction du nombre de sources.

2. Une autre possibilité donnée par la connaissance de $\mathbf{A}$ est d'estimer les sources par une méthode *spatiale* (par ex DUET) et d'apprendre les modèles MMG à partir de ces estimations par l'algorithme EM de Benaroya. Dans ce cas, les modèles MMG des sources sont appris séparément. L'avantage de cette approche par rapport à l'approche Moulines-Attias, est qu'elle a une complexité algorithmique linéaire

en fonction du nombre de sources. Cependant l'estimation spatiale des sources n'est pas parfaite, sinon il n'y aurait pas d'intérêt à recourir à des modèles de sources pour effectuer la séparation de source. Ainsi on a peu de chance de pouvoir améliorer la séparation de sources en apprenant des modèles de sources à partir de sources d'apprentissage corrompues aux endroits où l'on aurait le plus besoin des modèles de sources.

Contributions : Nous proposons en plus de l'approche MGL, deux approches EM à complexité linéaire exploitant les informations des matrices de covariances locales, afin d'estimer les MMG spectraux des sources :

- La première se base sur l'approche EM de Moulines-Attias, et consiste à apprendre les MMG à partir de l'estimation des variances fournies par la méthode MGL ; Nous appelons `Spat-MMG` cette méthode que nous illustrons par la figure 8.2 et que nous présentons à la section 8.2.1.

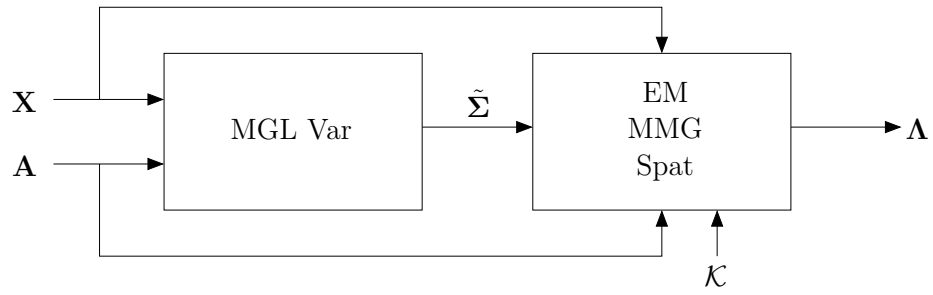


FIG. 8.2 – Schéma bloc de l'approche `Spat-MMG`, où les modèles MMG $\mathbf{\Lambda}$ des sources sont appris *en aveugle*, c.-à-d. uniquement à partir de l'observation $\mathbf{X}$ et de la matrice de mélange $\mathbf{A}$. Le bloc *MGL Var* fait une estimation $\tilde{\mathbf{\Sigma}}$ des variances des sources, en chaque point temps fréquence. Le bloc *EM MMG Spat* estime les MMG $\mathbf{\Lambda}$ à l'aide d'un algorithme d'apprentissage EM qui se fait à partir des variances $\tilde{\mathbf{\Sigma}}$ et des données $\mathbf{X}$ et $\mathbf{A}$. $\mathcal{K}$ désigne le nombre d'états des modèles MMG.

- La seconde se base sur l'approche d'adaptation de Ozerov [Oze06], elle même basée sur l'approche d'apprentissage de Benaroya, et consiste à apprendre les MMG à partir d'estimations des sources pour lesquelles on est capable de quantifier pour chaque point temps-fréquence l'erreur de ces estimations (voir figure 8.3).
  Nous proposons deux variantes de cette approche :
  1. La première se base sur une estimation des sources par la méthode DUET légèrement modifiée. Cette modification concerne l'estimation des sources non-dominante qui dans la méthode DUET sont estimées par une valeur nulle. Nous appelons `DUET-MMG` cette méthode que nous présentons à la section 8.2.2.1.
  2. La seconde se base sur la méthode MGL que nous allons présenter, et s'appelle `MGL-MMG`. Nous présentons cette méthode à la section 8.2.2.2.

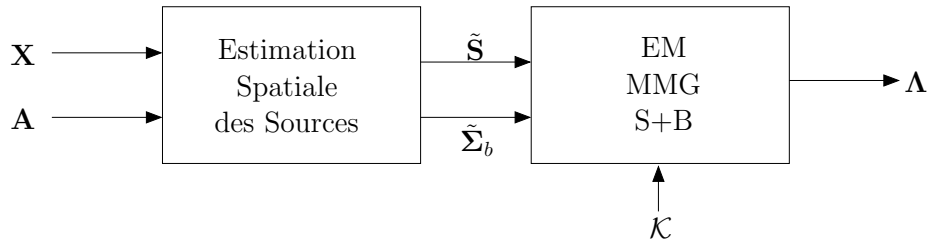


FIG. 8.3 – Schéma bloc des approches `DUET-MMG` et `MGL-MMG`, où les modèles MMG $\mathbf{\Lambda}$ des sources sont appris *en aveugle*, c.-à-d. uniquement à partir de l'observation $\mathbf{X}$ et de la matrice de mélange $\mathbf{A}$. Le bloc *Estimation spatiale des sources* fait une estimation $\tilde{\mathbf{S}}$ des sources, ainsi que de la variance $\tilde{\mathbf{\Sigma}}_b$ de l'erreur d'estimation des sources en chaque point temps fréquence. Les approches `DUET-MMG` et `MGL-MMG` se dinstinguent par une implémentation différente de ce premier bloc. Le bloc *EM MMG S+B* estime les MMG $\mathbf{\Lambda}$ à l'aide d'un algorithme d'apprentissage EM qui prend comme entrées les données $\tilde{\mathbf{S}}$ et $\tilde{\mathbf{\Sigma}}_b$. $\mathcal{K}$ désigne le nombre d'états des modèles MMG.

8.1 Le Modèle Gaussien Local (MGL)

Nous discutons dans cette section du modèle MGL, et plus particulièrement :

1. de la taille et de la forme des régions temps-fréquence sur lesquelles les variances des sources sont estimées.
2. des méthodes permettant d'estimer les variances des sources dans le cas où celles-ci sont identifiables, c.-à-d. quand le nombre d'inconnues du système linéaire de l'équation (8.1) n'est pas supérieur au nombre d'équations qui est donnée par le nombre de degrés de liberté Q de la matrice de covariance locale $\widehat{\mathbf{R}}_x$. La matrice $\widehat{\mathbf{R}}_x$ étant hermitienne, et $\mathbf{\Sigma}$ étant diagonale, nous disposons de $Q = \frac{M(M+1)}{2}$ équations linéaires pour résoudre les N inconnues σ_n^2.

   Nous pouvons remarquer que contrairement à la matrice $\mathbf{R}_x$, la matrice de covariance locale empirique $\widehat{\mathbf{R}}_x$ est à valeurs complexes car elle est calculée à partir des coefficents complexes $\mathbf{X}(t,f)$. Nous expliquons à la section C.1 de l'annexe que nous pouvons uniquement considérer la partie réelle $\Re\mathrm{e}\widehat{\mathbf{R}}_x$ de la matrice de covariance empirique. Dans les développements qui suivent, nous considérons la partie réelle de $\widehat{\mathbf{R}}_x$ mais nous gardons la notation $\widehat{\mathbf{R}}_x$ pour ne pas alourdir les notations.
3. d'une extension de la méthode, basée sur l'hypothèse de parcimonie, au où nombre de sources est supérieur à $Q = \frac{M(M+1)}{2}$.

8.1.1 Nature du modèle MGL

L'hypothèse du modèle MGL est que dans le voisinage temps-fréquence où est estimée la matrice de covariance locale, les sources sont des Gaussiennes stationnaires et indépendantes.

Compromis sur la taille du voisinage : La taille et la forme du voisinage a une importance particulière. En effet afin que l'hypothèse de stationnarité locale soit valide, on a intérêt à choisir un petit voisinage. Cependant afin que le matrice de covariance locale du mélange soit bien estimée, on a besoin d'un grand nombre d'échantillons, et donc d'un grand voisinage.

8.1.2 Les méthodes d'estimation des variances lorque $N \leq M(M+1)/2$

Nous cherchons à estimer localement les variances des sources au maximum de vraisemblance à partir de la matrice de covariance locale du mélange. Il est possible d'optimiser explicitement le critère de vraisemblance, cependant les méthodes connues (EM, Newton,...) sont itératives et particulièrement lentes. Nous proposons alors une méthode heuristique rapide qui permet d'obtenir des résultats équivalents à ceux des méthodes itératives. Nous appelons `MGL` cette méthode MGL « heuristique » que nous présentons à la section 8.1.2.2 et nous appelons `MGL Iter` la méthode « itérative » que nous présentons ci-dessous. Nous comparerons les performances de ces deux méthodes sur des mélanges stéréophoniques sous-déterminés à la section 8.3.4.1.

8.1.2.1 Méthodes itératives de maximisation de la vraisemblance

L'estimation au maximum de vraisemblance des variances $\mathbf{\Sigma}(t,f)$ des sources gaussiennes est équivalent à minimiser la divergence de Kullback-Leibler entre la distribution du mélange et celle du modèle [Car07] :

$$-\log P\left(\mathbf{X}(t,f)\,|\mathbf{\Sigma}(t,f)\right) = \mathrm{D}_{\mathrm{KL}}\left(\widehat{\mathbf{R}}_x(t,f), \mathbf{A}\mathbf{\Sigma}(t,f)\mathbf{A}^T\right) + cst \tag{8.2}$$

La divergence de Kullback-Leibler dans le cas de distributions gaussiennes a une expression explicite donnée par l'équation (A.23) que nous rappelons ci-dessous :

$$\mathrm{D}_{\mathrm{KL}}\left(\widehat{\mathbf{R}}_x, \mathbf{R}_x\right) = \frac{1}{2}\left(\mathrm{tr}\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) - \log\det\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) - M\right)$$

Il est donc possible d'estimer les variances des sources au maximum de vraisemblance par une technique d'optimisation itérative (de type Newton) basée sur le gradient et le hessien de la divergence de Kullback-Leibler. Il est aussi possible de maximiser la vraisemblance par un algorithme EM. Cependant, ces optimisations itératives doivent être conduites sur chacun des points temps-fréquence, et par conséquent nécessitent un temps de calcul très important (plusieurs heures de calcul pour un mélange de 10 secondes).

8.1.2.2 Méthode heuristique rapide :

Nous proposons une méthode directe beaucoup plus rapide que les méthodes itératives que nous venons de mentionner (quelques secondes comparées à quelques heures). Sous certaines conditions, il est possible d'estimer les variances des sources par identification du système linéaire (8.1). Pour chaque point temps-fréquence, afin de simplifier la lecture des équations, nous omettons par la suite les indices (t,f).

Soit (i_q, j_q) les $Q = M(M+1)/2$ indices des coefficients non nuls d'une matrice triangulaire supérieure de taille $M \times M$. Alors grâce à la symétrie de $\widehat{\mathbf{R}}_x$, on peut réécrire l'équation (8.1) sous la forme d'un système linéaire :

$$\hat{\mathbf{r}}_x \approx \mathbf{L}\mathbf{v} \tag{8.3}$$

où $\hat{\mathbf{r}}_x = \begin{bmatrix} \widehat{\mathbf{R}}_x(i_1, j_1) \\ \vdots \\ \widehat{\mathbf{R}}_x(i_Q, j_Q) \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_N^2 \end{bmatrix}$ et $\mathbf{L}$ est une matrice de taille $Q \times N$ telle que :

$$\mathbf{L}(q, n) = \left(\mathbf{a}_n \mathbf{a}_n^T\right)(i_q, j_q). \tag{8.4}$$

On peut ainsi estimer les variances σ_n^2 par résolution du système linéaire (8.3).

La résolution de ce système linéaire n'impose pas de contrainte sur la positivité des variances et il est par conséquent possible d'obtenir des estimations négatives des variances. Dans ces cas là, on peut alors soit estimer les variances par une méthode itérative minimisant la divergence de Kullback-Leibler, ou bien pour éviter le recours à une minimisation itérative, on peut faire l'heuristique consistant à tronquer les variances négatives à zéro. Nous évaluerons expérimentalement à la section 8.3.4.1 ces deux approches, que l'on appelle `MGL Iter` pour la version itérative et `MGL` pour la version heuristique, afin de savoir si l'on peut adopter cette heuristique sans dégrader les perfomances par rapport à l'approche itérative de maximisation de la vraisemblance.

Si le nombre d'équations est supérieur ou égal au nombre d'inconnues, c.a.d. si $M(M+1)/2 \geq N$, alors le système linéaire n'a pas forcément de solution, dans ce dernier cas il est nécessaire de recourir à une méthode d'optimisation itérative pour optimiser le critère de vraisemblance. Remarquons cependant que dans le cas stéréophonique le problème ci-dessus ne se pose pas car, soit le mélange est (sur-)déterminé et il suffit d'inverser la matrice de mélange $\mathbf{A}$ pour estimer les sources, soit le nombre de source est supérieur ou égale à $Q = 3$.

8.1.3 Extension au cas où le nombre de source est supérieur à $M(M+1)/2$

Bien que la méthode MGL proposée ne fonctionne que dans le cas où $M(M+1)/2 \geq N$, on peut toujours étendre la méthode au cas où le nombre de sources est supérieur à $M(M+1)/2$, en s'appuyant sur un critère de parcimonie sur les variances locales des sources. Le critère à minimiser devient alors :

$$J(\mathbf{v}) = \mathrm{D}_{\mathrm{KL}}\left(\widehat{\mathbf{R}}_x, \mathbf{A}\,\mathrm{diag}(\mathbf{v})\mathbf{A}^T\right) + \lambda\|\mathbf{v}\|_p \tag{8.5}$$

avec $\lambda \to 0$ et $\mathrm{diag}(\mathbf{v}) = \mathbf{\Sigma}$ une matrice diagonale à coefficients positifs et $D(.,.)$ est la divergence de Kullback entre deux distributions gaussiennes de même moyenne et dont la définition est donnée par l'équation (A.3). Le critère $J(\mathbf{v})$ ci-dessus signifie que l'on cherche le vecteur des variances $\mathbf{v}$ qui minimise la divergence de Kullback c.-à-d. qui maximise la vraisemblance. Dans le cas où il y a plusieurs solutions à ce critère de vraisemblance, alors celle qui a la plus petite norme l_p est la solution du critère $J(\mathbf{v})$.

8.1.3.1 Implémentation algorithmique dans le cas stéréophonique

Nous supposons que nous sommes dans le cas stéréophonique où nous disposons de $M = 2$ canaux. Afin de trouver une solution algorithmique au problème d'optimisation de l'équation (8.5), nous faisons l'hypothèse qu'il y a un maximum de $M(M+1)/2$ sources actives en chaque voisinage temps-fréquence.

On note alors par $\mathcal{J}$ le sous-ensemble, appelé *pattern d'activité*, des $M(M+1)/2$ sources qui ont une variance non nulle. L'algorithme proposé est le suivant :

Pour chaque point temps-fréquence (t, f) :

1. Calculer la matrice de covariance empirique $\widehat{\mathbf{R}}_x$;
2. Calculer le vecteur des variances $\mathbf{v}_\mathcal{J} = \mathbf{L}_\mathcal{J}^{-1}\hat{\mathbf{r}}_x$ pour chaque *pattern d'activité* $\mathcal{J}$; où : $\mathbf{L}_\mathcal{J} = [\mathbf{l}_j]_{j\in\mathcal{J}}$ et $\mathbf{l}_j$ étant les vecteurs colonnes de la matrice $\mathbf{L} = [\mathbf{l}_j]_{j=1}^N$.
   - Si il existe un ou plusieurs *patterns d'activité* tels que $\mathcal{J} \in \mathcal{P}$, où $\mathcal{P}$ désigne l'ensemble des *patterns d'activité* pour qui le vecteur $\mathbf{v}_\mathcal{J}$ est à composantes positives :
     sélectionner le pattern $\mathcal{J} \in \mathcal{P}$ qui minimise la norme l_p.
   - Sinon tronquer à zéro les valeurs des coefficients $v_{\mathcal{J},j}$ des vecteurs $\mathbf{v}_\mathcal{J}$ qui sont négatives :
     $\mathbf{v}_\mathcal{J} \leftarrow \text{tronc}(\mathbf{v}_\mathcal{J})$, où $\text{tronc}(\mathbf{v}_\mathcal{J}) = [\text{tronc}(v_{\mathcal{J},j})]_j$ et si x est un scalaire :
     $$\text{tronc}(x) = \begin{cases} x & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases} .$$
     Sélectionner le vecteur $\mathbf{v}_\mathcal{J}$ tronqué qui minimise la divergence de Kullback $\text{D}_{\text{KL}}\left(\widehat{\mathbf{R}}_x, \mathbf{A}\,\text{diag}(\mathbf{v}_\mathcal{J})\mathbf{A}^T\right)$;
3. Estimer les sources par filtrage de Wiener à partir de la matrice de covariance $\tilde{\mathbf{\Sigma}} = \text{diag}(\mathbf{v}_\mathcal{J})$ pour le pattern $\mathcal{J}$ qui a été sélectionné :
$$\tilde{\mathbf{S}} = \tilde{\mathbf{\Sigma}}\,\mathbf{A}^T\left(\mathbf{A}\,\tilde{\mathbf{\Sigma}}\,\mathbf{A}^T\right)^{-1}\mathbf{X}$$

8.2 Estimation des MMG spectraux à partir du mélange

Afin de tenir compte de la structure spectrale des sources, et aussi pour avoir des estimations des variances se basant sur un nombre d'échantillons plus important que pour le modèle MGL, nous proposons de modéliser les sources par des MMG spectraux comme au chapitre 4. Remarquons tout d'abord que la réalisation d'une variable aléatoire MMG peut être interprété comme une Gaussienne associée à un état (caché) particulier, et que par conséquent l'hypothèse que les sources soient localement gaussiennes revient à faire l'hypothèse que les sources sont chacune dans un état, qu'il s'agit de décoder, qui demeure constant sur toute la région temps-fréquence concernée.

Nous pouvons ainsi nous baser sur l'estimation locale des variances du modèle MGL expliquée à la section 8.1, afin d'inférer les paramètres des MMG des sources. Nous proposons trois approches permettant l'apprentissage des paramètres des MMG spectraux *en aveugle*, c.-à-d. sans faire appel à d'autres connaissances que le mélange $\mathbf{X}$ et la matrice de mélange $\mathbf{A}$.

8.2.1 Apprentissage des MMG dans l'espace du mélange

Les modèles MMG spectraux des sources peuvent être appris au sens du maximum de vraisemblance en considérant le modèle de mélange $\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f)$. On peut alors utiliser une extension de l'algorithme EM de Moulines et Attias (voir section 1.2) pour des MMG spectraux. L'algorithme EM d'estimation des MMG spectraux, connaissant la matrice de mélange, est donné à la section B.3 de l'annexe.

Principe de l'approche : Nous appelons `Spat-MMG` cette approche que nous illustrons par la figure 8.2. Ici, en plus de fixer les paramètres du mélange, c.-à-d. la matrice de mélange $\mathbf{A}$ qui est connue, nous fixons les variances $\tilde{\mathbf{\Sigma}}$ de l'ensemble des sources, qui ont été estimées par la méthode MGL, exceptée celle de la source S_n pour laquelle on souhaite apprendre le MMG. Autrement dit, la méthode d'apprentissage du MMG de la source S_n est celle qui est décrite par l'algorithme 3 de la section B.3, en remplaçant la matrice $\mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) = \mathrm{diag}\left(\left[\sigma^2_{k_1}(f), \cdots, \sigma^2_{k_N}(f)\right]\right)$ par la matrice $\mathbf{\Sigma}_{n,k}^{(l)}(t,f) = \mathrm{diag}\left(\left[\tilde{\sigma}^2_1(t,f), \cdots, \tilde{\sigma}^2_{n-1}(t,f), \sigma^2_{n,k}(f), \tilde{\sigma}^2_{n+1}(t,f), \cdots, \tilde{\sigma}^2_N(t,f)\right]\right)$, où $\tilde{\sigma}^2_k(t,f)$ est la variance de la source S_k estimée par la méthode MGL au point temps-fréquence (t,f), et $\sigma^2_{n,k}(f)$ est la variance de l'état k du modèle MMG de la source n que l'on cherche à estimer.

Complexité calculatoire : On peut ainsi estimer les modèles de chacune des sources alternativement, avec une complexité linéaire en fonction du nombre de sources. Il faut cependant noter que le nombre de matrices $\mathbf{A}\,\mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f)\mathbf{A}^T$ à inverser était de $\mathrm{T\,F}\,\mathcal{K}^N$ dans l'algorithme EM initial de la section B.3, alors que maintenant il faut inverser pour chaque sources les $\mathrm{T\,F}\,\mathcal{K}$ matrices $\mathbf{A}\,\mathbf{\Sigma}_{n,k}^{(l)}(t,f)\mathbf{A}^T$, où T est le nombre de trames de la TFCT du mélange, F le nombre de fréquences, et $\mathcal{K}$ le nombre d'états des MMG de chaque source. Le nombre total de matrices à inverser pour l'apprentissage des N sources est donc de $\mathrm{T\,F}\,\mathcal{K}\,N$.

Comme on le verra dans la partie expérimentale (section 8.3.3.2), on atteind vite les limites en terme de capacité mémoire des machines quand le nombre d'états $\mathcal{K}$ augmente. Par la suite nous désignerons par `Spat-MMG` cette méthode d'apprentissage de MMG.

Limites de l'approche : Cette méthode d'estimation des MMG repose sur le modèle MGL, et donc sur l'hypothèse que les sources sont localement gaussiennes et indépendantes. Les variances locales sont estimées à partir de l'ensemble des points d'un voisinage, ce qui constitue un échantillon de petite taille. Par conséquent, les estimations des variances seront probablement entachées d'une erreur, non prise en compte par la méthode d'apprentissage des MMG.

8.2.2 Apprentissage des MMG spectraux à partir d'un modèle source + bruit

Une seconde approche illustré par la figure 8.3 consiste à considérer une estimation *spatiale* des sources, et d'exploiter l'information de la matrice de covariance locale des observations afin d'établir une mesure de la qualité de l'estimation *spatiale* des sources. On considère ainsi le modèle suivant où l'estimation *spatiale* $\tilde{S}_n(t,f)$ de la source n au point temps-fréquence (t,f) est une observation bruitée de la source $S_n(t,f)$:

$$\tilde{S}_n(t,f) \approx S_n(t,f) + B_n(t,f) \tag{8.6}$$

On cherche ainsi à estimer les paramètres du MMG de la source S_n à partir des observations $\tilde{S}_n$ données par l'estimation *spatiale* de la source et la distribution du bruit B_n. Le bruit est supposé gaussien et de moyenne nulle. En plus de l'estimation de la source $\tilde{S}_n$, il s'agit alors d'estimer la variance du bruit $\tilde{\sigma}^2_{b,n}$ qui peut être vue comme la variance de l'erreur d'estimation *spatiale* de la source. Celle-ci peut être estimée grâce aux informations conjointes de la matrice de mélange $\mathbf{A}$ et de la matrice de covariance locale des observations $\widehat{\mathbf{R}}_x$.

Une fois les valeurs de $\tilde{S}_n(t,f)$ et de $\tilde{\sigma}^2_{b,n}(t,f)$ estimées, on peut apprendre les paramètres du modèle MMG de la source S_n au maximum de vraisemblance, grâce à l'algorithme EM. Celui-ci découle naturellement de l'algorithme EM pour les MMG détaillé à la section 5.2.2, en considérant des MMG spectraux, une matrice de mélange de taille 1×1 ($\mathbf{A} = [1]$) fixe, ainsi qu'un bruit de variance $\tilde{\sigma}^2_{b,n}$ fixée.

L'algorithme EM résultant, détaillé à la section B.2 de l'annexe, est assez proche de l'algorithme d'adaptation acoustique d'Ozerov [Oze06] appliqué à la séparation voix-/musique d'un enregistrement monophonique. Ozerov apprend un modèle MMG de musique au maximum de vraisemblance tandis que le modèle de voix, lui aussi MMG, est fixé. Cela correspond à une interprétation « dégénérée » du critère MAP, où l'on considère que la loi a priori sur les paramètres du MMG de la musique est une loi uniforme, tandis que la loi a priori sur les paramètres du MMG de la voix est une distribution de Dirac. Dans notre situation, le modèle de bruit qui d'ailleurs n'est pas un MMG mais un MGL est fixé, tandis que l'on apprend le modèle MMG de la source S_n au maximum de vraisemblance.

Nous proposons dans les sections 8.2.2.1 et 8.2.2.2 deux approches différentes pour estimer les paramètres $\left\{\tilde{S}_n, \tilde{\sigma}^2_{b,n}\right\}_{(t,f)}$ du modèle source + bruit.

8.2.2.1 Estimations des paramètres du modèle source + bruit : l'approche DUET-MMG

Une première approche consiste à reprendre le modèle de mélange local utilisé par DEMIX. Le modèle de mélange local de DEMIX détaillé à la section 6.2, contrairement au modèle MGL, ne tient pas compte de la connaissance sur la matrice de mélange. Il considère qu'au voisinage (t,f), une source S_n gaussienne domine les autres. Ces dernières étant modélisées par un bruit gaussien isotrope :

$$\mathbf{X}(t,f) \approx \mathbf{a}_{n(t,f)} S_{n(t,f)}(t,f) + \mathbf{B}(t,f) \tag{8.7}$$

Par la suite, pour faciliter la lisibilité, nous omettons les indices (t,f).

Estimation de la variance du bruit : La densité de probablilité du bruit est $\mathbf{B} \sim N_c(\mathbf{B}; \mathbf{0}, \tilde{\sigma}_b^2 \mathbf{I}_M)$, où $N_c(.;.,.)$ est la loi gaussienne circulaire définie à l'équation (B.2). Suivant un tel modèle, les $M-1$ plus petites valeurs propres de la matrice de covariance locale $\mathbf{R}_x$ sont alors égales à $\tilde{\sigma}_b^2$. Autrement dit, dans le cas stéréophonique, on peut estimer la valeur de $\tilde{\sigma}_b^2 = \lambda_2$, où $\lambda_1 > \lambda_2$ sont les valeurs propres de $\mathbf{R}_x$, par la plus petite valeur propre $\widehat{\lambda}_2$ de $\widehat{\mathbf{R}}_x$: $\tilde{\sigma}_b^2 \approx \widehat{\lambda}_2$.

Estimation de la source dominante : Pour estimer S_n, nous proposons de projeter l'observation du mélange $\mathbf{X}$ sur la direction $\mathbf{a}_n$. Cela correspond à l'estimation de DUET si l'on suppose que la direction la plus proche du vecteur d'observation $\mathbf{X}$ est effectivement la direction de la source S_n. L'estimation de la source S_n par DUET peut alors s'écrire :

$$\tilde{S}_n = \mathbf{a}_n^H \mathbf{X} \approx S_n + B_n \tag{8.8}$$

avec $B_n \sim N_c(B_n; 0, \tilde{\sigma}_b^2)$

Pour l'estimation de la variance de l'erreur d'estimation de la source dominante S_n, nous prenons la valeur $\tilde{\sigma}_b^2 = \widehat{\lambda}_2$.

Estimation des sources non dominantes : Pour l'estimation des autres sources, on propose l'heuristique suivante :

$$\tilde{S}_{i\neq n}(t,f) = \sqrt{\frac{M}{N-1}\left(\|\mathbf{X}(t,f)\|^2 - \|\tilde{S}_n(t,f)\|^2\right)} \tag{8.9}$$

On peut donner l'interprétation suivante de cette heuristique :

Comme $\|\sum_{i\neq n} \mathbf{a}_i S_i\|^2 = \|\mathbf{X} - \mathbf{a}_n S_n\|^2$, et comme $\tilde{S}_n = \mathbf{a}_n^H \mathbf{X}$, d'après Pythagore $\|\mathbf{X} - \mathbf{a}_n \tilde{S}_n\|^2 = \|\mathbf{X}\|^2 - \|\tilde{S}_n\|^2$. Supposons maintenant que les S_i soient des réalisations de variables aléatoires iid gaussiennes de moyenne nulle et de variance σ^2, et que $\mathbf{a}_{i\neq n}$ soient des réalisations de vecteurs aléatoires iid tel que $\|\mathbf{a}_i\|^2 = 1$ et $\mathbb{E}\{a_{m,i}^2\} = 1/M, \forall m$. Par conséquent, $\mathbb{E}\{\sum_{i\neq n} a_{m,i}^2\} = \frac{N-1}{M}$. Or, le modèle local suppose que la contribution des sources non dominantes soit la réalisation d'une Gaussienne de moyenne nulle et de covariance isotrope : c.a.d. $\mathbb{E}\{\sum_{i\neq n} \mathbf{a}_i \mathbf{a}_i^T \sigma^2\} = \sigma^2 \mathbb{E}\{\sum_{i\neq n} \mathbf{a}_i \mathbf{a}_i^T\} = \sigma^2 \cdot \alpha \cdot \mathbf{I}_M$. Par conséquent $\mathbb{E}\{\sum_{i\neq n} a_{m,i} \cdot a_{l,i}\} = \alpha \delta[m,l] = \frac{N-1}{M}\delta[m,l]$ et donc $\mathbb{E}\{\sum_{i\neq n} \mathbf{a}_i \mathbf{a}_i^T\} = \frac{N-1}{M}\mathbf{I}_M$ et $\mathbb{E}\{\|\sum_{i\neq n} \mathbf{a}_i S_i\|^2\} = \frac{N-1}{M}\sigma^2$. Par conséquent : $\sigma^2 \approx \frac{M}{N-1}\left(\|\mathbf{X}\|^2 - \|\tilde{S}_n\|^2\right)$. Notre heuristique consiste à prendre la racine carrée de σ^2 comme estimation des sources non dominantes.

Pour l'estimation de la variance de l'erreur d'estimation de la source $S_{i\neq n}$, nous prenons la valeur $\tilde{\sigma}_b^2 = \widehat{\lambda}_2$ comme pour la source dominante.

Résumé : Soit n l'indice de la source non nulle estimée par la méthode DUET au point temps-fréquence (t, f). Les estimateurs de l'approche DUET-MMG sont finalement :

$$\tilde{S}_n(t,f) = \mathbf{a}_n^H \mathbf{X}(t,f) \tag{8.10}$$

$$\tilde{S}_{i\neq n}(t,f) = \sqrt{\frac{M}{N-1}\left(\|\mathbf{X}(t,f)\|^2 - \|\tilde{S}_n(t,f)\|^2\right)} \tag{8.11}$$

$$\tilde{\sigma}^2_{b,k}(t,f) = \widehat{\lambda}_2(t,f), \forall k \tag{8.12}$$

où $\widehat{\lambda}_2(t,f)$ est la seconde valeur propre de la matrice de covariance empirique $\widehat{\mathbf{R}}_x(t,f)$. Par la suite nous désignerons par `DUET-MMG` cette méthode d'apprentissage de MMG.

8.2.2.2 Estimations des paramètres du modèle source + bruit : l'approche MGL-MMG

Une autre approche consiste à faire l'estimation spatiale locale des sources $\tilde{S}_n(t,f)$ par la méthode MGL (estimation de $\tilde{\mathbf{\Sigma}}(t,f)$, puis filtrage de Wiener). Afin de simplifier la lecture des équations, nous omettons par la suite les indices (t,f). L'estimation de la variance du bruit $\tilde{\sigma}^2_{b,n}$ est obtenue à partir la formule de la covariance $\tilde{\mathbf{C}} = \mathbb{E}\left[(\tilde{\mathbf{S}} - \mathbf{S})(\tilde{\mathbf{S}} - \mathbf{S})^H \middle| \mathbf{X}\right]$ des sources gaussiennes conditionnellement à l'observation du mélange qui est donnée par la formule (A.27) dans le cas scalaire, et qui peut s'écrire :

$$\tilde{\mathbf{C}} = \tilde{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\,\mathbf{A}^T \left(\mathbf{A}\,\tilde{\mathbf{\Sigma}}\,\mathbf{A}^T\right)^{-1} \mathbf{A}\,\tilde{\mathbf{\Sigma}}\,. \tag{8.13}$$

Comme nous souhaitons apprendre les MMG des sources séparément les uns des autres, nous ne nous intéressons qu'aux probabilités marginales des sources. En particulier nous ne nous intéressons pas aux covariances entre sources. Par conséquent la variance du bruit correspondant à une source S_n est obtenue à partir de la valeur $\tilde{\mathbf{C}}_{n,n}$ de la matrice $\tilde{\mathbf{C}}$ de covariance de $\mathbf{S}$ conditionnellement à $\mathbf{X}$:

$$\tilde{\sigma}^2_{b,n} = \mathbb{E}\left[\left|\tilde{S}_n - S_n\right|^2 \middle| \mathbf{X}\right] = \tilde{\mathbf{C}}_{n,n}\,. \tag{8.14}$$

Pour résumer, nous avons donc les estimateurs suivants :

$$\tilde{\mathbf{S}} = \tilde{\mathbf{W}}\,\mathbf{X} \tag{8.15}$$

$$\tilde{\mathbf{W}} = \tilde{\mathbf{\Sigma}}\,\mathbf{A}^T \left(\mathbf{A}\,\tilde{\mathbf{\Sigma}}\,\mathbf{A}^T\right)^{-1} \tag{8.16}$$

$$\tilde{\sigma}^2_{b,n} = \tilde{\mathbf{C}}_{n,n} \tag{8.17}$$

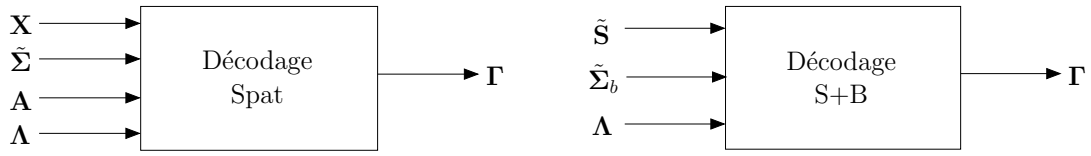
et $\tilde{\mathbf{\Sigma}}$ est obtenue par la méthode la méthode MGL expliquée à la section 8.1, et $\tilde{\mathbf{C}}$ est donnée par la formule (8.13).

Par la suite nous désignerons par `MGL-MMG` cette méthode d'apprentissage de MMG.

8.2.3 Décodage aveugle des états du MMG

Etant donné que les méthodes d'apprentissage de MMG présentées dans cette section prennent comme entrée le mélange, à partir duquel on souhaite séparer les sources, il est possible de décoder les états du MMG lors de l'étape d'apprentissage EM. En effet le décodage des états consiste à calculer les probabilités $\mathbf{\Gamma} = [\gamma_{\mathbf{k}}]_{\mathbf{k}}$ des états, or celles-ci sont calculées à chaque étape E de l'algorithme EM à partir des paramètres du MMG estimés à l'étape précédente. L'approche que nous proposons consiste alors ré-itérer une étape E de l'algorithme EM après l'apprentissage afin de calculer les valeurs $\gamma_{\mathbf{k}}^{(L+1)}$, où L désigne le nombre d'itérations effectuées par l'algorithme EM. Cette approche dépend par conséquent de la méthode d'apprentissage EM utilisée. Nous donnons sur la figure 8.4 les schémas blocs des modules de décodage correspondant aux deux approches EM que nous avons présenté dans cette section.

L'intérêt principale de cette approche de décodage aveugle est que contrairement au décodage sur le mélange décrit au chapitre 4, la **complexité calculatoire est linéaire** en fonction du nombre de sources.



(a) Décodage Spat pour la méthode `Spat-MMG`. $\tilde{\mathbf{\Sigma}}$ est la matrice de variance des sources estimée par l'approche MGL.

(b) Décodage S+B pour les méthodes `DUET-MMG` et `MGL-MMG`. $\tilde{\mathbf{S}}$ est l'ensemble des sources $\tilde{S}_n$ et $\tilde{\mathbf{\Sigma}}_b$ est l'ensembles des variances d'erreur $\tilde{\sigma}_{b,n}$.

FIG. 8.4 – Schéma bloc des méthodes de décodage aveugle de MMG à complexité linéaire que nous proposons. La sortie $\mathbf{\Gamma}$ du module de décodage est l'ensemble des estimations des probabilités des états des modèles MMG spectraux des différentes sources.

8.3 Résultats expérimentaux

Nous proposons dans cette partie d'évaluer les méthodes d'estimation des sources introduites dans ce chapitre.

8.3.1 Choix des mélanges test

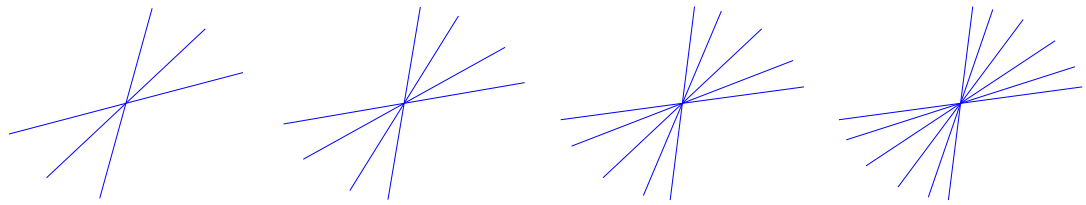
Afin de tester les algorithmes d'estimation des sources dans différents contextes, nous évaluons l'ensemble des algorithmes sur un ensemble de mélanges ayant des propriétés différentes. En particulier, nous souhaitons d'une part avoir des mélanges ayant différents nombres de sources dans le cas sous-déterminé, et d'autre part avoir des mélanges composés de sources musicales et d'autres composés de sources de parole. De plus, afin de pouvoir comparer les résultats en séparation avec d'autres méthodes, nous

souhaitons utiliser des mélanges standards qui ont déjà été utilisés auparavant. Notre choix est d'utiliser les signaux de parole et de musique utilisés dans l'article [VGP07].

Les enregistrements musicaux sont constitués de 10 ensembles de 3 sources synchronisés et en harmonie, tandis que les enregistrements de parole sont constitués de 10 ensembles de 3 sources issues de lectures de livres en langue anglaise. Les 3 premiers ensembles de sources de parole sont constitués de locuteurs masculins, les 3 suivants de locuteurs féminins et les 4 derniers sont mixtes.

La durée de chaque mélange est de 11 secondes, la fréquence d'échantillonnage des signaux musicaux est de $22,05$ kHz tandis que celle des signaux de parole n'est que de 8 kHz, car les signaux de parole ont une bande passante plus réduite que les signaux musicaux. Dans toutes les experiences que nous évaluons dans cette section, nous utilisons des fenêtres sinusoïdales de 512 échantillons pour la TFCT calculée pour les signaux échantillonés à 8 kHz, ce qui correspond à une fenêtre de 64 ms, et des fenêtres sinusoïdales de 2048 échantillons pour les signaux échantillonés à $22,05$ kHz, ce qui correspond à une fenêtre d'environs 93 ms. Dans tous les cas le recouvrement des fenêtres est d'un facteur $1/2$.

Nous faisons varier le nombre de sources N de 3 à 6, et pour chaque N, une matrice de mélange est construite comme dans l'article [VM01], avec un angle de $50-5N$ degrées entre sources successives. Une représentation graphique de ces 4 matrices de mélanges est présentée à la figure 8.5.



(a) 3 sources (b) 4 sources (c) 5 sources (d) 6 sources

FIG. 8.5 – Les quatre matrices de mélanges instantanées utilisées dans nos expériences

Nous les appliquons à chacun des 20 ensembles de sources (10 de parole et 10 de musique) que nous venons de décrire. Il en résulte un total de $4 \times 20 = 80$ mélanges classés en 8 groupes de 10 mélanges, 4 de musique, et 4 de parole.

Les résultats que nous allons présenter par la suite seront à chaque fois les moyennes en SDR (voir [GBVF03]) calculées sur les $N \times 10$ sources de chaque groupe de 10 mélanges. Rappelons que nous évaluons uniquement l'estimation des sources, et par conséquent la matrice de mélange est supposée connue pour l'ensemble des méthodes.

8.3.2 Évaluation des voisinages optimaux

Intention expérimentale : La méthode MGL ainsi que les méthodes d'apprentissage de MMG que nous proposons dans ce chapitre, s'appuient sur l'analyse de la matrice de covariance locale calculée sur les voisinages des points temps-fréquence. Nous

souhaitons évaluer la taille et la forme optimale des voisinages temps-fréquence pour les différentes méthodes d'estimations des sources présentées dans ce chapitre.

8.3.2.1 Voisinages de la méthode MGL

La méthode MGL, contrairement aux méthodes qui apprennent des MMG, est une méthode rapide qui exploite directement les matrices de covariance locales afin d'estimer les variances des sources utilisées lors de la séparation par filtrage de Wiener. Nous proposons donc une première expérience ayant pour but de déterminer la taille et la forme du voisinage qui permet d'obtenir les meilleurs résultats avec la méthode MGL.

Ainsi, nous souhaitons évaluer les performances en séparation de la méthode MGL, pour différentes tailles de voisinage temps-fréquence ($\mathrm{L_T} \times \mathrm{L_F}$) :

$$\mathcal{V} = \{1 \times 1, 1 \times 3, 1 \times 5, 3 \times 1, 3 \times 3, 3 \times 5, 5 \times 1, 5 \times 3, 5 \times 5\} \tag{8.18}$$

Le premier paramètre $\mathrm{L_T}$ représente la largeur du voisinage en trames, tandis que le deuxième paramètre $\mathrm{L_F}$ représente la hauteur en nombre de bins fréquentiels.

La méthode MGL estime les variances en chaque point temps-fréquence (t, f). Il parait alors intuitif de donner plus de poids aux points du voisinage de (t, f) qui sont proches du point (t, f). Pour cette raison, nous comparons les performances obtenues en séparation pour une fenêtre de pondération de Hanning et pour une fenêtre de pondération rectangulaire. Cette dernière correspond à la matrice de covariance obtenue avec un poids identique pour l'ensemble des points du voisinage :

$$\widehat{\mathbf{R}}_x(\Omega_{t,f}) = \frac{1}{|\Omega_{t,f}|} \sum_{(\tau,\omega)\in\Omega_{t,f}} \mathbf{X}(\tau,\omega)\mathbf{X}^H(\tau,\omega) \tag{8.19}$$

tandis que l'utilisation d'une fenêtre de Hanning $h_L(t)$ de taille L correspond au calcul de la matrice de covariance :

$$\widehat{\mathbf{R}}_x(\Omega_{t,f}) = \frac{1}{\sum\limits_{(\tau,\omega)\in\Omega_{t,f}} h_{\mathrm{L_T}}(t-\tau)h_{\mathrm{L_F}}(f-\omega)} \sum_{(\tau,\omega)\in\Omega_{t,f}} h_{\mathrm{L_T}}(t-\tau)h_{\mathrm{L_F}}(f-\omega)\mathbf{X}(\tau,\omega)\mathbf{X}^H(\tau,\omega) \tag{8.20}$$

A titre d'exemple, nous donnons ci-dessous les poids obtenus pour les voisinages $\{3 \times 1\}$ et $\{3 \times 3\}$ avec fenêtrage de Hanning :

(a) fenêtrage de Hanning $\{3 \times 1\}$

0.5
1
0.5

(b) fenêtrage de Hanning $\{3 \times 3\}$

0.25	0.5	0.25
0.5	1	0.5
0.25	0.5	0.25

Les résultats (en terme de SDR) de cette expérience sont présentés dans le tableau 8.1. Dans le cas des signaux musicaux, les résultats obtenus sur les voisinages $\mathcal{V}$, suggérant d'utiliser des voisinages ayant un nombre de trames plus important que le nombre

de bins fréquentiels, nous rajoutons à $\mathcal{V}$ les voisinages $\{9 \times 1, 9 \times 3, 9 \times 5\}$ avec fenêtrage de Hanning.

Analyse des réultats : Nous remarquons premièrement que les résultats obtenus varient très peu en fonction du type de voisinage.

Dans le cas des signaux de parole, le type de voisinage qui permet d'obtenir les meilleurs résultats en séparation est le voisinage 3×3 avec une fenêtre de Hanning, et ce quelle que soit le nombre de sources du mélange. La différence entre l'utilisation d'un fenêtrage de Hanning ou rectangulaire pour la taille de voisinage 3×3, n'est que de 0.2 dB au maximum.

Pour les signaux de musique, cependant, le type de voisinage optimal varie selon le nombre de sources du mélange. En effet, les voisinages qui permettent d'obtenir, pour les signaux de musique, les meilleurs résultats en séparation sont :

- pour 3 sources : le voisinage 9×3 avec une fenêtre de Hanning ;
- pour 4 sources : le voisinage 9×1 avec une fenêtre de Hanning ;
- pour 5 et 6 sources : le voisinage 5×1 avec une fenêtre de Hanning ;

Cependant les différences entre ces voisinages sont très faibles : Entre les voisinages $9 \times 1H$ et $5 \times 1H$ la différence maximale est de 0.1 dB, et entre $9 \times 1H$ et $9 \times 3H$ elle est de 0.2 dB.

Interprétation des résultats :

1. Dans tous le cas, le fenêtrage de Hanning est meilleur que le fenétrage rectangulaire. Le principe qui consiste à attribuer un poids plus important aux points d'une region temps-fréquence $\Omega_{t,f}$ qui sont proches du point (t, f), est donc validé expérimentalement ; Cependant la différence de performance entre ces deux types de fenêtre est relativement faible et au plus de 0.2 dB en SDR.
2. Pour les signaux de musique, plus il y a de sources, plus le voisinage optimal est petit et vice versa. Comme on l'a vu dans la section 8.1.1, la taille du voisinage est un compromis entre : avoir un grand voisinage afin que le nombre d'échantillons soit grand et que l'hypothèse d'indépendance des sources soit valide, et avoir un petit voisinage afin que les hypothèses de parcimonie et de stationarité locale soient valides. Les résultats expérimentaux suggèrent que plus le nombre de sources devient grand, moins les hypothèses de parcimonie et de stationarité locales sont valides, et qu'il est par conséquent conseillé de choisir des voisinages plus petits quand le nombre de sources augmente.
3. Pour les signaux de musique, le voisinage optimal contient un plus grand nombre de trames que de bins fréquentiels. Cela suggère une continuité temporelle plus importante pour les signaux de musique que pour les signaux de parole.

8.3.2.2 Voisinages des méthodes MMG

Protocole expérimental : Afin de déterminer expérimentalement la forme optimale des voisinages des méthodes d'apprentissage de MMG décrites aux sections 8.2.1 et

voisinage MGL	parole				musique			
nbr de sources →	3	4	5	6	3	4	5	6
$1 \times 1R$	12.7	7.9	4.8	2.7	14.0	9.4	6.7	4.4
$1 \times 3R$	13.1	8.0	4.8	2.7	14.0	9.3	6.4	4.1
$1 \times 5R$	12.8	7.6	4.2	2.0	14.0	9.0	5.8	3.7
$3 \times 1R$	13.0	7.9	4.8	2.6	14.3	9.8	7.1	4.7
$3 \times 3R$	13.5	8.2	4.9	2.7	14.3	9.6	6.7	4.4
$3 \times 5R$	13.3	7.8	4.3	2.1	14.3	9.3	6.1	4.0
$5 \times 1R$	12.8	7.6	4.4	2.2	14.4	9.8	7.0	4.5
$5 \times 3R$	13.2	7.8	4.4	2.2	14.4	9.6	6.7	4.2
$5 \times 5R$	13.0	7.3	3.8	1.6	14.4	9.3	6.0	3.8
$1 \times 1H$	12.8	7.9	4.8	2.7	14.0	9.4	6.7	4.4
$1 \times 3H$	13.1	8.1	4.9	2.8	14.1	9.4	6.5	4.3
$1 \times 5H$	13.1	8.0	4.7	2.5	14.1	9.3	6.3	4.1
$3 \times 1H$	13.1	8.1	4.9	2.8	14.3	9.8	7.1	4.7
$3 \times 3H$	**13.5**	**8.4**	**5.1**	**2.9**	14.3	9.7	6.9	4.6
$3 \times 5H$	13.5	8.2	4.8	2.7	14.3	9.6	6.6	4.4
$5 \times 1H$	13.1	8.0	4.8	2.6	14.4	9.8	**7.2**	**4.7**
$5 \times 3H$	13.5	8.2	4.9	2.7	14.5	9.8	7.0	4.6
$5 \times 5H$	13.5	8.1	4.6	2.6	14.4	9.6	6.7	4.3
$9 \times 1H$					14.5	**9.9**	7.1	4.6
$9 \times 3H$					**14.5**	9.8	6.9	4.4
$9 \times 5H$					14.5	9.6	6.5	4.2

TAB. 8.1 – Evaluation des performances en séparation (en terme de SDR en dB) de la méthode MGL, pour différents voisinages $L_T \times L_F$, où L_T est la largeur du voisinage en trames, tandis que le paramètre L_F est la hauteur en nombre de bins fréquentiels du voisinage. Le type de fenétrage appliqué au voisinage est indiqué par la lettre majuscule qui suit : « R » pour une fenêtre rectangulaire et « H » pour une fenêtre de Hanning. Les meilleurs résultats de chaque groupe de mélanges sont écrits en caractère gras.

8.2.2, nous partons du principe que la taille globale de 9 points qui est « optimale » pour la méthode MGL, est aussi optimale pour le cas des méthodes d'apprentissage MMG. En fixant la taille du voisinage à 9 points, nous réduisons ainsi la combinatoire des paramètres que nous cherchons à déterminer.

Nous proposons alors d'évaluer les voisinages suivants :

$\{1 \times 9, 3 \times 3, 9 \times 1\}$ avec des fenêtres de Hanning ou bien rectangulaires.

Afin de déterminer le voisinage optimal pour les différentes méthodes d'apprentissages de MMG, nous évaluons les perfomance en séparation (SDR), en faisant le décodage des MMG sur les sources oracles.

Voisinages adaptatifs : Pour la méthode `DUET-MMG`, qui se base sur le modèle local de DEMIX où l'on cherche à exploiter les régions temps-fréquence où une seule source est active, on peut utiliser le critère de fiabilité afin de sélectionner de façon adaptative le voisinage $\widehat{V}(t,f)$ le plus adapté parmi une combinaison $\mathcal{V}$ de voisinages différents :

$$\widehat{V}(t,f) = \arg\max_{V \in \mathcal{V}} \widehat{\mathcal{T}}\left(\Omega_{t,f}^{V}\right) \tag{8.21}$$

Afin d'évaluer cette approche nous proposons de la tester avec la combinaisons des voisinages $\mathcal{V} = \{1 \times 9, 9 \times 1\}$.

Analyse des résultats : Les résultats obtenus pour la méthode `DUET-MMG` (MMG appris à partir de la séparation spatiale de type DUET) résumés dans le tableau 8.2, nous indiquent que :

- Pour les signaux de parole, à part dans le cas des mélanges de 6 sources, l'utilisation du voisinage adaptatif consistant à choisir, parmi la combinaison des états 9×1 et 1×9, celui qui a la plus grande fiabilité, obtient de meilleurs résultats qu'en choisissant le voisinage 9×1 ou bien 1×9 de façon fixe. Cette remarque est aussi valable pour les signaux de musique dans le cas du mélange à 3 sources. Ces résultats suggèrent donc que l'utilisation d'un critère pour choisir de façon adaptative le voisinage de chaque point temps-fréquence, puisse permettre d'améliorer les résultats par rapport à une méthode basée sur un seul type de voisinage.
- L'utilisation du fenêtrage de Hanning obtient les meilleurs résultats pour les signaux de parole, mais pas pour les signaux musicaux. Cependant comme pour la méthode MGL, les différences de résultats entre les deux types de fenêtrage sont très faibles (entre 0 et 0.1 dB).

Les résultats obtenus pour la méthode MGL-MMG (MMG appris à partir de la séparation spatiale MGL) résumés dans le tableau 8.3, nous indiquent que :

- Dans le cas des mélanges à 3 sources, le voisinage optimal est le même que dans le cas de la méthode MGL, c'est à dire le voisinage 3×3 avec une fenêtre de Hanning ;
- Quand le nombre de sources est supérieur à 3, il n'y a pas un type de voisinage pour lequel les résultats se démarquent des autres. Le voisinage optimal est tantôt 9×1, tantôt 1×9 ou 3×3, tantôt avec une fenêtre de Hanning, tantôt avec une fenêtre rectangulaire.

voisinage DUET-MMG	parole				musique			
nbr de sources →	3	4	5	6	3	4	5	6
$\{9\times1, 1\times9\}R$	11.4	7.2	4.7	2.9	**16.1**	11.8	8.5	7.0
$9\times1R$	10.8	7.0	4.7	3.0	15.4	11.8	8.9	7.2
$1\times9R$	10.9	7.0	4.8	3.0	15.3	11.7	8.8	7.2
$3\times3R$	10.8	6.9	4.7	2.9	15.3	**11.9**	**9.0**	7.3
$\{9\times1, 1\times9\}H$	**11.5**	**7.4**	**4.9**	3.0	16.0	11.8	8.5	7.1
$9\times1H$	10.7	7.0	4.8	2.9	15.5	11.8	9.0	7.2
$1\times9H$	10.7	6.9	4.7	**3.0**	15.4	11.8	9.0	7.3
$3\times3H$	10.8	6.9	4.7	3.0	15.3	11.8	9.0	**7.3**

TAB. 8.2 – Evaluation des performances en séparation (en terme de SDR en dB) de la méthode `DUET-MMG` (MMG appris à partir de la séparation spatiale de type DUET), pour différents voisinages $\mathrm{L_T}\times\mathrm{L_F}$, où $\mathrm{L_T}$ est la largeur du voisinage en trames, tandis que le paramètre $\mathrm{L_F}$ est la hauteur en nombre de bins fréquentiels du voisinage. Le type de fenêtrage appliqué au voisinage est indiqué par la lettre majuscule qui suit : « R » pour une fenêtre rectangulaire et « H » pour une fenêtre de Hanning. Le décodage des états est effectué sur les sources oracles. Les meilleurs résultats de chaque groupe de mélanges sont écrits en caractère gras.

– Contrairement à la méthode MGL, l'utilisation de la fenêtre de Hanning n'améliore pas toujours les résultats par rapport à la fenêtre rectangulaire.

8.3.3 Évaluation des MMG spectraux

Intention expérimentale : Nous souhaitons évaluer les limites des méthodes MMG spectrales d'estimation des sources que nous avons proposées à la section 8.2. En particulier, nous souhaitons :

1. évaluer le nombre d'états $\mathcal{K}$ optimal des différentes méthodes proposées ;
2. évaluer l'écart de performance entre les méthodes d'estimation *aveugle* des MMG et les méthodes MMG *oracles* obtenues à partir des sources du mélange.

Estimateur dur : Plutôt que d'estimer les sources par un filtre de Wiener pondéré par les probabilités a posteriori de chaque état :

$$\widehat{\mathbf{S}}(t,f) = \sum_{\mathbf{k}} \gamma_{\mathbf{k}}(t)\mathbf{W}_{\mathbf{k}}(f)\mathbf{X}(t,f)$$

(où $\mathbf{W}_{\mathbf{k}}(f)$ est défini à l'équation (4.9)) nous préférons, afin de limiter la complexité calculatoire, *décoder* l'état :

$$\mathbf{k}^*(t) = \arg\max_{\mathbf{k}} \gamma_{\mathbf{k}}(t)$$

voisinage MGL-MMG	parole				musique			
nbr de sources →	3	4	5	6	3	4	5	6
$9 \times 1R$	11.6	6.4	3.1	1.2	15.5	10.9	6.8	4.9
$1 \times 9R$	11.9	7.4	4.3	**2.2**	15.9	**11.1**	7.5	5.6
$3 \times 3R$	11.9	7.4	3.9	2.0	16.1	11.1	7.5	**6.1**
$9 \times 1H$	11.8	6.6	3.2	1.6	16.2	10.3	**7.7**	5.2
$1 \times 9H$	12.0	**7.6**	**4.4**	2.0	16.3	10.6	7.7	5.3
$3 \times 3H$	**12.1**	7.6	4.0	2.1	**16.4**	11.0	7.4	6.0

TAB. 8.3 – Evaluation des performances en séparation (en terme de SDR en dB) de la méthode MGL-MMG (MMG appris à partie de la séparation spatiale de la méthode MGL), pour différents voisinages $\mathrm{L_T} \times \mathrm{L_F}$, où $\mathrm{L_T}$ est la largeur du voisinage en trames, tandis que le paramètre $\mathrm{L_F}$ est la hauteur en nombre de bins fréquentiels du voisinage. Le type de fenétrage appliqué au voisinage est indiqué par la lettre majuscule qui suit : « R » pour une fenêtre rectangulaire et « H » pour une fenêtre de Hanning.
Le décodage des états est effectué sur les sources oracles. Les meilleurs résultats de chaque groupe de mélanges sont écrits en caractère gras.

le plus probable à chaque trame et estimer les sources par un filtre de Wiener simple :

$$\widehat{\mathbf{S}}(t,f) = \mathbf{W}_{\mathbf{k}^*(t)}(f)\mathbf{X}(t,f)$$

La distinction entre ces deux approches a été discutée à la section 4.2 sous les noms d'*estimateur doux* et d'*estimateur dur*.

Protocole expérimental : Les performances en séparation sont évaluées pour chaque méthode en terme de SDR [GBVF03], en fonction du nombre $\mathcal{K} = \{1, 2, 4, 8, 16, 32, 64, 128\}$ de Gaussiennes (d'états) du MMG des sources.

L'estimation des variances des filtres de Wiener s'obtenant en deux étapes, une étape d'apprentissage des MMG, et une étape de décodage, nous pouvons évaluer distinctement ces deux étapes en effectuant l'une des étapes à partir des sources de référence (oracles).

8.3.3.1 Évaluation des MMG oracles

Intention expérimentale : Nous souhaitons dans cette expérience :

1. évaluer les performances obtenues par les MMG oracles (`MMG OO`), c.-à-d. appris et décodés sur les sources de références, en fonction du nombre $\mathcal{K}$ d'états des MMG des sources ; La figure 8.6 donne le schéma bloc de l'apprentaissage oracle, et la figure 8.7(a) donne le schéma bloc du décodage oracle.
2. évaluer les performances obtenues par les MMG oracles décodés sur le mélange (`MMG OM`), afin d'évaluer l'éventuelle perte de performance dûe au décodage des

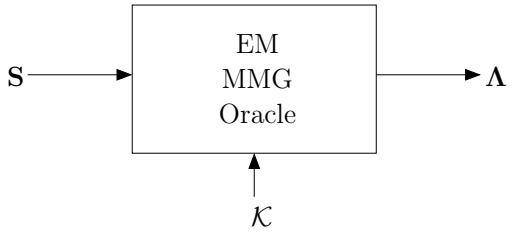
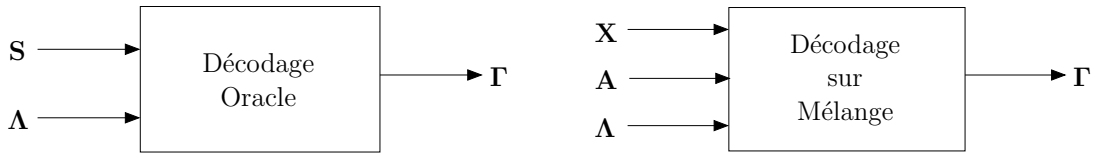


FIG. 8.6 – Schéma bloc de l'apprentissage oracle à partir des sources de référence. L'apprentissage est effectué par l'algorithme EM donné en annexe à la section B.1. $\mathcal{K}$ désigne le nombre d'états des modèles MMG.

MMG sur le mélange, plutôt que sur les sources oracles qui ne sont pas disponibles dans les situations réelles ; La figure 8.7(b) donne le schéma bloc du décodage sur le mélange.



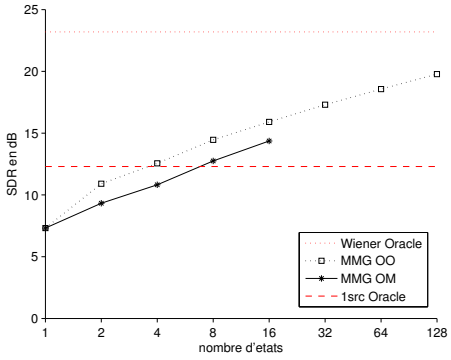
(a) Décodage Oracle (b) Décodage sur le mélange

FIG. 8.7 – Schéma bloc des méthodes de décodage de MMG de l'état de l'art qui sont présentées au chapitre 4

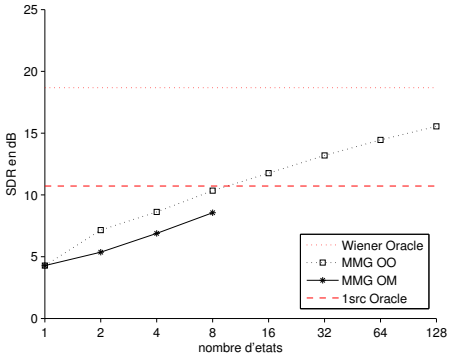
3. comparer ces résultats avec les performances oracles [VGP07] afin d'évaluer les performances maximales atteignables par les méthodes à base de MMG-spectraux que nous proposons, et les méthodes standard de masquage binaire. Les oracles que nous souhaitons évaluer sont donc :
   - Le filtrage de Wiener adaptatif oracle (`Wiener Oracle`),c.-à-d. la séparation de chaque point temps-fréquence par le filtre de Wiener $\mathbf{W}(t,f)$ qui connaissant les sources minimise l'EQM spectrale $\mathbb{E}\left\{\left\|\widehat{\mathbf{S}}(t,f)-\mathbf{S}(t,f)\right\|^2\right\}$;
   - Le masquage binaire oracle (`1src Oracle`), c.-à-d. la séparation de chaque point temps-fréquence par le masque binaire (voir section 3.2.1) qui minimise l'EQM spectrale.

Analyse des résultats : Suite à l'observation des figures 8.8 et 8.9, nous faisons les remarques suivantes :

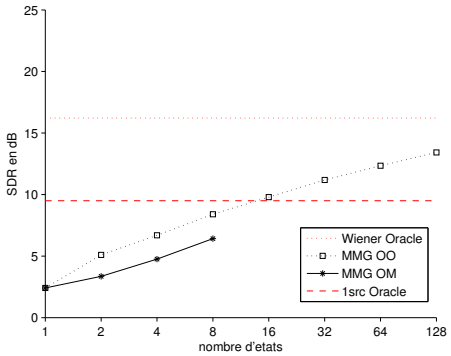
- La méthode `MMG OM` n'a peu être calculée pour toutes les valeurs de $\mathcal{K}$. En effet étend donné que le nombre d'états à décoder par cette méthode est exponentiel en fonction du nombre de sources ($\mathcal{K}^N$), nous avons vite attend les limites de nos machines (PC Bipro Intel Xeon 3.40GHz et disposant de 4 GigaOctets de mémoire) en terme de capacité mémoire.



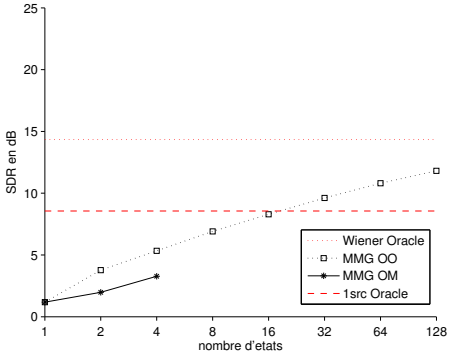
(a) 3 sources de parole



(b) 4 sources de parole



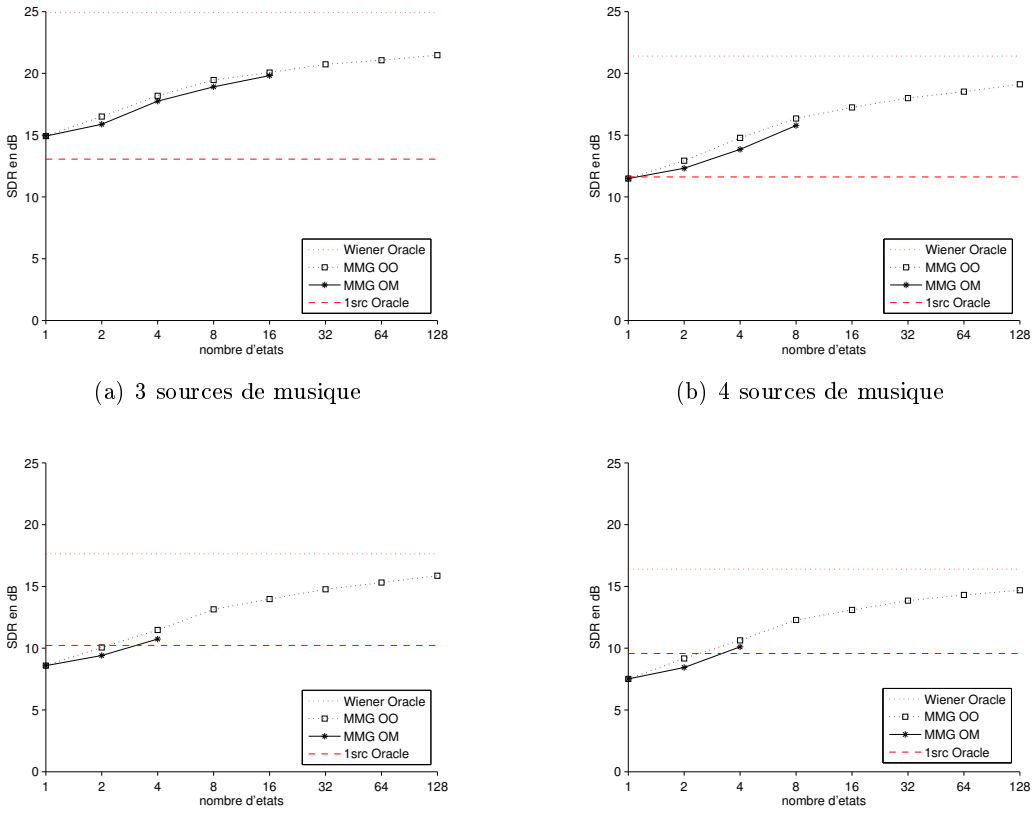
(c) 5 sources de parole



(d) 6 sources de parole

FIG. 8.8 – Performances en séparation (SDR) des méthodes MMG oracles en fonction du nombre d'états sur les mélanges de parole. `Wiener Oracle` est le filtre de Wiener oracle, `MMG OO` est la séparation par MMG appris et décodés sur les sources, `MMG OM` est la séparation par MMG appris sur les sources et décodés sur le mélange.

- L'écart de performance en séparation entre le décodage sur le mélange (pour les valeurs qui ont pu être calculées) et le décodage sur les sources oracles ne s'accroit pas quand le nombre d'état augmente, et est plus important pour les signaux de parole (entre $1,5$ et 2 dB) que pour les signaux musicaux (moins de 1 dB) ; Remarquons que la valeur $\mathcal{K} = 1$ est un cas particulier, car il correspond au filtre de Wiener classique, et par conséquent ne nécessite pas d'étape de décodage.
- Les courbes des MMG des sources de musique sont plus proches de leur borne maximale donnée par `Wiener Oracle`, que ne le sont les MMG des sources de parole. Aussi, la pente des courbes des MMG des sources de parole étant plus importante que pour les MMG des sources de musique, le gain obtenu par l'augmentation du nombre d'état des MMG est plus important pour les sources de paroles que pour les sources de musique.



(a) 3 sources de musique

(b) 4 sources de musique

(c) 5 sources de musique

(d) 6 sources de musique

FIG. 8.9 – Performances en séparation (SDR) des méthodes MMG oracles en fonction du nombre d'états sur les mélanges de musique. `Wiener Oracle` est le filtre de Wiener oracle, `MMG OO` est la séparation par MMG appris et décodés sur les sources, `MMG OM` est la séparation par MMG appris sur les sources et décodés sur le mélange.

En conclusion, pour un même niveau de qualité, il est nécessaire d'avoir un nombre d'états par MMG spectral plus important pour séparer les signaux de parole que pour séparer les signaux de musique.

Ces résultats ne sont cependant valables que lorsque les MMG spectraux estimés sont très représentatifs des sources. Dans les cas réels on ne dispose pas des sources du mélange, et l'estimation des MMG spectraux est un problème difficile. En particulier, il est peu probable que pour des MMG spectraux appris par les approches aveugles que nous proposons dans ce chapitre, les performances en séparation croissent, comme dans le cas oracle, de façon monotone en fonction du nombre d'états.

8.3.3.2 Évaluation des MMG aveugles

Intention expérimentale : Nous souhaitons dans cette expérience :

1. évaluer le nombre d'états $\mathcal{K}$ optimal pour les méthodes d'apprentissage aveugles de MMG spectraux que nous avons présenté dans ce chapitre, c.-à-d. les méthodes `Spat-MMG`, `DUET-MMG` et `MGL-MMG`, présentées respectivement aux sections 8.2.1, 8.2.2.1 et 8.2.2.2 ;
2. évaluer les méthodes de décodage aveugle des états que nous avons proposées à la section 8.2.3 en la comparant au décodage oracle (sur les sources de référence).

Méthodes comparées : Nous comparons pour différentes valeurs de $\mathcal{K}$ les performances en séparation des méthodes de séparation par MMG spectraux pour les différentes combinaisons des étapes d'apprentissage des MMG et de décodage des états :

décodage\apprentissage	`Spat-MMG`	`DUET-MMG`	`MGL-MMG`
oracle	`Spat-MMG O`	`DUET-MMG O`	`MGL-MMG O`
aveugle	`Spat-MMG A`	`DUET-MMG A`	`MGL-MMG A`

Choix des voisinages : L'apprentissage des MMG par les méthodes que nous avons présentées à la section 8.2 dépend du type de voisinage choisi. Au regard des résultats des expériences menées à la section 8.3.2.2 sur le type de voisinage des méthodes MMG, nous évaluons les MMG appris sur les sources spatiales avec les voisinages suivants :

- 3×3 avec une fenêtre de Hanning pour la méthode `MGL-MMG` ;
- $\{9 \times 1, 1 \times 9\}$ avec une fenêtre de Hanning pour la méthode `DUET-MMG`. Le critère de sélection du voisinage est la fiabilité.
- 3×3 avec une fenêtre de Hanning pour la méthode `Spat-MMG` sur les signaux de parole, et 9×1 avec une fenêtre de Hanning pour la méthode `Spat-MMG` sur les signaux de musique.

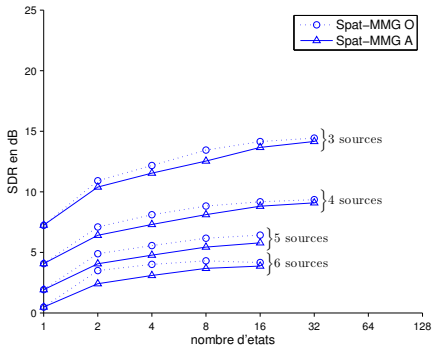
Evaluation de la méthode Spat-MMG Les résultats de la méthode `Spat-MMG`, introduite à la section 8.2.1 sont présentés à la figure 8.10

Evaluation de la méthode DUET-MMG Les résultats de la méthode `DUET-MMG`, introduite à la section 8.2.2.1 sont présentés à la figure 8.11.

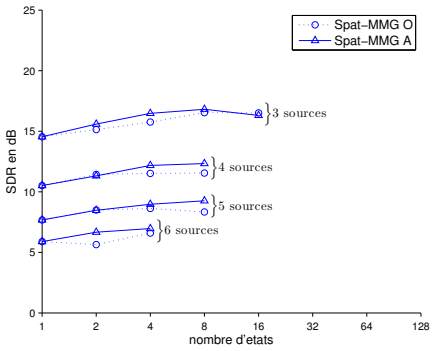
Evaluation de la méthode MGL-MMG Les résultats de la méthode `MGL-MMG`, introduite à la section 8.2.2.2 sont présentés à la figure 8.12

Analyse des résultats Comme on pouvait s'y attendre, contrairement aux MMG oracle, les performances n'augmentent pas systématiquement avec le nombre d'états des MMG. Au contraire, au delà d'un certain nombre d'états, les performances chutent dans la plupart des cas, ce qui montre la difficulté de l'apprentissage des MMG en aveugle quand le nombre d'états augmente.

Les résultats des expériences permettent néanmoins d'estimer le nombre d'états optimal des MMG pour chacune des méthodes et pour différentes configurations du mélange. Quelle que soit la méthode employée, les résultats confirment que, même dans

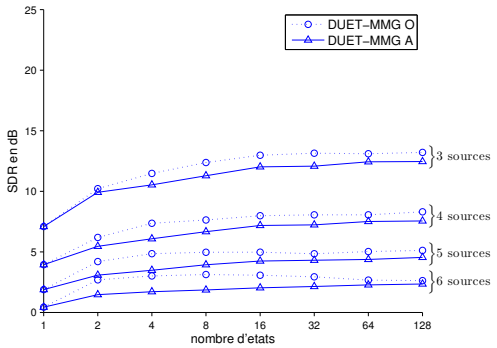


(a) signaux de parole

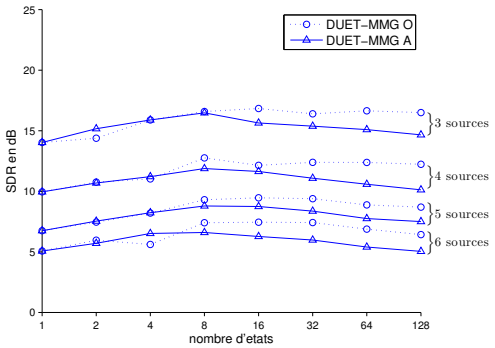


(b) signaux de musique

FIG. 8.10 – Performances en séparation (SDR) de la méthode `Spat-MMG` en fonction du nombre d'états. Pour la méthode `Spat-MMG-O`, le décodage des états a été effectué sur les sources oracles, tandis que pour la méthode `Spat-MMG-A`, le décodage des états a été effectué suivant le même critère que lors de l'étape d'apprentissage du MMG. Le voisinage est 3×3 avec fenêtre de Hanning pour les signaux de parole, et 9×1 avec fenêtre de Hanning pour les signaux de musique.



(a) signaux de parole

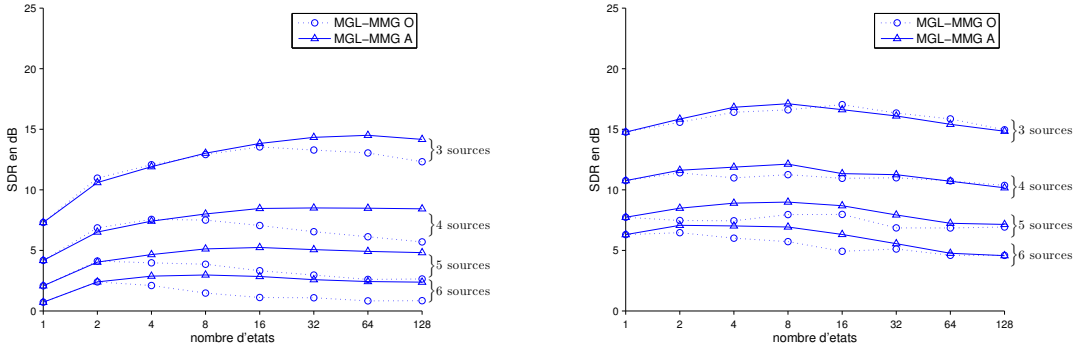


(b) signaux de musique

FIG. 8.11 – Performances en séparation (SDR) de la méthode `DUET-MMG` en fonction du nombre d'états. Pour la méthode `DUET-MMG-O`, le décodage des états a été effectué sur les sources oracles, tandis que pour la méthode `DUET-MMG-A`, le décodage des états a été effectué suivant le même critère que lors de l'étape d'apprentissage du MMG. Le voisinage utilisé est $\{9 \times 1, 1 \times 9\}$ avec la fiabilité comme critère de sélection, et un fenêtrage de Hanning.

le cas où les MMG sont appris en aveugle, un plus grand nombre d'états par source est souhaitable pour les signaux de parole ($\mathcal{K} = 64$) que pour les signaux musicaux ($\mathcal{K} = 8$).

La méthode de décodage des MMG ne semble pas poser de difficulté particulière car



(a) signaux de parole (b) signaux de musique

FIG. 8.12 – Performances en séparation (SDR) de la méthode `MGL-MMG` en fonction du nombre d'états. Pour la méthode `MGL-MMG-O`, le décodage des états a été effectué sur les sources oracles, tandis que pour la méthode `MGL-MMG-A`, le décodage des états a été effectué suivant le même critère que lors de l'étape d'apprentissage du MMG. Le voisinage utilisé est 3×3 avec une fenêtre de Hanning.

les résultats sont proches et même parfois supérieurs à ceux obtenus par le décodage sur les sources oracles.

8.3.4 Comparaison des méthodes par rapport à l'état de l'art

Intention experimentale : Nous souhaitons comparer entre elles les méthodes proposées dans ce chapitre, c.-à-d. les méthodes `MGL`, `Spat-MMG`, `DUET-MMG` et `MGL-MMG`, ainsi que comparer ces méthodes avec l'état de l'art. Les algorithmes de l'état de l'art auquels nous souhaitons comparer nos méthodes sont :

1. l'algorithme DUET [YR04] car il est très connu et sert donc de référence ;
2. l'algorithme *Lp* d'E. Vincent [Vin07] car il a obtenu les meilleurs résultats en séparation sur les mélanges instantanés lors de la campagne d'évaluation d'ICA'07 [VSB+07].

8.3.4.1 Évaluation de la méthode MGL

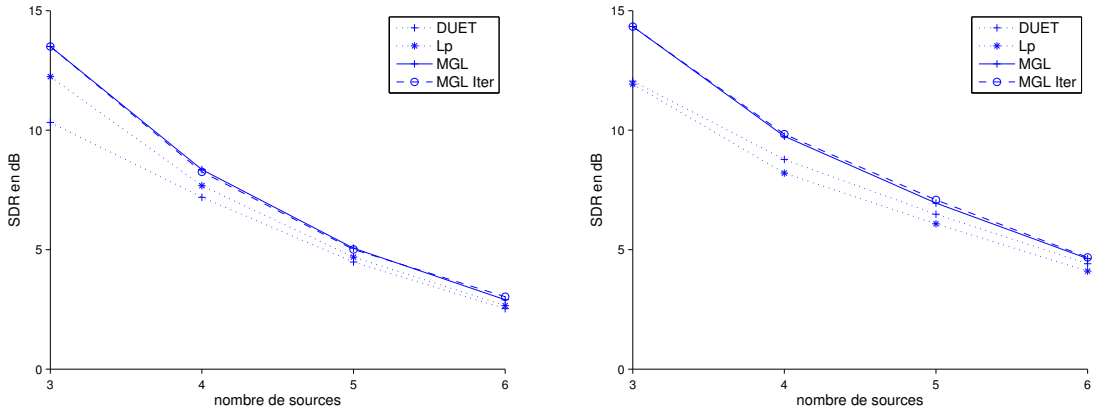
Intention expérimentale : Nous souhaitons évaluer la méthode MGL que nous avons présentée à la section 8.1. En particulier, nous souhaitons :

- Comparer les résultats en séparation de notre méthode MGL « heuristique » (notée `MGL`) présentée à la section 8.1.2.2, avec la méthode MGL « iterative » (`MGL Iter`), présentée à la section 8.1.2.1. Afin d'effectuer cette optimisation itérative, nous utilisons la fonction `fmincon` de la Toolbox d'optimisation de Matlab.
- Comparer la méthode MGL avec les algorithmes DUET et *Lp* de l'état de l'art.

Analyse des résultats : Les résultats présentés à la figure 8.13 montrent que :

Comparaison de MGL heuristique et MGL Itératif : La méthode MGL heuristique (`MGL`) présentée à la section 8.1.2.2 obtient des **performances en séparation similaires** à la méthode MGL itérative (`MGL Iter`). Cependant la méthode heuristique nécessite un **temps de calcul 1000 fois moins long** que pour la méthode itérative. En effet, contrairement à la méthode MGL itérative dont le temps de calcul pour séparer les sources d'un mélange de 10 secondes, se compte en heures, la méthode heuristique que nous proposons ne nécessite que quelques secondes pour effectuer la même tâche sur la même machine.

Comparaison de MGL à DUET et Lp : Dans tous les cas, les méthodes MGL obtiennent des meilleurs résultats en séparation que les méthodes DUET et Lp. Le gain en performance est d'environs **3 dB par rapport à DUET** pour les mélanges à 3 sources, tandis qu'il est d'environ **1 dB par rapport à Lp** pour les mêmes mélanges de parole, et d'environ 2,5 dB pour les signaux de musique.



(a) parole (b) musique

FIG. 8.13 – Performances en séparation (SDR) de la méthode `MGL` avec la méthode `MGL Iter`, `DUET` et `Lp` en fonction du nombre de sources. les voisinages utilisés pour les méthodes `MGL` et `MGL Iter` sont les voisinages 3×3 avec fenêtrage de Hanning pour les mélanges de paroles et 9×1 avec fenêtrage de Hanning pour les mélanges de musique.

8.3.4.2 Évaluation des MMG spectraux appris en aveugle

Intention expérimentale : Nous souhaitons évaluer les méthodes de séparation par MMG spectraux appris en aveugle `Spat-MMG`, `DUET-MMG` et `MGL-MMG` et :

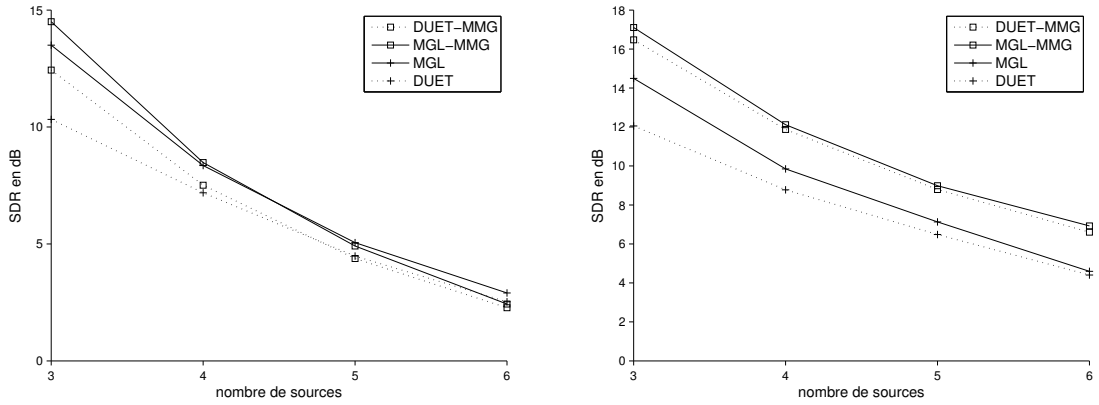
1. comparer les méthodes DUET et MGL avec `DUET-MMG` et `MGL-MMG` afin d'évaluer le gain en performance dû à l'introduction de la structure du modèle MMG ;
2. comparer les méthodes `Spat-MMG`, `DUET-MMG` et `MGL-MMG` entre elles afin de pouvoir choisir la méthode optimale ;

3. comparer ces méthodes avec l'état de l'art. La méthode MGL donnant des meilleurs résultats en séparation que la méthode Lp sur les mélanges testés, nous ne comparons pas les performances de cette dernière avec les méthodes MMG pour ne pas encombrer les figures.
4. comparer ces méthodes avec l'oracle MMG (MMG spectraux appris et décodés sur les sources de référence) et le masquage binaire oracle `1src Oracle` (application du masque binaire minimisant l'EQM spectrale lorsqu'on connait les sources), c.-à-d. l'approche oracle sur laquelle est basée la méthode DUET. Les performances oracles donnant une borne maximale sur les résultats qu'il est possible d'atteindre par la même approche en aveugle, nous voulons savoir si, et le cas échéant quand, les méthodes MMG spectrales en aveugle obtiennent de meilleurs résultats que le masquage oracle binaire.

Configuration des méthodes : Pour l'ensemble des méthodes à base de MMG nous choisissons le nombre d'états par source suivant :

- $\mathcal{K} = 64$ états par source pour les signaux de parole ;
- $\mathcal{K} = 8$ états par source pour les signaux de musique.

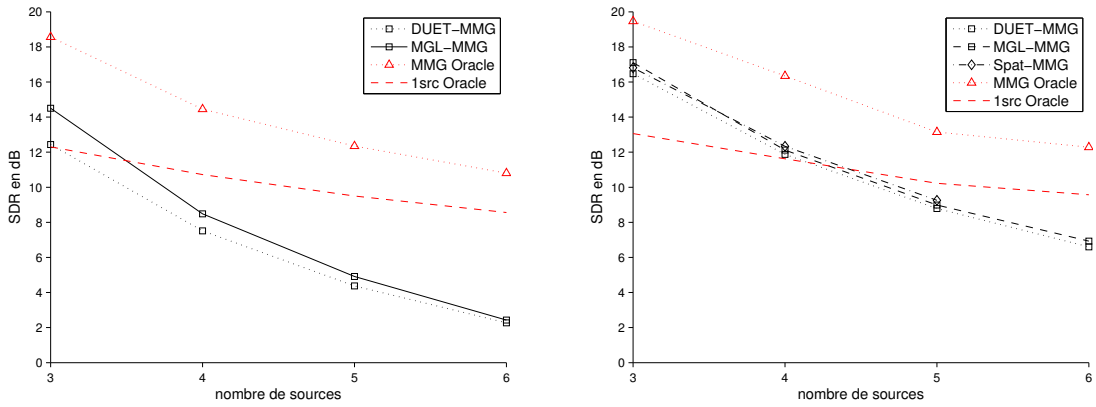
Le voisinage utilisé pour la méthode `DUET-MMG` est $\{9 \times 1, 1 \times 9\}$ avec la fiabilité comme critère de sélection, et un fenêtrage de Hanning. Le voisinage utilisé pour la méthode `MGL-MMG` est 3×3 avec fenêtrage de Hanning. Le voisinage utilisé pour la méthode `Spat-MMG` est 9×1 avec fenêtrage de Hanning.



(a) parole (b) musique

FIG. 8.14 – Performances en séparation (SDR) en fonction du nombre de sources, des méthodes `DUET`, `DUET-MMG`, `MGL` et `MGL-MMG`. Pour les signaux de paroles les MMG ont 64 états par source, tandis que pour les signaux de musique, les MMG ont 8 états par source.

Analyse des résultats : Sur la figure 8.14 nous avons affiché les performances des méthodes `DUET`, `MGL`, `DUET-MMG` et `MGL-MMG`, afin d'évaluer l'impacte de l'utilisation de



(a) parole (b) musique

FIG. 8.15 – Performances en séparation (SDR) en fonction du nombre de sources, des méthodes `Spat-MMG` (musique seulement), `DUET-MMG` et `MGL-MMG` ainsi que celles de l'oracle MMG et du masquage binaire oracle `1src Oracle`. Pour les signaux de paroles les MMG ont 64 états par source, tandis que pour les signaux de musique, les MMG ont 8 états par source.

la structure imposée par le modèle MMG par rapport aux mêmes méthodes sans cette structure.

Les résultats affichés sur la figure 8.14 font apparaître les points suivants :

- Les méthodes `DUET-MMG` et `MGL-MMG` obtiennent respectivement de meilleurs résultats que les méthodes `DUET` et `MGL`, à part pour les signaux de parole quand le nombre de sources du mélange est supérieur ou égal à 5. La supériorité des méthodes MMG se fait davatange sentir sur les signaux musicaux, où la méthode `DUET-MMG` est entre 2 dB et 4 dB supérieur à `DUET`, tandis que `MGL-MMG` est environ 2 dB supérieur à `MGL`.
- Dans tous les cas `MGL-MMG` **est meilleur que** `DUET-MMG`, bien que l'écart soit plus serré sur les signaux musicaux.

Sur la figure 8.15 nous avons affiché les performances des méthodes MMG aveugles proposées dans ce chapitre, c.-à-d. les méthodes `DUET-MMG`, `MGL-MMG` et `Spat-MMG`, ainsi que les performances des oracles MMG et celui du masquage binaire.

Notons que nous n'avons pas pu calculer les résultats de la méthode `Spat-MMG` pour les signaux de parole car nous n'avions pas la puissance de calcul nécessaire pour effectuer l'apprentissage des MMG de 64 états.

Les résultats affichés sur la figure 8.15 font apparaître les points suivants :

- La méthode `Spat-MMG` obtient des résultats parfois meilleurs que `MGL-MMG` sur les signaux de musique, mais la complexité calculatoire de la méthode la rend difficilement utilisable en pratique.
- Les méthodes MMG spectrales obtiennent pour les mélanges de 3 sources, des performances, **d'environs 2 dB pour les signaux de paroles et 4 dB pour les signaux de musique, supérieures à l'oracle du masquage binaire**.

Dans le cas des signaux de musique, les résultats des méthodes MMG spectrales sont supérieures au masquage binaire oracle pour les mélanges ayant moins de 5 sources.

- L'écart de performance des méthodes MMG par rapport au MMG oracle est important, en particulier pour les signaux de parole où il est de 4 à 8 dB, contre 2 à 5 dB pour les signaux de musique. Remarquons aussi que cet écart est d'autant plus important que le nombre de sources du mélange est grand. Autrement dit il est plus difficile d'apprendre correctement les MMG en aveugle dans le cas des signaux de parole, et quand le nombre de source du mélange est important.

Remarque sur la performance des méthodes : Nous donnons à titre indicatif les temps de calcul des différentes méthodes (aveugles) que nous avons évaluées dans ce chapitre pour estimer les sources d'un mélange de 4 sources extrait de :

- la base de test des signaux de parole de durée 11 secondes et échantillonnés à 8 KHz ;
- la base de test des signaux musicaux de durée 11 secondes et échantillonnés à 22,05 KHz ;

Le nombre d'états des méthodes MMG est fixé à $\mathcal{K} = 64$ pour le mélange de parole et à $\mathcal{K} = 8$ pour le mélange de musique. La machine utilisée est un PC Bipro Intel Xeon 3.40GHz et disposant de 4 GigaOctets de mémoire.

méthodes	`Spat-MMG`	`DUET-MMG`	`MGL-MMG`	`MGL`	`DUET`	`Lp`
parole	Nc	103 s	108 s	27 s	0.6 s	0.8 s
musique	54 min	132 s	133 s	120 s	1.9 s	1.8 s

8.4 Conclusion

Les méthodes d'apprentissage de MMG que nous avons étudiées dans ce chapitre se basent sur des estimations spatiales des sources ainsi que sur l'analyse des matrices de covariances locales qui ont déjà été mises à profit par la méthode DEMIX d'estimation des directions présenté au chapitre 6. Contrairement aux approches de l'état de l'art des méthodes d'apprentissage de MMG à partir du mélange, les approches proposées dans ce chapitre ont une complexité qui reste linéaire en fonction du nombre d'états du MMG.

Les approches proposées permettent d'améliorer de plusieurs dB en SDR les résultats en séparation par rapport aux méthodes de l'état de l'art comme DUET et Lp. Les résultats obtenus sont plus flagrant sur les signaux musicaux, ou quand le nombre de source n'est pas trop grand. En effet quand le nombre de source augmente, la parcimonie du mélange diminue, et il devient plus difficile de trouver des régions temps-fréquence où peu de sources sont actives, rendant difficile l'exploitation des matrices de covariance locale. Les résultats sur les signaux de parole sont moins bons que pour la musique, parce que le nombre d'états nécessaire à la séparation des sources de parole est plus important pour la parole que pour la musique, et parce que la difficulté de l'apprentissage MMG augmente avec le nombre d'états.

La comparaison des approches proposées suggère que l'on utilise la méthode `MGL-MMG` car :

- elle donne de meilleurs résultats en séparation que les méthodes `MGL` et `DUET-MMG`, pour un temps de calcul ayant le même ordre de grandeur ;
- elle améliore la séparation de 2 à 5 dB (selon le nombre de sources et la nature des signaux) en SDR par rapport à la méthode DUET de l'état de l'art, et dépasse même les performances du masquage binaire oracle quand il y a moins de 4 sources.

Cependant il faut noter que la méthode `MGL-MMG` s'appuie sur la méthode MGL qui a été définie uniquement pour des mélanges stéréophoniques instantanés. Il n'est pas évident que la généralisation de la méthode MGL au cas multicanal en général ($M > 2$) se fasse sans qu'il soit nécessaire de recourir à une minimisation itérative du critère de vraisemblance, ce qui qui aurait pour conséquence de la rendre rédhibitoire. La généralisation de l'approche `MGL-MMG` aux conditions des mélanges anéchoïques et convolutifs n'est également possible que si l'on est capable, d'estimer les paramètres du mélange et de généraliser la méthode MGL, à ces conditions.

Au delà des approches `MGL-MMG` et `DUET-MMG` que nous avons proposées dans ce chapitre, nous avons proposé une architecture pour l'apprentissage aveugle des MMG spectraux, pour laquelle les méthodes `MGL-MMG` et `DUET-MMG` ne sont que des exemples d'implémentation. Nous pouvons ainsi espérer que d'autres approches pourront exploiter cette architecture ou s'en inspirer.

Troisième partie

Conclusion et perspectives

Chapitre 9

Conclusion

Nous nous somme intéressés au problème de séparation de sources audio à partir d'un enregistrement multicanal, et en particulier dans le cas sous-déterminé où le nombre de sources est supérieur au nombre de canaux. Dans ce dernier cas, le problème se divise généralement en 2 étapes distinctes : l'estimation des directions du mélanges, puis l'estimation des sources à partir du mélange et des directions. Les approches permettant de traiter le cas sous-déterminé s'appuient en général sur les hypothèses d'indépendance et de parcimonie des sources dans le plan temps-fréquence. Pour les points temps-fréquence où ces hypothèses sont valides, le mélange est alors localement sur-déterminé, et il est possible d'identifier localement les directions et les sources par l'exploitation de l'information spatiale.

Ces hypothèses peuvent en pratique être considérées comme valides pour une partie des points temps-fréquence, mais pour les autres points temps-fréquence, les estimations locales des directions et des sources sont très peu fiables.

Inspiré par les travaux de Deville, qui se base sur une hypothèse de parcimonie plus faible et plus réaliste que celle de l'approche standard et qui propose une approche d'analyse spatiale qui soit locale dans le plan temps-fréquence, plutôt que ponctuelle, nous avons exploité les matrices de covariance locales afin d'inférer des informations relatives à la distribution locale des sources. En particulier, l'analyse des matrices de covariance locales permet d'estimer localement le nombre de sources actives si celui-ci est inférieur au nombre de canaux. Autrement dit on est capable d'avoir une estimation locale de la parcimonie du mélange. Ainsi on est en mesure d'exploiter les estimations locales (de directions et de sources) des méthodes basées sur la parcimonie avec une mesure de la fiabilité de ces estimations.

Nous exploitons cette mesure de fiabilité, à travers un algorithme de clustering appelé DEMIX qui permet de fusionner entre elles les estimations locales des directions et enfin d'estimer à la fois le nombre de sources du mélange (tâche qui a été très peu abordée par l'état de l'art) et les directions « globales » du mélange. Les évaluations expérimentales montrent que l'algorithme est beaucoup plus robuste que DUET pour l'estimation des mélanges instantané, car il permet d'estimer des directions du mélange qui sont très proches, et permet d'estimer des délais de plusieurs dizaines d'échantillons.

Pour la deuxième étape du problème de séparation de source, qui concerne l'estimation des sources connaissant les directions du mélange, nous avons exploité les informations des matrices de covariance locales conjointement à l'information donnée par la connaissance des directions du mélange pour inférer les variances des sources que nous avons supposées localement gaussiennes. Contrairement aux méthodes basées sur la parcimonie qui supposent que les sources sont identiquement distribuées, et qui reposent uniquement sur la diversité spatiale pour séparer les sources, notre approche suppose que les sources sont distribuées différemment les unes des autres, et par conséquent, notre approche se base en plus de la diversité spatiale sur la diversité spectrale des sources. L'estimation des sources par filtrage de Wiener, avec les variances ainsi estimées (méthode MGL), permet d'obtenir des résultats en séparation supérieurs de plusieurs dB en SDR, par rapport aux méthodes tels que DUET ou Lp qui sont basées uniquement sur la parcimonie des sources. Nous avons également proposé une méthode d'apprentissage de MMG spectraux, afin d'améliorer les estimations de variances obtenus localement par la méthode MGL, en imposant une contrainte structurelle globale. La méthode d'apprentissage MMG proposée, contrairement aux méthodes d'apprentissage MMG de l'état de l'art, n'a pas besoin de connaissance a priori sur les sources et garde une complexité calculatoire linéaire en fonction du nombre d'états des MMG. La méthode proposée permet d'améliorer de quelques dB en SDR la séparation par rapport à la méthode MGL. A la vision des résultats des méthodes MMG oracles qui ont été présentés dans ce manuscript, il reste une marge de plusieurs dB pour l'amélioration de l'étape d'apprentissage des MMG.

Chapitre 10

Perspectives

Dans cette thèse, nous avons surtout évoqué la séparation de sources dans le cas stéréophonique sous-déterminé, pour des mélanges instantané ou bien anéchoïque. Nous présentons dans ce chapitre quelques pistes possibles afin d'étendre et d'améliorer les approches étudiées dans cette thèse.

L'estimation des directions dans le cas multicanal où $M \geq 2$: La méthode DEMIX d'estimation des directions du mélange a été formulée dans le cas multicanal avec $M \geq 2$ dans les cas instantané et anéchoïque, mais les expériences n'ont été faite que dans le cas stéréophonique. Aussi dans DEMIX, on cherche à l'aide de la mesure de fiabilité les régions temps-fréquence où une seule source est active, afin d'obtenir une estimation locale d'une des directions du mélange. Dans le cas multicanal avec $M \geq 2$, on peut généraliser la mesure de fiabilité, pour détecter les régions où il y $J < M$ sources actives par une méthode dans l'esprit de Wax et Kailath [WK85], et identifier, non plus des directions locales, mais des sous-espaces dits de *concentration* dans lesquels se situent les coefficients du mélange. A partir de l'identification des sous-espaces de *concentration*, il est alors possible, soit pour une région temps-fréquence où le nombre de sources actives est inférieur à M, d'estimer les directions de ces sources actives au maximum de vraisemblance à partir de la matrice de covariance locale [ZW88], ou bien d'estimer les directions du mélange par clustering des sous-espaces de *concentration* [NMBZJ07, NBZJ07].

Extensions au cas convolutif : La méthode DEMIX d'estimation des directions du mélange a été formulée est évaluée dans le cas anéchoïque, qui est une forme simple du cas convolutif. L'extension de la méthode au cas convolutif en général n'est pas évidente. Une piste pour étendre la méthode DEMIX au cas convolutif serait de se baser sur l'hypothèse que les filtres sont parcimonieux dans le domaine temporel, le cas anéchoïque étant le cas particulier d'un filtre parcimonieux de degré 1. Bien évidemment, dans le cas général où la parcimonie des filtres (le nombre de valeurs non-nulles) est supérieur à 1, la différence d'intensité d'une direction n'est plus constante à travers les fréquences, si bien que d'autre techniques doivent être trouvées afin résoudre le « problème de

permutation », c.-à-d. afin de regrouper les estimées locales d'une direction à travers les fréquences.

L'approche en deux étapes qui a été exploitée dans cette thèse, consistant à d'abord estimer les paramètres du mélange, puis les modèles MMG des sources, pourrait sans doute être utilisée dans l'autre sens, pour estimer les paramètres du mélange convolutif à partir des modèles MMG des sources. A. Ozerov a proposé un formalisme d'*adaptation contrainte* [Oze06] qui permet d'estimer dans le cas monophonique, le filtre « de mélange » d'une des sources, à partir de la connaissance des MMG des sources. Ce formalisme est basé sur une extension de l'algorithme EM pour maximiser un critère MAP au lieu du critère de vraisemblance. Il est sans doute possible de généraliser cette approche au cas multicanal afin d'estimer les paramètres du mélange dans le cas convolutif.

Les méthodes d'estimation des sources MGL et MMG que nous avons proposées, ont été présentées dans le cas instantané, et il n'est pas évident que l'approche MGL puisse être étendu au cas convolutif sans avoir à recourir à une approche itérative rédhibitoire, cependant l'extension de l'approche d'apprentissage et de séparation par MMG ne semble pas soulever de difficultés théoriques majeures.

Amélioration de l'étape d'apprentissage des MMG : Les expériences effectuées montrent qu'il y a un écart significatif entre les résultats en séparation obtenus par les MMG appris sur les sources et ceux appris en aveugle. Afin d'améliorer l'étape d'apprentissage des MMG en aveugle, plusieurs pistes sont possibles.

On a vu avec la méthode DUET-MMG que le fait d'utiliser non pas une seule forme de voisinage, mais de choisir pour chaque point temps-fréquence une forme de voisinage parmi un ensemble de deux formes différentes de voisinage, selon un critère comme la fiabilité, permettait d'augmenter les performances de la méthode. Cette amélioration se comprend dans la mesure ou l'hypotèse de DUET est qu'il y ait une seule source active en chaque point temps-fréquence, et que le critère de fiabilité mesure en quelque sorte la validité de cette hypothèse. Par conséquent, à condition de trouver des critères qui permettent de valider localement les hypothèses de la méthode spatiale qui est utilisée pour l'apprentissage du MMG, il est sans doute possible d'améliorer l'apprentissage MMG, par un choix adaptatif de la forme du voisinage en chaque point temps-fréquence. Pour la méthode MGL il s'agirait alors de trouver un critère pour valider localement les hypothèses d'indépendance, de gaussianité et d'une certaine parcimonie des sources.

Une autre piste non exclusive avec la première, est d'utiliser une approche multi-échelle. En effet, bien que l'estimation des sources soit optimale pour une échelle assez grande (pour une taille de fenêtre de plusieurs milliers d'échantillons quand la fréquence d'échantillonnage est celle d'un CD audio), estimer des modèles MMG ayant des centaines voir des milliers de fréquences, n'est par une tâche facile, surtout si le nombre d'états du modèle est important, et la durée du signal est courte. Par contre l'apprentissage d'un MMG à une échelle ou il n'y a que quelques fréquences est beaucoup simple, car le modèle est plus simple et le nombre de trames disponibles pour l'estimation du modèle est plus élevé. Une approche visant à d'abord estimer les MMG à une petite échelle, puis à rafiner le modèle à des échelles plus grande pour les fréquences où l'in-

formation est « fiable » serait sûrement bénéfique. Le principe d'adaptation de modèle d'Ozerov [Oze06] pourrait être mis à profit pour formaliser cette approche.

Approche multimodale : Dans cette thèse nous avons essentiellement exploité : l'information spatiale pour apprendre les informations spectrales qui caractérisent les sources sonores, et exploité conjointement ces deux types d'information pour discriminer les sources de la scène sonore.

Comme évoqué dans l'introduction, l'homme fait appel à d'autres sens que l'audition pour analyser une scène sonore et en particulier la vue, et ceci d'autant plus que l'information sonore est dégradée. Ainsi par exemple, des personnes malentendantes sont souvent capables de lire sur les lèvres d'un locuteur pour en comprendre le sens.

Ainsi quand l'information spatiale est difficilement exploitable, à cause par exemple de l'effet d'une forte réverbération, une autre piste pour améliorer la séparation de source ou permettre l'apprentissage de modèles statistiques sur les sources, consiste à exploiter, conjointement à l'audio, l'information d'un autre media comme la vidéo. Les travaux précurseurs de Sodoyer [SSG+02] et Rivet et al [Riv06, RAG+] ont montrés l'intérêt d'exploiter l'information vidéo, en particulier l'observation des lèvres du locuteur, afin de détecter les silences, et permettre la séparation de sources dans le cas convolutif. Egalement, les travaux de Monaci et al [MV06] et Llagostera et al [LMVG08] ont montrés l'intérêt d'exploiter les corrélations entre l'énergie du signal sonore et les mouvements d'atomes vidéo afin de détecter les sources sonores dans l'image.

Annexes

Annexe A

Exemples de modèles de sources pour l'ACI

Dans cette annexe, nous détaillons certains résultats concernant l'ACI déterminée présentée à la section 1.1 et l'ACI sous-déterminée présentée à la section 1.2.

A.1 ACI déterminée

Dans cette section, nous dérivons le modèle de l'ACI déterminée dans les cas particuliers d'une distribution gaussienne des sources, et d'une distribution gaussienne des sources diagonale dans une base de Fourier.

A.1.1 Spécification d'une distribution gaussienne des sources

Nous discutons ici de l'identification du mélange (ou de façon équivalente des sources) quand les sources sont supposées avoir une distribution gaussienne (dont les paramètres sont inconnus). Nous avons vu (1.4) que l'information mutuelle est équivalente à la corrélation dans le cas où les sources sont Gaussiennes. Maintenant, si l'on suppose que les sources sont distribuées selon une distribution gaussienne iid de moyenne nulle et de covariance $\mathbf{\Sigma}$, $\mathbf{s}(t) \sim \mathcal{N}(0, \mathbf{\Sigma})$, alors on peut montrer [Car07] que la quantité à minimiser dans l'équation (1.3) peut s'écrire :

$$K(P_{\mathbf{Y}}|P_{\mathbf{S}}) = K(\mathcal{N}(0, \bar{\mathbf{R}}_y)|\mathcal{N}(0, \mathbf{\Sigma})) \tag{A.1}$$

$$= \mathrm{D}_{\mathrm{KL}}\left(\bar{\mathbf{R}}_y, \mathbf{\Sigma}\right) \tag{A.2}$$

où $\mathrm{D}_{\mathrm{KL}}(\mathbf{R}_1, \mathbf{R}_2)$ est la divergence de Kullback entre deux distributions gaussiennes de même moyenne, ayant les matrices de covariances $\mathbf{R}_1$ et $\mathbf{R}_2$ de taille $d \times d$ et est définie par :

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{R}_1, \mathbf{R}_2) \stackrel{\triangle}{=} \frac{1}{2}\left(\mathrm{tr}\left(\mathbf{R}_1\mathbf{R}_2^{-1}\right) - \log\det\left(\mathbf{R}_1\mathbf{R}_2^{-1}\right) - d\right) \tag{A.3}$$

Or, les matrices de covariances $\bar{\mathbf{R}}_y$ et $\mathbf{\Sigma}$ étant symétriques, l'identification de $\bar{\mathbf{R}}_y$ avec $\mathbf{\Sigma}$ ne fournit que $M(M+1)/2$ équations tandis que l'identification de $\mathbf{A}$ nécessite dans

le cas général M^2 équations. Il n'est pas possible par conséquent d'identifier la matrice de mélange si les sources sont iid gaussiennes.

A.1.2 Spécification d'une distribution gaussienne des sources diagonale dans une base de Fourier

Nous discutons ici de l'identification du mélange (ou de façon équivalente des sources) quand les sources sont supposées avoir une distribution gaussienne dans une base de Fourier.

Si les signaux sont stationnaires, ce qui est une hypothèse souvent admise pour les signaux audio, du moins localement, alors les coefficients de la transformée de Fourier discrète sont décorrélés, et la matrice de covariance de chaque source est donc diagonale. Les valeurs de cette diagonale correspondent à la densité spectrale de puissance de cette source.

Ainsi supposons que les sources soient distribuées par un mélange de Gaussiennes (complexes circulaires) :

$$P_{S_n}(S_n) = \prod_f N_c\left(S_n(f); \sigma_n^2(f)\right)$$

où $N_c\left(S; \sigma^2\right)$ est la densité de probabilité d'une variable aléatoire gaussienne complexe circulaire de moyenne nulle définie par :

$$N_c\left(S; \sigma^2\right) = \prod_f \frac{1}{\pi\sigma^2} \exp\left[-\frac{|S|^2}{\sigma^2}\right].$$

Les sources étant indépendantes, la probablilité jointe s'écrit :

$$P_{\mathbf{S}}(\mathbf{S}) = \prod_f \prod_n N_c\left((S_n(f); \sigma_n^2(f)\right).$$

Alors on peut montrer [Car07] que :

$$I(\mathbf{Y}) = \sum_f C(\mathbf{Y}(f)) = C(\mathbf{Y}) - \sum_n G(\mathbf{Y}_n) + cst \tag{A.4}$$

$$G(\mathbf{Y}_n) = \sum_f \mathrm{D_{KL}}\left(\sigma_n^2(f), \frac{1}{F}\sum_{f'} \sigma_n^2(f')\right) \tag{A.5}$$

F est le nombre de fréquences de la base de Fourier. Ainsi, contrairement au cas iid gaussien, l'information mutuelle n'est pas égale à la corrélation, mais à la somme des corrélations à chaque fréquence. Pour minimiser l'information mutuelle, il s'agit alors de décorréler conjointement (sur l'ensemble des fréquences) les sources. Autrement dit, il s'agit de diagonaliser conjointement les matrices de covariances spectrales (voir [Pha01]). La mesure de "non-iid-gaussianité", qui apparait dans l'équation (A.5) est une mesure

de "couleur" spectrale, c.-à-d. d'écart du spectre à la blancheur (spectre constant). Il apparaît [Car07] que la condition nécessaire à la séparabilité des sources est que celles-ci n'aient pas des densitées spectrales proportionnelles ($\{\sigma_n^2(f)\}_f \neq \alpha\{\sigma_j^2(f)\}_f$ pour $\forall n \neq j$). Autrement dit, une diversité des densité spectrales des sources est nécessaire pour pouvoir séparer les sources en aveugle.

Si l'on suppose connu les densités spectrales des sources, $\mathbf{\Sigma}(f) = \mathrm{diag}\left(\left[\sigma_n^2(f)\right]_{n=1}^N\right)$, alors la maximisation de la vraisemblance qui dans le cas iid gaussien s'écrivait suivant l'équation (A.1) revient à minimiser le contraste de vraisemblance :

$$K(P_{\mathbf{Y}}|P_{\mathbf{S}}) = \frac{1}{F}\sum_f \mathrm{D_{KL}}\left(\widehat{\mathbf{R}}_y(f), \mathbf{\Sigma}(f)\right) \tag{A.6}$$

où $\widehat{\mathbf{R}}_y(f) = \widehat{\mathbb{E}}\left\{\mathbf{y}(f)\mathbf{y}(f)^T\right\}$. $\widehat{\mathbb{E}}$ est une espérance empirique. Ici nous n'avons qu'un seul échantillon pour calculer l'espérance empirique, ce qui n'est pas très satisfaisant, cependant, si nous faisions une TFCT (Transformée de Fourier à Court Terme), cette espérance empirique serait alors le moyennage sur l'ensemble des trames.

A.2 ACI sous-déterminée

A.2.1 Estimation des paramètres du modèle bruité de l'ACI sous-déterminée

Nous dérivons dans cette section les estimations au maximum de vraisemblance des paramètres $\xi = (\mathbf{A}, \mathbf{R}_b, \mathbf{\Lambda})$ du modèle bruité de la section 1.2.

Afin d'estimer les paramètres du mélange $\xi = (\mathbf{A}, \mathbf{R}_b, \mathbf{\Lambda})$, où $\mathbf{\Lambda}$ est l'ensemble des paramètres de la densité de probabilité $P_{\mathbf{s}}$ des sources, on a besoin de définir un critère à optimiser. Comme dans le cas de l'ICA (1.3), on peut facilement établir la relation entre la vraisemblance du mélange $\mathbf{X}$ et la divergence de Kullback-Leibler entre les observations et le modèle :

$$K\left(P(\mathbf{X})\,|P(\mathbf{X};\xi)\right) = -\mathbb{E}\{\log P(\mathbf{X};\xi)\} - H(\mathbf{X}) \tag{A.7}$$

Le second terme $H(\mathbf{X})$ est l'entropie du mélange, qui est indépendante des paramètres ξ. La distribution du mélange étant supposé iid, la vraisemblance peut s'écrire :

$$\mathbb{E}\{\log P(\mathbf{X};\xi)\} = -T\left(K\left(P(\mathbf{x})\,|P(\mathbf{x};\xi)\right) + H(\mathbf{x})\right) \tag{A.8}$$

Par conséquent maximiser la vraisemblance est équivalent à minimiser la divergence de Kullback-Leibler entre la densité de probabilité des observations et celle postulée par le modèle avec les paramètres ξ.

Contrairement au cas (déterminé) sans bruit vu à la section précédante, la maximisation de la vraisemblance ne consiste plus à minimiser l'information mutuelle comme dans le cas de l'ACI [Car07] (équation (1.3)).

La dérivée partielle de la log-vraisemblance est donnée par [Car07] :

$$\frac{\partial \log P(\mathbf{X};\xi)}{\partial \xi} = \mathbb{E}\left\{ \frac{\partial \log P(\mathbf{X},\mathbf{S};\xi)}{\partial \xi} \middle| \mathbf{X};\xi \right\} \tag{A.9}$$

$$= \mathbb{E}\left\{ \frac{\partial \log P(\mathbf{X}|\mathbf{S};\xi_x)}{\partial \xi_x} \middle| \mathbf{X};\xi \right\} \tag{A.10}$$

$$+ \mathbb{E}\left\{ \frac{\partial \log P(\mathbf{S};\xi_s)}{\partial \xi_s} \middle| \mathbf{X};\xi \right\} \tag{A.11}$$

où $\xi_x = (\mathbf{A}, \mathbf{R}_b)$ et $\xi_s = (\mathbf{\Lambda})$ sont les paramètres de ξ concernant les densités de probabilité respectivement de $P(\mathbf{x}|\mathbf{s};\xi)$ et $P(\mathbf{s};\xi)$.

Les paramètres ξ_x et ξ_s étant disjoints, l'optimisation de la vraisemblance globale donnée par l'annulation de l'équation (A.9) se fait indépendamment pour les paramètres du mélange ξ_x en annulant l'équation (A.10), et les paramètres des sources ξ_s, en annulant l'équation (A.11).

Le bruit $\mathbf{b}$ étant gaussien, la densité de probabilité de $P(\mathbf{x}|\mathbf{s};\xi_x)$ est elle aussi Gaussienne. Ainsi, $P(\mathbf{x}|\mathbf{s};\xi_x) = P_{\mathbf{b}}(\mathbf{x} - \mathbf{A}\mathbf{s}) = \mathcal{N}(\mathbf{x}; \mathbf{A}\mathbf{s}, \mathbf{R}_b)$. Il est maintenant possible d'estimer les paramètres ξ_x au maximum de vraisemblance, dont les formules obtenues par simple dérivation de l'équation (A.10) sont données entre autres par Attias [Att99] et Cardoso [Car07]. La distribution du mélange $\mathbf{X}$ étant iid, $\log P(\mathbf{X};\xi) = \sum_t \log P(\mathbf{x}(t);\xi)$ et donc :

$$\frac{\partial \log P(\mathbf{X};\xi)}{\partial \mathbf{A}} = \sum_t \frac{\partial \log P(\mathbf{x}(t);\xi)}{\partial \mathbf{A}}, \quad \frac{\partial \log P(\mathbf{X};\xi)}{\partial \mathbf{R}_b} = \sum_t \frac{\partial \log P(\mathbf{x}(t);\xi)}{\partial \mathbf{R}_b} \tag{A.12}$$

avec :

$$\frac{\partial \log P(\mathbf{x};\xi)}{\partial \mathbf{A}} = \mathbf{R}_b^{-1}\left(\mathbf{x}\mathbb{E}\left\{\mathbf{s}^T|\mathbf{x};\xi\right\} - \mathbf{A}\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}\right) \tag{A.13}$$

$$\frac{\partial \log P(\mathbf{x};\xi)}{\partial \mathbf{R}_b} = \frac{1}{2}\mathbf{R}_b^{-1}\left(\mathbf{x}\mathbf{x}^T - \mathbf{A}\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\}\mathbf{x}^T - \mathbf{x}\mathbb{E}\left\{\mathbf{s}^T|\mathbf{x};\xi\right\}\mathbf{A}^T + \mathbf{A}\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}\mathbf{A}^T - \mathbf{R}_b\right)\mathbf{R}_b^{-1} \tag{A.14}$$

Ainsi, à condition que la covariance du bruit ne soit pas nulle, ce qui rendrait la densité de probabilité de $P(\mathbf{x}|\mathbf{s};\xi_x)$ singulière, on est en mesure d'estimer les paramètres $\mathbf{A}$ et $\mathbf{R}_b$ du moment que l'on sait estimer les quantitées (dites statistiques suffisantes) $\mathbb{E}\{\mathbf{s}|\mathbf{x};\xi\}$ et $\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\}$.

L'annulation des dérivées par rapport à $\mathbf{A}$ (équation (A.13)) et $\mathbf{R}_b$ (équation (A.14)) nous fournit les formules d'estimation :

$$\widehat{\mathbf{A}} = \widehat{\mathbf{R}}_{xs|x,\xi}\widehat{\mathbf{R}}_{s|x,\xi}^{-1} \tag{A.15}$$

$$\widehat{\mathbf{R}}_b = \widehat{\mathbf{R}}_x - \mathbf{A}\widehat{\mathbf{R}}_{sx|x,\xi} - \widehat{\mathbf{R}}_{xs|x,\xi}\mathbf{A}^T + \mathbf{A}\widehat{\mathbf{R}}_{s|x,\xi}\mathbf{A}^T \tag{A.16}$$

avec

$$\widehat{\mathbf{R}}_x = \frac{1}{T}\sum_t \mathbf{x}(t)\mathbf{x}(t)^T \tag{A.17}$$

$$\widehat{\mathbf{R}}_{sx|x,\xi} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}|\mathbf{x}(t);\xi\right\}\mathbf{x}(t)^T \tag{A.18}$$

$$\widehat{\mathbf{R}}_{s|x,\xi} = \frac{1}{T}\sum_t \mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x}(t);\xi\right\} \tag{A.19}$$

$$\widehat{\mathbf{R}}_{xs|x,\xi} = \widehat{\mathbf{R}}_{sx|x,\xi}^T \tag{A.20}$$

L'estimation $\widehat{\mathbf{R}}_b$ donnée à l'équation (A.16) dépend de $\mathbf{A}$, mais si l'on ne connait pas $\mathbf{A}$, et que l'on souhaite estimer $\widehat{\mathbf{R}}_b$ et $\mathbf{A}$ conjointement, alors on peut substituer la valeur de $\mathbf{A}$ par son estimation donnée par l'équation (A.15). Ce qui donne après simplification :

$$\widehat{\mathbf{R}}_b = \widehat{\mathbf{R}}_x - \widehat{\mathbf{R}}_{xs|x,\xi}\widehat{\mathbf{R}}_{s|x,\xi}^{-1}\widehat{\mathbf{R}}_{sx|x,\xi} = \widehat{\mathbf{R}}_x - \widehat{\mathbf{R}}_{xs|x,\xi}\widehat{\mathbf{A}}^T \tag{A.21}$$

A.2.2 Spécification d'une distribution gaussienne des sources

On suppose que les sources $\mathbf{s}$ sont des Gaussiennes iid de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$:

$$P(\mathbf{s}) = \mathcal{N}(\mathbf{s};\boldsymbol{\mu},\boldsymbol{\Sigma})$$

Les sources étant des Gaussiennes indépendantes, la matrice de covariance $\boldsymbol{\Sigma}$ est diagonale : $\boldsymbol{\Sigma} = \mathrm{diag}\left(\left[\sigma_1^2,\ldots,\sigma_N^2\right]\right)$.

Identifiabilité du modèle : Il est intéressant de remarquer ce que devient l'expression de la vraisemblance lorsque la distribution des sources est une Gaussienne. Les sources étant maintenant gaussiennes, le mélange devient lui aussi gaussien, et maximiser la vraisemblance devient équivalent [Car07] à minimiser la divergence de Kullback-Leibler (A.8) qui dans le cas gaussien a une forme explicite :

$$K\left(P\left(\mathbf{x}\right)|P\left(\mathbf{x}|\,\xi\right)\right) = K\left(\mathcal{N}\left(\bar{\mathbf{x}},\widehat{\mathbf{R}}_x\right)\middle|\,\mathcal{N}\left(\mathbf{A}\boldsymbol{\mu},\mathbf{R}_{x,\xi}\right)\right) \tag{A.22}$$

avec $\bar{\mathbf{x}} \triangleq \frac{1}{T}\sum_t \mathbf{x}(t)$, $\mathbf{R}_{x,\xi} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{R}_b$ et $\widehat{\mathbf{R}}_x$ est la matrice de covariance empirique du mélange définie à l'équation (A.17).

Pour simplifier le problème, nous considérons que les sources sont centrées ($\boldsymbol{\mu} = \mathbf{0}$). Par conséquent, le mélange est lui aussi centré $\mathbb{E}\left\{\mathbf{x}\right\} = \mathbf{0}$, et la divergence de Kullback-Leibler peut s'écrire :

$$K\left(P\left(\mathbf{x}\right)|P\left(\mathbf{x}|\,\xi\right)\right) = \mathrm{D}_{\mathrm{KL}}\left(\widehat{\mathbf{R}}_x,\mathbf{R}_{x,\xi}\right) \tag{A.23}$$

où $\mathrm{D}_{\mathrm{KL}}(.,.)$ est définie par l'équation (A.3).

La maximisation de la vraisemblance consiste donc à identifier la covariance du modèle $\mathbf{R}_{x,\xi}$ avec la covaraince empirique $\widehat{\mathbf{R}}_x$. Afin que cette identification soit unique, il est nécessaire d'avoir un nombre d'équations supérieur ou égale au nombre d'inconnues. La matrice $\widehat{\mathbf{R}}_x$ est symétrique et fournit donc $M(M+1)/2$ équations, tandis que le nombre d'inconnues est de $(M-1)N$ pour la matrice de mélange (si les directions sont normées), N pour la matrice de covariance des sources qui est diagonale, et M pour le bruit (en considérant un bruit blanc décorrélé). Le modèle gaussien, comme dans le cas de l'ACI, pose des problèmes d'identifiabilité, car il impose que le nombre d'inconnues soit inférieur ou égale à $M(M+1)/2$, ce qui est peu étant donné le nombre d'inconnues du modèle.

Distribution *a posteriori* des sources : Par contre il est aisé de calculer les espérances conditionnelles des équations (A.18) et (A.19), car la distribution à posteriori de $\mathbf{s}$ sachant $\mathbf{x}$ est alors elle aussi gaussienne et vaut (en utilisant le théorème 10.3 de Kay [Kay93]) :

$$
\begin{aligned}
P(\mathbf{s}|\mathbf{x};\xi) &= \mathcal{N}\left(\mathbf{s};\boldsymbol{\rho}(\mathbf{x}),\mathbf{C}\right) && \text{(A.24)}\\
\boldsymbol{\rho}(\mathbf{x}) &= \boldsymbol{\mu} + \mathbf{W}\left(\mathbf{x} - \mathbf{A}\boldsymbol{\mu}\right) && \text{(A.25)}\\
\mathbf{W} &= \boldsymbol{\Sigma}\mathbf{A}^T\mathbf{B} && \text{(A.26)}\\
\mathbf{C} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T\mathbf{B}\,\mathbf{A}\boldsymbol{\Sigma} && \text{(A.27)}\\
\mathbf{B} &= \left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{R}_b\right)^{-1} && \text{(A.28)}
\end{aligned}
$$

La moyenne et la covariance de la distribution de $P(\mathbf{s}|\mathbf{x};\xi)$, définies par les équations (A.26) et (A.27) peuvent s'écrire sous une autre forme en utilisant le lemme d'inversion matricielle [Kay93] :

$$
\begin{aligned}
\mathbf{W} &= \mathbf{C}\mathbf{A}^T\mathbf{R}_b^{-1} && \text{(A.29)}\\
\mathbf{C} &= \left(\boldsymbol{\Sigma}^{-1} + \mathbf{A}^T\mathbf{R}_b^{-1}\,\mathbf{A}\right)^{-1} && \text{(A.30)}
\end{aligned}
$$

Il est plus avantageux d'utiliser les expressions de $\mathbf{W}$ et $\mathbf{C}$, données par les équations (A.26) et (A.27) dans le cas sous-déterminé, car la matrice à inverser $\mathbf{B}$ est de taille $M \times M$. Par contre dans le cas sur-déterminé, il est plus intéressant d'utiliser les équations (A.29) et (A.30), car les matrices $\boldsymbol{\Sigma}$, $\mathbf{R}_b$ et $\mathbf{C}$ à inverser sont de taille $N \times N$.

Dans tous les cas, si $\boldsymbol{\Sigma}$ et $\mathbf{R}_b$ sont inversibles, alors $\mathbf{B}$ est inversible si et seulement si la forme $\mathbf{C}$ de l'équation (A.30) est inversible.

Calcul des espérances conditionnelles : On peut maintenant donner la valeur des espérances conditionelles nécessaires aux calcul des équations (A.18) et (A.19) :

$$
\begin{aligned}
\mathbb{E}\left\{\mathbf{s}|\mathbf{x};\xi\right\} &= \boldsymbol{\rho}(\mathbf{x}) && \text{(A.31)}\\
\mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x};\xi\right\} &= \boldsymbol{\rho}(\mathbf{x})\boldsymbol{\rho}(\mathbf{x})^T + \mathbf{C} && \text{(A.32)}
\end{aligned}
$$

Estimation des paramètres des sources : Maintenant que la distribution a posteriori sur les sources est définie, on peut estimer les paramètres $\xi_s = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ des sources par annulation de l'équation (A.11). Ce qui conduit aux estimations :

$$\hat{\boldsymbol{\mu}} = \frac{1}{T} \sum_t \mathbb{E}\left\{\mathbf{s}|\mathbf{x}(t); \xi\right\} \tag{A.33}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_t \mathbb{E}\left\{\mathbf{s}\mathbf{s}^T|\mathbf{x}(t); \xi\right\} - \mathrm{diag}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T\right) \tag{A.34}$$

A.2.3 Comportement de l'algorithme EM quand le bruit tend vers zéro

Nous étudions maintenant le fonctionnement de l'algorithme EM, quand la covariance du bruit tend vers zéro. Nous considérons la limite $\mathbf{R_b} = \sigma_b^2\mathbf{I}, \sigma_b^2 \to 0$.

A.2.3.1 Image des sources

Nous démontrons ici que, quand le bruit est nul, l'image de la covariance des sources conditionnellement aux observations $\mathbf{C}$ est égal au noyau de la matrice de mélange $\mathbf{A}$ (on suppose que les colonnes de $\mathbf{A}$ sont linéairement indépendantes) :

$$\mathrm{Im}\mathbf{C} = \mathrm{Ker}\mathbf{A} \tag{A.35}$$

Dans les cas (sur-)déterminés, cette relation est évidente car $\mathbf{C} = \mathbf{0}$ et Ker$\mathbf{A}$ est un ensemble vide d'après le théorème du rang.

Dans le cas sous-déterminé, la matrice de covariance des sources s'écrit suivant l'équation (1.16) :

$$\mathbf{C} = \left(\mathbf{I}_N - \mathbf{W}\mathbf{A}\right)\boldsymbol{\Sigma}$$

$\mathbf{W}$ étant une inverse à droite de $\mathbf{A}$ (c.a.d. $\mathbf{A}\mathbf{W} = \mathbf{I}_M$), et comme les valeurs propres non nulles de $\mathbf{W}\mathbf{A}$ sont les mêmes que celles de $\mathbf{A}\mathbf{W}$, Il en découle que la matrice $\mathbf{W}\mathbf{A}$ possède M valeurs propres égales à 1 et $N - M$ valeurs propres nulles. D'autre part on peut remarquer que le noyau de $\mathbf{A}$, de dimension $N - M$ est aussi le noyau de $\mathbf{W}\mathbf{A}$. Il est alors facile de vérifier que la matrice $\mathbf{I}_N - \mathbf{W}\mathbf{A}$ possède M valeurs propres nulles et $N - M$ valeurs propres égales à 1, et dont l'espace propre associé, est le noyau de $\mathbf{A}$. Par concéquent l'image de $\mathbf{C}$ est le noyau de $\mathbf{A}$.

A.2.3.2 Formule de ré-estimation de la matrice de mélange

Nous étudions ici ce que devient la formule de ré-estimation de la matrice de mélange de l'algorithme EM défini à la section 1.3 par de l'équation (1.11).

cas d'un mélange (sur-)déterminé : Dans les cas où les sources sont gaussiennes ou des mélanges de Gaussiennes, la formule de ré-estimation de la matrice de mélange,

donnée par l'équation (1.11) s'écrit [Att99] dans le cas d'un mélange sur-déterminé et quand le bruit devien nul :

$$\widehat{\mathbf{R}}_{xs|x,\xi^{(l)}}\widehat{\mathbf{R}}^{-1}_{s|x,\xi^{(l)}} = \widehat{\mathbf{R}}_x\mathbf{A}\big(\mathbf{A}^T\widehat{\mathbf{R}}_x\mathbf{A}\big)^{-1}\cdot\big(\mathbf{A}^T\mathbf{A}\big) \tag{A.36}$$

Si bien que si $\mathbf{A}$ est carré (cas déterminé), la formule de réestimation devient :

$$\mathbf{A}^{(l+1)} = \mathbf{A}^{(l)} \tag{A.37}$$

Dans le cas sur-déterminé, il est assez facile de remarquer qu'il y a une infinité de matrice $\mathbf{A}$ qui sont des points stationnaires de la formule de réestimation (1.11). En effet, si l'on suppose que $\mathbf{A}$ est un point stationnaire, c.a.d. $\mathbf{A} = \widehat{\mathbf{R}}_x\mathbf{A}\big(\mathbf{A}^T\widehat{\mathbf{R}}_x\mathbf{A}\big)^{-1}\cdot\big(\mathbf{A}^T\mathbf{A}\big)$, alors toute matrice $\mathbf{AU}$, avec $\mathbf{U}$ matrice carrée $N\times N$ inversible est un point stationnaire de l'algorithme.

cas d'un mélange sous-déterminé : Dans le cas de sources Gaussiennes, on peut remarquer en développant les formules de $\widehat{\mathbf{R}}_{xs|x,\xi^{(l)}}$ et $\widehat{\mathbf{R}}_{s|x,\xi^{(l)}}$ et en appliquant la relation $\mathbf{A}^{(l)}\mathbf{W}^{(l)} = \mathbf{I}_M$ que lorsque le bruit est nul : $\widehat{\mathbf{R}}_{xs|x,\xi^{(l)}} = \mathbf{A}^{(l)}\widehat{\mathbf{R}}_{s|x,\xi^{(l)}}$. Par conséquent la formule de réestimation (1.11) ne change pas les valeurs de la matrice de mélange :

$$\mathbf{A}^{(l+1)} = \widehat{\mathbf{R}}_{xs|x,\xi^{(l)}}\widehat{\mathbf{R}}^{-1}_{s|x,\xi^{(l)}} = \mathbf{A}^{(l)} \tag{A.38}$$

Il est facile de vérifier que la même chose se produit avec les MMG (en remarquant que $\forall\mathbf{k}, \mathbf{A}^{(l)}\mathbf{W}^{(l)}_{\mathbf{k}} = \mathbf{I}_M$), et par conséquent l'algorithme EM avec le modèle non bruité ne permet pas d'estimer la matrice de mélange quelque soit la détermination du mélange. En fait, l'estimation de la matrice de mélange ne fonctionnera pas quelque soit la distribution des sources. Pour s'en rendre compte, il faut remarquer que la distribution des sources à l'étape (l) connaissant l'observation $\mathbf{x}(t)$, dans le cas non bruité, impose une probabilité non nulle uniquement pour les valeurs des sources $\mathbf{s}$ qui vérifient $\mathbf{A}^{(l)}\mathbf{s} = \mathbf{x}(t)$. Si bien que

$$\mathbf{A}^{(l)}\mathbb{E}\{\mathbf{s}\mathbf{s}^T|\mathbf{x}(t);\xi^{(l)}\} = \mathbf{x}(t)\mathbb{E}\{\mathbf{s}^T|\mathbf{x}(t);\xi^{(l)}\}$$

et la relation (A.38) est vérifiée quelque soit la distribution des sources.

Annexe B

Algorithmes EM pour l'apprentissage de MMG spectraux

Nous présentons quelques versions utilisés dans cette thèse de l'algorithme EM [DLR] pour l'apprentissage des MMG spectraux.

B.1 L'apprentissage oracle

Le problème de l'apprentissage du MMG spectrale d'une source λ à partir de la source oracle $S(t) = [S(t,f)]_f$, consiste à chercher le modèle MMG λ qui maximise le critère du Maximum de Vraisemblance (MV) suivant :

$$\lambda = \arg\max_{\lambda'} P(S|\,\lambda') \tag{B.1}$$

où $S = [S(t,f)]_{t,f=1}^{T,F}$ $\lambda = \{\pi_k, \Sigma_k\}_{k=1}^{\mathcal{K}}$, où $\mathcal{K}$ est le nombre d'états, π_k sont les poids des Gaussiennes, et Σ_k sont des matrices de covariances diagonales $\Sigma_k = \mathrm{diag}\,[\sigma_k^2(f)]_{f=1}^F$.

L'algorithme EM d'estimation des MMG utilisé par Benaroya [Ben03] est décrit par l'algorithme 1. L'initialisation de l'algorithme EM est effectuée à l'aide de l'algorithme K-Means [Mac67].

Rappelons que la réalisation d'une variable aléatoire $V = [V(f)]_f$ Gausienne complexe circulaire de moyenne $\mu = [\mu(f)]_f$ et de matrice de covariance diagonale $\Sigma = \mathrm{diag}\,[\sigma^2(f)]_f$, $P(V|\,\mu,\Sigma) = N_c(V;\mu,\Sigma)$, est définie à l'équation (B.2).

$$N_c(V;\mu,\Sigma) = \prod_f \frac{1}{\pi\sigma^2(f)} \exp\left[-\frac{|V(f)-\mu(f)|^2}{\sigma^2(f)}\right], \tag{B.2}$$

Algorithm 1 Algorithme EM pour l'estimation oracle du MMG d'une source au sens du MV (l'indice d'exposant (l) représente la l-ième itération de l'algorithme)

1. Calculer les poids $\gamma_k^{(l)}(t)$ satisfaisant $\sum_k \gamma_k^{(l)}(t) = 1$ et

$$\gamma_k^{(l)}(t) \triangleq P(K(t) = k|S; \lambda^{(l)}) \propto \pi_k^{(l)} N_c(S(t); \bar{0}, \Sigma_k^{(l)}) \tag{B.3}$$

   où $K(t)$ est l'état courant du MMG λ à l'instant t, et $N_c(V; \mu, \Sigma)$ est définie par (B.2).

2. Réestimer les poids des Gausiennes

$$\pi_k^{(l+1)} = \frac{1}{T} \sum_t \gamma_k^{(l)}(t) \tag{B.4}$$

3. Réestimer les matrices de covariances

$$\sigma_k^{2,(l+1)}(f) = \frac{\sum_t \left\langle |S(t,f)|^2 \right\rangle_k^{(l)} \gamma_k^{(l)}(t)}{\sum_t \gamma_k^{(l)}(t)} \tag{B.5}$$

B.2 Le modèle source + bruit

Nous avons le modèle de mélange monophonique :

$$\tilde{S}(t,f) = S(t,f) + \tilde{E}(t,f), \tag{B.6}$$

où à partir des observations de $\tilde{S}(t,f)$, et de la connaissance de la matrice de covariance diagonale $\tilde{\Sigma}_t$ de la variable aléatoire Gausienne de moyenne nulle $\tilde{E}(t) = [\tilde{E}(t,f)]_f$, on cherche à estimer le modèle MMG spectral λ de la source S.

Ce modèle MMG, dont la structure a été décrite par A.Ozerov [OPBG07], est paramétrisé par : $\lambda = \{\pi_k, \Sigma_k\}_{k=1}^{\mathcal{K}}$, où $\mathcal{K}$ est le nombre d'états, π_k sont les poids des Gaussiennes, et Σ_k sont des matrices de covariances diagonales $\Sigma_k = \text{diag}\,[\sigma_k^2(f)]_{f=1}^F$.

Nous cherchons alors à estimer λ suivant le critère de MV suivant :

$$\lambda = \arg\max_{\lambda'} P(\tilde{S}|\tilde{\Sigma}; \lambda') \tag{B.7}$$

où $\tilde{S} = [\tilde{S}(t,f)]_{t,f=1}^{T,F}$ et $\tilde{\Sigma} = [\tilde{\Sigma}_t]_{t=1}^T$. L'algorithme 2 décrit un algorithme EM (Expectation-Maximization) pour optimiser le critère (B.7) (voir [OPBG07] pour plus de détails concernant la dérivation de ces équations). L'initialisation de l'algorithme EM est effectuée à l'aide de de l'algorithme K-Means [Mac67] sur les énergies du spectre du mélange $\tilde{S}(t,f)$.

Algorithm 2 Algorithme EM pour l'estimation du MMG d'une source, à partir de son estimation corrompue par un bruit dont la distribution gaussienne est connue, au sens du MV (l'indice d'exposant (l) représente la l-ième itération de l'algorithme)

1. Calculer les poids $\gamma_k^{(l)}(t)$ satisfaisant $\sum_k \gamma_k^{(l)}(t) = 1$ et

$$\gamma_k^{(l)}(t) \triangleq P(K(t) = k|\tilde{S}, \tilde{\Sigma}; \lambda^{(l)}) \propto \pi_k^{(l)}\, N_c(\tilde{S}(t); \bar{0}, \Sigma_k^{(l)} + \tilde{\Sigma}_t) \tag{B.8}$$

   où $K(t)$ est l'état courant du MMG λ à l'instant t, et $N_c(V; \mu, \Sigma)$ est définie par (B.2).

2. Calculer les DSP pour l'état $K(t) = k$

$$\begin{aligned}\left\langle |S(t,f)|^2 \right\rangle_k^{(l)} &\triangleq \mathbb{E}_S\left[|S(t,f)|^2 \,\middle|\, K(t) = k, \tilde{S}, \tilde{\Sigma}; \lambda^{(l)} \right] = \\ &= \frac{\sigma_k^{2,(l)}(f)\tilde{\sigma}_t^2(f)}{\sigma_k^{2,(l)}(f) + \tilde{\sigma}_t^2(f)} + \left| \frac{\sigma_k^{2,(l)}(f)}{\sigma_k^{2,(l)}(f) + \tilde{\sigma}_t^2(f)} \tilde{S}(t,f) \right|^2 \end{aligned} \tag{B.9}$$

3. Réestimer les poids des Gausiennes

$$\pi_k^{(l+1)} = \frac{1}{T} \sum_t \gamma_k^{(l)}(t) \tag{B.10}$$

4. Réestimer les matrices de covariances

$$\sigma_k^{2,(l+1)}(f) = \frac{\sum_t \left\langle |S(t,f)|^2 \right\rangle_k^{(l)} \gamma_k^{(l)}(t)}{\sum_t \gamma_k^{(l)}(t)} \tag{B.11}$$

B.3 Le mélange multicanal

Nous avons le modèle de mélange multicanal :

$$\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f), \tag{B.12}$$

où à partir des observations $\mathbf{X}(t) = [\mathbf{X}(t,f)]_{f=1}^{F}$, et de la connaissance de la matrice de mélange $\mathbf{A}$, on cherche à estimer les modèles MMG spectraux $\mathbf{\Lambda} = [\lambda_n]_{n=1}^{N}$ des sources $\mathbf{S} = [\mathbf{S}(t,f)]_{t,f=1}^{T,F}$, avec $\mathbf{S}(t,f) = [S_n(t,f)]_{n=1}^{N}$.

Nous cherchons alors à estimer $\mathbf{\Lambda}$ suivant le critère de MV suivant :

$$\mathbf{\Lambda} = \arg\max_{\mathbf{\Lambda}'} P(\mathbf{X}|\mathbf{A};\mathbf{\Lambda}') \tag{B.13}$$

où $\mathbf{X} = [\mathbf{X}(t,f)]_{t,f=1}^{T,F}$. L'algorithme 3 décrit l'algorithme EM (Expectation-Maximization) pour optimiser le critère (B.13). Rappelons que $\Sigma_{\mathbf{k}}(f) = \operatorname{diag}[\sigma_{k_n}^2(f)]_{n=1}^{N}$ et $\mathbf{k} = [k_1, k_2, \cdots, k_N]^T$ est la variable cachée de l'état du MMG du mélange, et que k_n est la variable cachée de l'état de la source n.

Algorithm 3 Algorithme EM pour l'estimation des MMGs des sources au sens du MV à partir d'un mélange multicanal et de la connaissance de la matrice de mélange (l'indice d'exposant (l) représente la l-ième itération de l'algorithme)

1. Calculer les poids $\gamma_{\mathbf{k}}^{(l)}(t)$ satisfaisant
   $\sum_{\mathbf{k}} \gamma_{\mathbf{k}}^{(l)}(t) = 1$
   et
   $$\gamma_{\mathbf{k}}^{(l)}(t) \triangleq P(\mathbf{K}(t) = \mathbf{k} | \mathbf{X}(t), \mathbf{A}; \mathbf{\Lambda}^{(l)}) \propto \pi_{\mathbf{k}}^{(l)} \prod_f N_c(\mathbf{X}(t,f); \bar{0}, \mathbf{A}\, \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) \mathbf{A}^T), \tag{B.14}$$
   où $\mathbf{K}(t)$ est l'état courant du MMG $\mathbf{\Lambda}^{(l)}$ à l'instant t, et $N_c(V; \mu, \Sigma)$ est défini par (4.2).
2. Calculer les DSP pour l'état $\mathbf{K}(t) = \mathbf{k}$
   $$\left\langle |S_i(t,f)|^2 \right\rangle_{\mathbf{k}}^{(l)} \triangleq \mathbb{E}_{\mathbf{S}} \left[|S_i(t,f)|^2 \left| \mathbf{K}(t) = \mathbf{k}, \mathbf{X}, \mathbf{A}; \mathbf{\Lambda}^{(l)} \right. \right] = \left\langle \mathbf{S}(t,f) \mathbf{S}^T(t,f) \right\rangle_{\mathbf{k}}^{(l)} (i,i) \tag{B.15}$$
   $$\left\langle \mathbf{S}(t,f) \mathbf{S}^T(t,f) \right\rangle_{\mathbf{k}}^{(l)} \triangleq \mathbb{E}_{\mathbf{S}} \left[\mathbf{S}(t,f) \mathbf{S}^T(t,f) \left| \mathbf{K}(t) = \mathbf{k}, \mathbf{X}, \mathbf{A}; \mathbf{\Lambda}^{(l)} \right. \right] = = \mathbf{W}_{\mathbf{k}}^{(l)}(f) \mathbf{X}(t,f) \mathbf{X}^T(t,f) \mathbf{W}_{\mathbf{k}}^{T,(l)}(f) + \mathbf{C}_{\mathbf{k}}^{(l)}(f) \tag{B.16}$$
   avec
   $$\mathbf{C}_{\mathbf{k}}^{(l)}(f) = \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) - \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) \mathbf{A}^T \mathbf{R}_{\mathbf{k}}^{(l)}(f) \mathbf{A}\, \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) \tag{B.17}$$
   $$\mathbf{W}_{\mathbf{k}}^{(l)}(f) = \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) \mathbf{A}^T \mathbf{R}_{\mathbf{k}}^{(l)}(f) \tag{B.18}$$
   $$\mathbf{R}_{\mathbf{k}}^{(l)}(f) = (\mathbf{A}\, \mathbf{\Sigma}_{\mathbf{k}}^{(l)}(f) \mathbf{A}^T)^{-1} \tag{B.19}$$
3. Réestimer des poids des Gaussiennes
   $$\omega_{i,k_i}^{(l+1)} = \frac{1}{T} \sum_t \sum_{\{k_j\} j \neq i} \gamma_{\mathbf{k}}^{(l)}(t) \tag{B.20}$$
4. Réestimer des matrices de covariances
   $$\sigma_{i,k_i}^{2,(l+1)}(f) = \frac{\sum_t \sum_{\{k_j\} j \neq i} \left\langle |S_i(t,f)|^2 \right\rangle_{\mathbf{k}}^{(l)} \gamma_{\mathbf{k}}^{(l)}(t)}{\sum_t \sum_{\{k_j\} j \neq i} \gamma_{\mathbf{k}}^{(l)}(t)} \tag{B.21}$$

Annexe C

Remarques sur le modèle MGL

C.1 Remarque sur la partie imaginaire de la matrice de covariance locale

Remarquons aussi que, bien que les matrices de covariance du modèle $\mathbf{R}_x = \mathbf{A}\mathbf{\Sigma}(t,f)\mathbf{A}^T$ soient réelles dans le cas instantané, les matrices de covariance empirique des observations $\widehat{\mathbf{R}}_x$ sont complexes, car calculées à partir des points de la TFCT. La matrice du modèle étant réelle, nous proposons de ne considérer que la partie réelle de la matrice de covariance. Cette approche peut se justifier d'une part en remarquant que la partie imaginaire d'une matrice de covariance complexe représente uniquement les covariances entre les parties réelles et les parties imaginaires des éléments du vecteur complexe $\mathbf{X}$, qui sont supposées nulles par notre modèle ($\mathbf{A}\mathbf{\Sigma}(t,f)\mathbf{A}^T \in \mathbb{R}^{M\times M}$). D'autre part, on peut remarquer que la partie imaginaire de la matrice de covariance $\widehat{\mathbf{R}}_x$ n'apparait que comme une constante dans l'équation de la divergence de Kullback-leibler de l'équation (A.23) que nous rappelons ici :

$$\mathrm{D_{KL}}\left(\widehat{\mathbf{R}}_x, \mathbf{R}_x\right) = \frac{1}{2}\left(\mathrm{tr}\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) - \log\det\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) - M\right) \tag{C.1}$$

En effet, on peut montrer que si une matrice $\widehat{\mathbf{R}}_x$ est hermitienne et qu'une seconde matrice inversible $\mathbf{R}_x$ est réelle et symétrique, alors $\mathrm{tr}\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) = \mathrm{tr}\left(\Re\mathrm{e}\left(\widehat{\mathbf{R}}_x\right)\mathbf{R}_x^{-1}\right)$. Par conséquent, le premier terme de l'équation (C.1) ne dépend pas de la partie imaginaire de $\widehat{\mathbf{R}}_x$. Le second terme de l'équation (C.1), peut s'écrire :

$$\log\det\left(\widehat{\mathbf{R}}_x\mathbf{R}_x^{-1}\right) = \log\det\widehat{\mathbf{R}}_x - \log\det\mathbf{R}_x,$$

si bien que dans le second terme de l'équation (C.1) l'expression de la covariance du mélange $\widehat{\mathbf{R}}_x$ n'apparait que comme un terme constant.

Bibliographie

[AD03] F. Abrard and Y. Deville. Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach. In *ISSPA 2003*, Paris, France, July 2003. IEEE.

[AJ02] C. Avendano and J.-M. Jot. Frequency-domain techniques for stereo to multichannel upmix. In *Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pages 121–130, Espoo, Finland, 2002.

[Ant74] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist*, 2(6) :1152–1174, 1974.

[ARB05] Simon ARBERET. Estimation des directions des sources sonores dans le cas d'un mélange stéréophonique sous-déterminé. Master's thesis, Université Paris 6, 2005.

[Att99] H. Attias. Independent factor analysis. *Neural Comput.*, 11(4) :803–851, 1999.

[BB03] L. Benaroya and F. Bimbot. Wiener based source separation with HMM/GMM using a single sensor. *Proc. ICA*, pages 957–961, 2003.

[BBL+07] Michael W. Berry, Murray Browne, Amy N. Langville, Paul V. Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1) :155–173, September 2007.

[BBR07] N Bertin, R Badeau, and G Richard. Blind signal decompositions for automatic transcription of polyphonic music : Nmf and k-svd on the benchmark. In *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, pages 65–68. IEEE, April 2007.

[BC99a] O. Bermond and J-F. Cardoso. Méthodes de séparation de sources dans le cas sous-déterminé. In *GRETSI*, pages 749–752, Vannes,France, 1999.

[BC99b] Olivier Bermond and Jean-François Cardoso. Approximate likelihood for noisy mixtures. In *Proc. ICA '99, Aussois, France*, pages 325–330, 1999.

[Ben73] JP Benzecri. *L'Analyse des Données. Tome II : L'Analyse des Correspondances.* Paris, 1973.

[Ben03] Elie Laurent Benaroya. *Séparation de plusieurs sources sonores avec un seul microphone.* PhD thesis, Université de Rennes 1, 2003.

[Ber00] Olivier Bermond. *Méthodes statistiques pour la séparation de sources*. PhD thesis, ENST, 2000.

[BMC97] Olivier Bermond, Éric Moulines, and Jean-François Cardoso. Séparation et déconvolution aveugle de signaux bruités : modélisation par mélange de gaussiennes. In *Proc. GRETSI, Grenoble, France*, 1997.

[Bof03] P. Bofill. Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, 55, Issues 3-4(3) :627–641(15), October 2003.

[BZ01] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. In *Signal Processing*, volume 81, pages 2353–2362, 2001.

[Car91] Jean-François Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In *Proc. ICASSP*, pages 3109–3112, 1991.

[Car99] Jean-François Cardoso. High-order contrasts for independent component analysis. *Neural Comput.*, 11(1) :157–192, 1999.

[Car02] Jean-François Cardoso. Analyse en composantes indépendantes. In *Proc. of XXXIV Journées de Statistique*, Bruxelles, 2002. JSBL 2002. Conférence invitée.

[Car07] Jean-François Cardoso. *Vraisemblance*, volume Séparation de sources of *Traités Information-Commande-Communication (IC2)*, chapter 4, pages 115–168. Hermes, 2007. P. Comon and Ch. Jutten editors.

[CG92] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3) :167–174, 1992.

[COM94] P. COMON. Independent Component Analysis, a new concept ? *Signal Processing, Elsevier*, 36(3) :287–314, April 1994. Special issue on Higher-Order Statistics.

[Com98] P. Comon. Blind channel identification and extraction of more sources than sensors. In *SPIE Conf.*, pages 2–13, San Diego, CA, July 1998.

[CS93] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6) :362–370, December 1993.

[CS96] Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1) :161–164, January 1996.

[CSDP02] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon. Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging. In *Proc. EUSIPCO*, volume 1, pages 561–564, 2002.

[DCB05] K. Palomäki D. Campbell and G. Brown. A matlab simulation of ?shoebox ? room acoustics for use in research and teaching. *Computing and Information Systems Journal*, 9(3) :48–51, October 2005.

[DE03] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5) :2197–2202, 2003.

[Dev03] Y. Deville. Temporal and time-frequency correlation-based blind source separation methods. In *Proc. ICA2003*, pages 1059–1064, 2003.

[DH73] R.O. Duda and P.E. Hart. *Pattern classification ans Scene Analysis*. John Wiley, New York, 1973.

[DLR] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm.

[DM04] M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ica. *Vision, Image and Signal Processing, IEE Proceedings* -, 151(1) :35–43, Feb. 2004.

[DT07] M. DAVY and J.Y. TOURNERET. Classification bayésienne supervisée par processus de Dirichlet. 2007.

[EK04] J Eggert and E Körner. Sparse coding and nmf. In *in Neural Networks, IEEE International Conference on*, pages 2529–2533. IEEE, 2004.

[Eph92] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *Signal Processing, IEEE Transactions on*, 40(4) :725–735, Apr 1992.

[FCC05a] D FitzGerald, M Cranitch, and E Coyle. Non-negative tensor factorisation for sound source separation. In *Irish Signals and Systems Conference*, pages 8–12. IEEE, September 2005.

[FCC05b] D FitzGerald, M Cranitch, and E Coyle. Shifted non-negative matrix factorisation for sound source separation. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 1132–1137. IEEE, July 2005.

[Fer73] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist*, 1(2) :209–230, 1973.

[FG06] C. Fevotte and S.J. Godsill. A bayesian approach for blind separation of sparse sources. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6) :2174–2188, Nov. 2006.

[GBVF03] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003)*, pages 763–768, Nara, Japan, april 2003.

[GG88] Stuart German and Donald German. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. pages 611–634, 1988.

[GGR95] I.F. Gorodnitsky, J.S. George, and B.D. Rao. Neuromagnetic source imaging with FOCUSS : a recursive weighted minimum norm algorithm. *Electroencephalography and Clinical Neurophysiology*, 95(4) :231–251, 1995.

[Gri03] R. Gribonval. Piecewise linear source separation. In M. A. Unser, A. Aldroubi, and A. F. Laine, editors, *Wavelets : Applications in Signal and Image Processing X. Edited by Unser, Michael A.; Aldroubi, Akram; Laine, Andrew F. Proceedings of the SPIE, Volume 5207, pp. 297-310 (2003).*, volume 5207 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 297–310, November 2003.

[Gri07] R. Gribonval. *Séparation de sources 2*, chapter 10 : Séparation de sources basée sur la parcimonie, pages 395–441. hermes-science, 2007.

[GTC05] P Georgiev, FJ Theis, and A Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4) :992–996, 2005.

[HS77] E. Hoyer and R. Stork. The zoom fft using complex modulation. In *ICASSP*, volume 2, pages 78–81, May 1977.

[HS03] W. Härdel and L. Simar, editors. *Applied multivariate statistical analysis.* Spinger-Verlag, 2003.

[HW79] JA Hartigan and MA Wong. A K-means clustering algorithm. *JR Stat. Soc. Ser. C-Appl. Stat*, 28 :100–108, 1979.

[Hyv99] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3) :626–634, May 1999.

[JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. pages 105–161, 1999.

[JH91] Christian Jutten and Jeanny Herault. Blind separation of sources, part 1 : an adaptive algorithm based on neuromimetic architecture. *Signal Process.*, 24(1) :1–10, 1991.

[JK97] McLachlan Geoffrey J. and T. Krishnan. *The EM algorithm and extensions / Geoffrey J. McLachlan, Thriyambakam Krishnan.* Wiley, New York :, 1997.

[Jol02] I.T. Jolliffe. *Principal component analysis.* Springer New York, 2002.

[Kay93] Steven M. Kay. *Fundamentals of statistical signal processing : estimation theory.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[KC76] C. H. Knapp and G.C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustics,Speech, and Signal Processing*, 24(4) :320–327, 1976.

[KDR99] K. Kreutz-Delgado and BD Rao. Sparse basis selection, ICA, and majorization : towards a unifiedperspective. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, volume 2, 1999.

[KDRE+99] Kenneth Kreutz-Delgado, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Convex/Schur–Convex (CSC)

Log-Priors and Sparse Coding. In *Proc. 6th Joint Symposium on Neural Computation*, Caltech, Pasadena, California, May 1999. http ://cairo.ucsd.edu/∼kreutz/publications.htm.

[LLGS99] Te-Won Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *Signal Processing Letters, IEEE*, 6(4) :87–90, Apr 1999.

[LMVG08] A. Llagostera, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation using overcomplete dictionaries. In *Proc. IEEE ICASSP*, 2008.

[LS99] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, October 1999.

[LS00a] Daniel D. Lee and Sebastian H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, volume 13, pages 556–562, 2000.

[LS00b] Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Comput.*, 12(2) :337–365, 2000.

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[Mal00] Stéphane Mallat. *Une exploration des signaux en ondelettes*. ellipse, 2000.

[MC07] Eric Moreau and Pierre Comon. *Contraste*, volume Séparation de sources of *Traités Information-Commande-Communication (IC2)*, chapter 3, pages 73–113. Hermes, 2007. P. Comon and Ch. Jutten editors.

[MCG97] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 5 :3617–3620 vol.5, Apr 1997.

[MV06] Gianluca Monaci and Pierre Vandergheynst. Audiovisual gestalts. In *CVPRW '06 : Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 200, Washington, DC, USA, 2006. IEEE Computer Society.

[NBZJ07] N. Noorshams, M. Babaie-Zadeh, and C. Jutten. Estimating the mixing matrix in sparse component analysis based on converting a multiple dominant to a single dominant problem. In *ICA 2007*, 2007.

[NMBZJ07] F. Movahedi Naini, G.H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Estimating the mixing matrix in sparse component analysis (sca) based on multidimensional subspace clustering. In *ICT'07*, Malaysia, May 2007.

[OM99] Bruno A. Olshausen and K. Jarrod Millman. Learning sparse codes with a mixture-of-gaussians prior. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *NIPS*, pages 841–847. The MIT Press, 1999.

[OPBG07] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of bayesian models for single-channel source separation and its application

to voice/music separation in popular songs. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on*, 15(5) :1564–1578, July 2007.

[OPR05] Paul D. O'Grady, Barak A. Pearlmutter, and Scott T. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15(1) :18 – 33, March 2005.

[Oze06] Alexey Ozerov. *Adaptation de modèles statistiques pour la séparation de sources mono-capteur Application à la séparation voix/musique dans les chansons.* PhD thesis, Université de Rennes 1, 2006.

[PD05] M. Puigt and Y. Deville. Time–frequency ratio-based blind separation methods for attenuated and time-delayed sources. *Mechanical Systems and Signal Processing*, 19(6) :1348–1379, 2005.

[PD06] M. Puigt and Y. Deville. A time-frequency correlation-based blind source separation method for time-delayed mixtures. In *ICASSP*, 2006.

[PE06] R. M. Parry and I. Essa. Estimating the spatial position of spectral components in audio. In *Independent Component Analysis and Blind Signal Separation*, Lecture Notes in Computer Science (LNCS), pages 666–673, Charleston, SC, March 2006. Springer.

[Pha01] Dinh-Tuan Pham. Blind separation of instantaneous mixture of sources via the gaussian mutual information criterion. *Signal Process.*, 81(4) :855–870, 2001.

[Pic94] B. Picinbono. *Signaux aléatoires T2 Fonctions aléatoires et modèles.* Dunod Université, 1994, 1994.

[PO05] S.T Rickard P.D O'grady, B.A Pearlmutter. Survey of sparse and non-sparse methods in source separation. *IJIST*, 2005.

[PR00] Giuseppe Patanè and Marco Russo. ELBG implementation. *International Journal of Knowledge based Intelligent Engineering Systems*, 4(2) :94–109, April 2000.

[PR01] Giuseppe Patanè and Marco Russo. The enhanced LBG algorithm. *Neural Networks*, 14(9) :1219–1237, November 2001.

[RAG+] B. Rivet, A. Aubrey, L. Girin, Y. Hicks, and C. Jutten. Development and comparison of two approaches for visual speech analysis with application to voice activity detection.

[Rao98] BD Rao. Signal processing with the sparseness constraint. In *Acoustics, Speech, and Signal Processing, 1998. ICASSP'98. Proceedings of the 1998 IEEE International Conference on*, volume 3, 1998.

[Ras00] C.E. Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12 :554–560, 2000.

[Riv06] Bertrand Rivet. *La bimodalité de la parole au secours de la séparation de sources.* PhD thesis, INP Grenoble, 2006.

[Sap90] G Saporta. *Probabilités, Analyse de données et Statistiques*. Technip, 1990.

[SB03] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180, 2003.

[SPMP+01] H. Snoussi, G. Patanchon, J. Macias-Perez, A. Mohammad-Djafari, and J. Delabrouille. Bayesian blind component separation for cosmic microwave background observations. In American Institute of Physics, editor, *Bayesian Inference and Maximum Entropy Methods, MaxEnt Workshops*, pages 125–140, New York, NY, USA, August 2001.

[SSG+02] David Sodoyer, Jean-Luc Schwartz, Laurent Girin, Jacob Klinkisch, and Christian Jutten. Separation of audio-visual speech sources : a new approach exploiting the audio-visual coherence of speech stimuli. *EURASIP J. Appl. Signal Process.*, 2002(1) :1165–1173, 2002.

[TK03] S. Theodoridis and K. Koutroumbas. *Pattern recognition*. Academic Press, 2003.

[VGP07] Emmanuel Vincent, Rémi Gribonval, and Mark D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Process.*, 87(8) :1933–1950, 2007.

[Vin04] Emmanuel Vincent. *Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux*. PhD thesis, Université Pierre et Marie Curie, 2004.

[Vin07] E. Vincent. Complex nonconvex l_p norm minimisation for underdetermined source separation. In *Independent Component Analysis and Blind Source Separation (ICA)*, pages 430–437. Springer, 2007.

[Vir07] T Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. 15(3) :1066 – 1074, March 2007.

[VM01] Pulkki V and Karjalainen M. Localization of amplitude-panned virtual sources. i : Stereophonic panning. *Journal of the Audio Engineering Society*, 49(9) :739–752, 2001.

[VR04] Emmanuel Vincent and Xavier Rodet. Underdetermined source separation with structured source priors. pages 327–334. 2004.

[VSB+07] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca. First stereo audio source separation evaluation campaign : Data, algorithms and results. In *Independent Component Analysis and Blind Source Separation (ICA)*, pages 552–559. Springer, 2007.

[Wie64] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[WK85] M. Wax and T. Kailath. Detection of signals by information theoric criteria. *IEEE Transactions On Acoustics, Speech, and Signal Processing*, ASSP-33(2) :387–392, April 1985.

[WSM05] S Winter, H Sawada, and S Makino. On real and complex valued l1-norm minimization for overcomplete blind source separation. In *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 86–89, 2005.

[XI05] Rui Xu and Donald Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3) :645–678, May 2005.

[XLG03] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA, 2003. ACM.

[XXF05] Ming Xiao, Shengli Xie, and Yuli Fu. A statistically sparse decomposition principle for underdetermined blind source separation. *Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on*, pages 165–168, Dec. 2005.

[YR04] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7) :1830–1847, July 2004.

[ZP01] Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13 :863–882, 2001.

[ZW88] I. Ziskind and M. Wax. Maximum likelihood localisation of multiple sources by alternating projection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(10) :1553–1560, October 1988.

Table des figures

Résumé

La séparation de sources aveugle dans le cas sous-déterminé est un problème mal posé pour lequel on suppose que les sources sont indépendantes et parcimonieuses dans le domaine temps-fréquence. La séparation se fait alors en deux étapes : une étape d'estimation des paramètres du mélange, suivi d'une étape d'estimation des sources.

Les hypothèses faites sur les sources ne sont cependant pas valides sur l'ensemble des points temps-fréquence, si bien que les approches qui traitent naïvement de l'ensemble des points de manière identiques et indépendantes, sont peu robustes pour estimer les paramètres du mélange et les sources.

L'objet de cette thèse est d'exploiter la distribution locale du mélange dans les voisinages de chaque point temps-fréquence, afin de :

- Détecter les régions temps-fréquence où une seule source est active et d'estimer la direction de la source dominante dans ces régions ;
- Estimer la distribution des sources en chaque point temps-fréquence à l'aide de la connaissance sur les paramètres du mélange.

L'approche locale que nous proposons est étayée par un algorithme de clustering appelé DEMIX, qui estime de façon robuste les paramètres du mélange dans les cas instantanés et anéchoïques. D'autre part, l'estimation locale de la distribution des sources peut être utilisée pour apprendre des MMG spectraux qui jusqu'à présent nécessitaient une étape d'apprentissage à partir d'exemples. Nous montrons que cette approche améliore l'estimation des sources de plusieurs dB en SDR.

Abstract

Blind source separation in the underdetermined case is an ill-posed problem where it is usually assumed that sources are independent and sparse in the time-frequency domain. Separation is then done in two steps : the estimation of the mixture parameters, followed by the estimation of the sources.

The assumptions made about the sources are not valid for all the time-frequency points, so that the approaches which naively address all the points identically and independently, are little robust in estimating the mixture parameters and the sources.

In this thesis we exploit the local distribution of the mixture in the neighborhood of each time-frequency point, to :

- Detect the time-frequency regions where only one source is active and to estimate the direction of the dominant source in these regions ;
- Estimate the distribution of the sources in each time-frequency point using the knowledge on the mixture parameters.

The proposed local approach is supported by a clustering algorithm called DEMIX, which robustly estimates the mixture parameters in the instantaneous and anechoic cases. On the other hand, the local spatial distribution of the sources can be used to learn Spectral-GMM which until now required a learning step with source examples. We show that this approach improve the source estimation performance of some dB in SDR.